Západočeská univerzita v Plzni
Fakulta aplikovaných věd

STATISTICKÝ AUTOMATICKÝ PŘEKLAD ČEŠTINA – ZNAKOVANÁ ŘEČ

Ing. Jakub Kanis

dizertační práce

k získání akademického titulu doktor

v oboru *Kybernetika*

Školitel: Doc. Ing. Luděk Müller, Ph.D.
Katedra: Katedra kybernetiky

Plzeň, 2009

University of West Bohemia in Pilsen
Faculty of Applied Sciences

STATISTICAL AUTOMATIC TRANSLATION CZECH – SIGNED SPEECH

Ing. Jakub Kanis

A dissertation submitted for the degree of

Doctor of Philosophy

in *Cybernetics*

Major advisor: Doc. Ing. Luděk Müller, Ph.D.
Department: Department of Cybernetics

Pilsen, 2009

Prohlášení

Předkládám tímto k posouzení a obhajobě dizertační práci zpracovanou na závěr doktorského studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že tuto práci jsem vypracoval samostatně s použitím odborné literatury a dostupných pramenů uvedených v seznamu, jež je součástí této práce.

V Plzni, 2.7.2009

Jakub Kanis

Poděkování

Tato dizertační práce vznikla za podpory:

- Ministerstva školství, mládeže a tělovýchovy v rámci centra MŠMT LC536: „Centrum komputační lingvistiky“
- Grantové agentury Akademie věd v rámci projektu GA AV ČR 1ET101470416: „Multimodální zpracování lidské znakové a mluvené řeči počítačem pro komunikaci člověk-stroj“

Dále bych chtěl poděkovat:

- svému školiteli Doc. Ing. Luďku Müllerovi, Ph.D. za cenné rady a připomínky,
- všem kolegům z oddělení umělé inteligence Katedry Kybernetiky, zvláště pak kolegům z laboratoře UL504 za vytvoření dobrých pracovních podmínek,
- mé rodině za její všestrannou podporu, kterou mi v průběhu celého studia věnovali. Speciální poděkování pak patří mé setře Janě, která mě naučila počítat do deseti.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

3D	Trojrozměrný.
AER	Alignment Error Rate (chybová míra přiřazení).
ASL	American Sign Language (Americký znakový jazyk).
ASR	Automatic Speech Recognition (automatické rozpoznávání řeči).
BLEU	BiLingual Evaluation Understudy (náhrada dvojjazyčného vyhodnocení).
BSL	British Sign Language (Britský znakový jazyk).
CAcc	Concept Accuracy (konceptová přesnost).
CCorr	Concept Correctness (konceptová korektnost).
CSC	Czech–Signed Czech parallel corpus (Čeština–Znakovaná čeština paralelní korpus).
CzEng	Czech–English parallel corpus (Česko-anglický paralelní korpus).
DAE	Dialogue Act Editor (editor dialogových aktů).
DAMSL	Dialogue Act Markup in Several Layer (značkování dialogových aktů v několika rovinách).
DATE	Dialogue Act Tagging for Evaluation (značkování dialogových aktů pro vyhodnocení).
DCL	Doll Control Language (jazyk pro řízení loutky).

DGS	German sign language (Německý znakový jazyk).
DRS	Discourse Representation Structures (struktury pro reprezentaci promluvy).
EB	Example-Based (založený na příkladech).
EM	Expectation Maximization (očekávaní a maximalizace).
FAV	Fakulta Aplikovaných Věd.
GIS	Generalized Iterative Scaling (obecné iterativní škálování).
GUI	Graphical User Interface (grafické uživatelské prostředí).
HamNoSys	Hamburg Notation System (Hamburský notační systém).
HDPS	Head-Driven Phrase Structure (frázová struktura řízená hlavním slovem).
HHTT	Human–Human Train Timetable dialog corpus (dialogový korpus z prostředí vlakového informačního centra).
HMM	Hidden Markov Model (skrytý Markovův model).
HVS	Hidden Vector State (skrytý stavový vektor).
ISL	Irish Sign Language (Irský znakový jazyk).
ITG	Inversion Transduction Grammar (inverzní transduktivní gramatika).
JSL	Japan Sign Language (Japonský znakový jazyk).
LFG	Lexical-Functional Grammar (lexikálně funkční gramatika).
LGPL	Lesser General Public License (nižší všeobecná veřejná licence).

MAP	Maximum A Posteriori (maximální aposteriorní).
MBR	Minimum Bayes Risk (minimální Bayesovský risk).
MEL	Minimal Expected Loss (minimální střední ztráta).
MER	Minimum Error Rate (minimální chybová míra).
MERT	Minimum Error Rate Training (trénování minimální chybové míry).
ML	Maximum Likelihood (maximální věrohodnost).
MMI	Maximal Mutual Information (maximální vzájemná informace).
MT	Machine Translation (automatický (strojový) překlad).
NGT	Sign language of the Netherlands (Nizozemský znakový jazyk).
NLP	Natural Language Processing (zpracování přirozeného jazyka).
OOV	Out-Of-Vocabulary words (slova vně slovníku).
PAR	Parameterized Action Representation (parametrizovaná reprezentace akce).
PCEDT	Prague Czech–English Dependency Treebank (Pražský česko-anglický závislostní korpus).
PDT	Prague Dependency Treebank (Pražský závislostní korpus).
PER	Position-independent Error Rate (pozičně nezávislá chybová míra).
POS	Part of Speech (slovní druh).
SDO	Semantic Dimension Overlap (překryv sémantických dimenzí).
SER	Sentence Error Rate (větná chybová míra).

SG	Syntactic Groups (syntaktické skupiny).
SiGML	Sign Gesture Markup Language (jazyk pro značkování znakových gest).
SMT	Statistical Machine Translation (statistický automatický (strojový) překlad).
SRO	Semantic Role Overlap (překryv sémantických rolí).
TER	Translation Edit Rate (míra překladové editace).
WER	Word Error Rate (slovní chybová míra).
ZČ	Znakovaná Čeština.
ZČU	ZápadoČeská Univerzita v Plzni.
ČZJ	Český Znakový Jazyk.

Přehled použitého značení

$\S$	paragraf
\$	začátek nebo konec věty
$\oplus$	spojení řetězců pomocí mezery
$\{\cdot\}$	množina prvků
$\emptyset$	prázdná množina
$< \cdot, \cdot >$	množina dvojic
$\lvert \cdot \rvert$	počet prvků množiny
$\cdot \ldots \cdot$	až
$\begin{pmatrix} \cdot \\ \cdot \end{pmatrix}$	kombinační číslo
$\cdot!$	faktoriál
$=$	rovno
$\neq$	různé
$:=$	je dáno jako
$\equiv$	odpovídá
$\approx$	přibližně odpovídá
$\hat{\cdot}$	označuje hledanou proměnnou (proměnné)
$\lvert\tilde{\cdot}\rvert$	délka fráze
δ	Kroneckerova delta funkce
ϕ	fertilita

ϕ_T	překladová pravděpodobnost
κ	kappa statistika
λ	váha log-lineárního modelu
λ_1^M	posloupnost M vah log-lineárního modelu
$\mathbf{a} = a_1^J = a_1...a_j...a_J$	slovní přiřazení
b_k	začátek k-té zdrojové fráze
d	distorzní pravděpodobnost
exp	exponenciální funkce
E	chybové kritérium
$h(\cdot)$	složka log-lineárního modelu (rys)
h_{PhP}	frázová penalizace
h_{WP}	slovní penalizace
i_k	konec k-té cílové fráze
j_k	konec k-té zdrojové fráze
log	logaritmická funkce
L_T	překladová ztráta
MI, MI_T	vzájemná informace
N	počet
$p(\cdot)$	pravděpodobnostní distribuce daná modelem
$Pr(\cdot)$	obecná pravděpodobnostní distribuce
p_{LM}	pravděpodobnost jazykového modelu
p_{MI}, p_{MI_T}	pravděpodobnost založená na vzájemné informaci
p_T	pravděpodobnost potencionálního překladu
$\mathbf{s} = s_1^J = s_1...s_j...s_J$	zdrojová věta
$\tilde{s_k} = s_{b_k}....s_{j_k}$	k-tá zdrojová fráze

s_j	j-té slovo zdrojové věty
$\mathbf{t} = t_1^I = t_1 ... t_i ... t_I$	cílová věta
$\tilde{t_k} = t_{i_{k-1}} .... t_{i_k}$	k-tá cílová fráze
t_i	i-té slovo cílové věty

Abstrakt

Tato dizertační práce se zabývá návrhem systému pro automatický překlad mezi češtinou a znakovanou řečí. Pojem znakovaná řeč je v této práci použit jako souhrnné pojmenování pro český znakový jazyk a znakovanou češtinu, které oba slouží ke komunikaci neslyšících v ČR. Hlavním cílem práce tedy bylo vytvořit obecný překladový systém, který by umožnil překlad pro oba zmíněné jazyky (resp. pro libovolnou dvojici jazyků). Za tímto účelem byly prozkoumány stávající možnosti přístupů ke konstrukci automatických překladových systému a také existující systémy pro překlad znakových jazyků ve světě. Jako nejvhodnější z hlediska splnění zvoleného cíle byl vybrán statistický přístup ke konstrukci automatického překladového systému založený na frázích. Tento přístup umožňuje konstrukci překladového systému pro libovolnou jazykovou dvojici a frázové systémy jsou v současné době jedním z nejpoužívanějších typů statistických překladových systémů a dosahují špičkových výsledků z hlediska přesnosti a rychlosti překladu.

Hlavním zdrojem informací o překládaných jazycích je v případě statistických systémů paralelní korpus, který obsahuje odpovídající si texty v obou jazycích. V případě znakových jazyků i znakované řeči je existence paralelního korpusu komplikována skutečností, že neexistuje oficiální psaná forma žádného znakového jazyka ani znakované češtiny (neexistuje tedy dosud ani žádný autorovi známý paralelní korpus znakované řeči). Pro potřeby této práce byl tedy vytvořen vlastní paralelní korpus znakované řeči – Czech – Signed Czech (CSC) korpus. Tento korpus vznikl přeložením existujícího Human-Human Train Timetable dialogového korpusu (viz [Jurčíček 05, Jurčíček 07]), který obsahuje přepisy telefonních dotazů do informačního centra vlakových jízdních řádů, do znakované češtiny. Ta byla zvolena především z hlediska jednodušší možnosti vytvoření její psané formy, která je reprezentována zapsáním znaků českého znakového jazyka v pořadí odpovídajícím českým slovům v překládané větě. Takto vytvořený korpus obsahuje 15 722 větných párů rozdělených do 1 109 dialogů s množstvím anotačních vrstev použitelných pro další zpracování (ke každé promluvě v korpusu je přítomna její transkripce a normalizovaná transkripce řeči, vyznačení pojmenovaných entit, sémantický popis ve formě dialogových značek a nově přidaný překlad do znakované češtiny).

Informace získané z paralelního korpusu jsou v případě frázového systému uloženy ve frázové tabulce, která obsahuje odpovídající si překladové páry. V rámci této práce byla navržena a otestována nová metoda pro výběr těchto překladových párů založená na principu minimální ztráty. Tato metoda spolu s jejími navrženými zlepšeními (resp. získaná frázová tabulka) byla dále porovnána s dalšími dvěma frázovými tabulkami získanými jednak z ručně vytvořeného frázového přiřazení vyznačeného při vytváření CSC korpusu a dále s tabulkou získanou standardní automatickou metodou pro výběr frází. Navržená zlepšení nové metody pro výběr frází spočívají v rozdělení výběru nejlepšího překladu podle četnosti výskytu zdrojové fráze, v kombinaci frázových tabulek pro oba směry překladu a dále ve filtraci výsledné tabulky prostřednictvím překladu vhodného textu.

Dále byl představen algoritmus pro monotónní a nemonotónní prohledávání založený na dy-

namickém programování a využívající frázový n-gram. Implementací algoritmu pro monotónní prohledávání byl vytvořen vlastní dekodér použitelný pro překlad mezi češtinou a znakovanou češtinou. Při návrhu dekodéru byl kladen důraz na jeho použitelnost a snadné zapojení v reálných aplikacích. Výkonnost vlastního dekodéru byla porovnána s výkonností frázového dekodéru, který představuje současný standard mezi frázovými dekodéry. Z porovnání dekodérů vyplývá, že poskytují srovnatelné výsledky. Mezi sebou byly také porovnány verze vlastního dekodéru využívající frázový bigram a trigram v kombinaci s bigramovým a trigramovým jazykovým modelem. Z hlediska přesnosti a rychlosti se jako nejlepší kombinace jeví použití frázového bigramu spolu s trigramovým jazykovým modelem.

V rámci experimentů byly porovnány přesnosti překladu mezi češtinou a znakovanou češtinou pro všechny tři dostupné frázové tabulky. Z tohoto porovnání vyplývá, že ručně vytvořená tabulka a tabulka vytvořená standardní automatickou metodou poskytují srovnatelné výsledky, zatímco tabulka vytvořená nově navrženou metodou za nimi v přesnosti překladu mírně zaostává (z hlediska reálných aplikací je tento rozdíl však zanedbatelný). Hlavní výhodou ručně vytvořené tabulky a tabulky vytvořené novou metodou je jejich několikanásobně menší velikost oproti tabulce získané standardní metodou (12 krát v případě ručně a 5 krát v případě nově vybrané tabulky). Dále byl také otestován základní systém pro překlad mezi češtinou a znakovanou češtinou, který dosáhl 81,22 bodu BLEU skóre a navržena a otestována jeho možná vylepšení využívající informace obsažené v bohaté anotaci CSC korpusu. Šlo především o použití třídního jazykového modelu a pozpracování výsledného překladu. Tato vylepšení přinesla nárůst přesnosti překladu o více než dva body BLEU skóre v závislosti na použité frázové tabulce. Výsledný nejlepší překladový systém byl pak vyzkoušen i při opačném směru překladu ze znakované češtiny do češtiny, kde bylo dosaženo nejlepšího výsledku 63,98 bodu BLEU skóre.

Abstract

This thesis deals with design of system for automatic translation between Czech and Signed Speech. The term Signed Speech is in this work used as a overall name for Czech Sign Language and Signed Czech, which both of them serve as communication facility of deaf in the Czech Republic. The main goal of this work was to create common translation system, which can be used for translation of both mentioned languages (for an arbitrary language pair respectively). The present possibilities to an automatic translation system construction and in the world existing systems for sign language translation were surveyed for this purpose. A statistical approach to the automatic translation system construction based on phrases was chosen as an optimal solution for the main goal accomplishment. This approach allow the construction of the translation system for the arbitrary language pair and the phrase based systems are one of the most used statistical translation systems at this time and represent state of the art in the accuracy and the speed of the automatic translation.

A parallel corpus containing corresponding texts in both languages is a main source of information about translated pair in the case of statistical machine translation systems. The existence of the parallel corpus is in the case of the sign languages and the Signed Speech complicated by the absence of an official written form neither any sign language nor Signed Czech (thus it does not exists no parallel corpus of the Signed Speech till now at least to author's best knowledge). The own parallel corpus of Signed Speech – Czech – Signed Czech (CSC) corpus was created for the purpose of this work. This corpus was created by the Signed Czech translation of the existing Human-Human Train Timetable dialog corpus, which contains transcriptions of telephone queries in a train time table information center. The Signed Czech was chosen mainly because of simpler possibility of the written form creation, which is represented as a sequence of Czech Sign Language signs in the same order as the Czech words in the translated sentence. The final corpus contains 15 722 sentence pairs organized in 1 109 dialogs with a rich annotation scheme useful for next processing (there is for each utterance in the corpus its transcription and normalized transcription of speech, named entity annotation, semantic annotation in form of dialog acts and newly added Signed Czech translation).

Information acquired from the parallel corpus is in the case of the phrase based system stored in a phrase table, which contains corresponding translation pairs. An original method for the translation pairs extraction based on minimal loss principle was proposed and tested in the scope of this work. This method together with its suggested improvements (acquired phrase table respectively) was compared with two other phrase tables obtained from a handcrafted phrase alignment established during CSC corpus collecting and from a standard method for the automatic phrase extraction. The suggested improvements of the new phrase extraction method consist in dividing of the best translation selection according to a source phrase occurring frequency, in combination of the phrase tables for both translation directions and further in a filtration of the resulting table through the translation of an appropriate text.

Further, it was introduced an algorithm for monotone and non-monotone searching based

on dynamic programming using phrase n-gram. The own decoder usable for translation between Czech and Signed Czech was created by the implementation of the algorithm for the monotone searching. The design of the decoder was oriented to easy using and connection in real applications. The performance of our decoder was compared to the performance of the state of the art phrase based decoder. The results of comparison indicate the comparable performance of both decoders. Additionally, the versions of our decoder using phrase bigram and trigram in combination with bigram and trigram language model were compared together. The best combination from the accuracy and the speed of translation point of view is the phrase bigram decoder with the trigram language model.

The translation accuracy between Czech and Signed Czech was compared for all three available phrase tables in the experiment stage of this work. The results show that the handcrafted table and the standard automatic method extracted table perform equal while the table extracted with the new method is behind them about one point of BLEU score (however this difference is negligible in the case of the real applications). The main advantage of the handcrafted and the new method extracted table against the standard automatic method extracted table is their multiple smaller size (12 times in the case of handcrafted and 5 times in the case of newly extracted table). The basic system for the translation from Czech to Signed Czech was tested and reached 81.22 point of BLEU score. Its possible improvements using informations included in the rich annotation of the CSC corpus (class based language model and post processing of the resulting translation) were suggested and tested further. These improvements bring the accrual of translation accuracy about more than two point of BLEU score in dependence of used table. The resulting best translation system was then used for the opposite translation direction from Signed Czech to Czech where it reached the best result of 63.98 point of BLEU score.

Kapitola 1

Úvod

Mezi námi žijí různé menšiny, které ke své komunikaci používají menšinové jazyky, jež ovládá jen zlomek většinové společnosti. Pokud navíc příslušníci této menšiny neovládají většinový jazyk, dochází u nich ke stíženému přístupu k informacím, které jsou primárně přístupné jen v majoritním jazyce. Tento stížený přístup k informacím pak přímo ovlivňuje jejich uplatnění a kvalitu života. Speciální menšinovou skupinu v každé národní společnosti pak tvoří neslyšící lidé. Primárním jazykem jejich komunikace je totiž znakový jazyk, který se vyjadřuje pomocí pohybu a postavení rukou, hlavy a horní části trupu v prostoru. Osvojení většinového mluveného jazyka je pak pro neslyšící velmi obtížné, neboť ke správnému osvojení je třeba sluch a navíc se mluvený a znakový jazyk diametrálně odlišují svojí povahou danou rozdílnou existencí (zvuk versus 3D obraz). Pro neslyšící děti neslyšících rodičů je např. majoritní jazyk až druhým jazykem, neboť jejich mateřským jazykem je právě jazyk znakový. Tyto obtíže s osvojením majoritního jazyka tak vedou k vytváření informačních a komunikačních bariér pro neslyšící. Jednou z cest, jak odstranit tyto bariéry, je využití tlumočníků, kteří mohou neslyšícím zprostředkovat kontakt se slyšícími a přístup k informacím v majoritním jazyce. Služby tlumočníků jsou však poměrně nákladné a hlavně, ne bezprostředně využitelné vždy, když je neslyšící potřebuje (tlumočníků je jen omezené množství a je třeba si jejich služby předem zajistit). Řešením těchto problémů by tak mohlo být využití automatického překladu z majoritního jazyka do znakového jazyka a opačně.

Už od počátku používání počítačů se lidé snaží o jejich využití při řešení různorodých problémů. V současné době jsme, díky rostoucí výpočetní síle počítačů a také jejich masivnímu rozšíření do všech oborů lidského konání, svědky jejich praktického použití pro dříve jen obtížně řešitelné problémy. Jedním z těchto problémů je také zpracování přirozeného jazyka, jehož konečným cílem je poskytnout člověku možnost komunikace s počítačem v přirozeném jazyce. Jde tedy o vytvoření takových nástrojů, které umožní uživatelům vést s počítačem dialog tak, jako by se jednalo o jiného člověka. Zpracování přirozeného jazyka a také stejnojmenný vědní obor (angl. natural language processing (NLP)) se tak primárně zabývá třemi hlavními problémy, které je pro využití komunikace a dialogu s počítačem v přirozeném jazyce potřeba řešit. První dva problémy souvisí se vstupem a výstupem v přirozeném jazyce a zabývají se rozpoznáváním mluvené řeči (vstup od uživatele) a syntézou mluvené řeči (výstup počítače). Třetím problém, jehož vyřešení je nejdůležitější, je pak porozumění dané promluvy v přirozeném jazyce. V případě rozpoznávání řeči je hlavním cílem převod akustického signálu na odpovídající text, s kterým lze dále pracovat. Úkolem syntézy je pak naopak převést daný text na zvukový signál, který by co nejlépe odpovídal tomu, jak by daný text přečetl a vyslovil člověk. V případě porozumění pak jde především o „pochopení“ toho, co bylo uživatelem řečeno a následné provedení očekávané nebo požadované akce. Tento způsob komunikace po-

mocí přirozeného jazyka by měl být přístupný všem uživatelům, tedy i těm hendikepovaným. V případě neslyšících uživatelů tak ještě do hry vstupuje použití automatického překladu tak, aby obsah počítače a informačních sítí, který je v majoritním jazyce, byl přístupný i pro tyto uživatele. K vedení plnohodnotného dialogu mezi počítačem a neslyšícím je tedy třeba také vyřešit problém syntézy a rozpoznávání znakového jazyka. V případě např. syntézy z textu se pak také uplatní systém automatického překladu mezi majoritním a znakovým jazykem. V případě komunikace mezi slyšícím a neslyšícím je třeba dále zajistit i opačný směr překladu, tj. ze znakového do majoritního jazyka. Další možností využití automatického systému pro překlad je také výuka majoritního jazyka. Kdy by použití automatického překladu mohlo vést k lepšímu osvojení a porozumění majoritního jazyka neslyšícími.

Tato dizertační práce je členěna následovně. V úvodní kapitole dále následuje popis jazyků a komunikačních systémů používaných neslyšícími v České republice a popis základních pojmů a přístupů z oblasti automatického překladu. Na konci úvodní kapitoly je pak přehled a popis existujících systémů pro automatický překlad různých světových znakových jazyků. V druhé kapitole je na základě poznatků z předešlé kapitoly, definován hlavní cíl dizertační práce a popsáno zvolené řešení pro jeho dosažení. Jsou diskutovány důvody pro volbu tohoto řešení a vytyčeny dílčí problémy, které je třeba v zájmu hlavního cíle vyřešit. Třetí kapitola tak popisuje tvorbu vlastního paralelního korpusu vhodného pro vytvoření automatického překladového systému pro češtinu a znakovanou řeč. Čtvrtá kapitola se zabývá výběrem frází do frázové tabulky, která tvoří jednu z klíčových součástí zvoleného řešení automatického překladu. V této kapitole jsou nejprve představeny stávající metody pro výběr frází a následně je podrobně popsána nová metoda pro výběr frází, která se snaží o eliminaci nedostatků stávajících metod. Je popsáno nové kritérium pro výběr frází a porovnáno s dalšími dvěma používanými kritérii. Pátá kapitola se zabývá konstrukcí dekodéru, který je klíčový z hlediska přesnosti a rychlosti vytvářeného překladového systému. Jde především o popis algoritmu prohledávání, který byl implementován při tvorbě vlastního dekodéru využitého pro překlad. V šesté kapitole jsou představena různá kritéria pro nastavení vah použitých v překladovém systému a je uveden optimalizační algoritmus používaný pro nalezení optimálních vah log-lineárního modelu, který je využíván jak při výběru frází tak i hledání nejlepšího překladu. Sedmá kapitola obsahuje výsledky všech provedených experimentů. Nejprve jsou představena kritéria použitá pro měření přesnosti vytvořeného překladu a je také popsáno nové automatické kritérium pro měření sémantické shody mezi referenčním a automaticky vytvořeným překladem. První část experimentů se pak zabývá nově navrženou metodou pro výběr frází popsanou ve čtvrté kapitole. Je vyzkoušeno základní nastavení metody a ověřena možná vylepšení navržená na konci čtvrté kapitoly. V následných experimentech jsou pak porovnány tři různé frázové tabulky. Jde o ručně vytvořenou frázovou tabulku, která vznikla při vytváření paralelního korpusu a dále pak dvou automaticky vytvořených frázových tabulek. Z nichž první je vytvořena standardní metodou pro výběr frází popsanou na začátku čtvrté kapitoly a druhá pak v předešlé části experimentů nalezenou nejlepší modifikací nové metody pro výběr frází navrženou v této práci. Po srovnání různých frázových tabulek následuje srovnání výkonnosti vlastního dekodéru s výkonností standardního dekodéru a také porovnání různých modifikací obou těchto dekodérů použitých v experimentech. Následně jsou představena a porovnána možná vylepšení základního vytvořeného překladového systému. Jsou uvedeny výsledky přesnosti překladu pro oba směry překladu. Nakonec jsou uvedeny výsledky experimentů při použití rysů založených na slovním přiřazení v nové metodě pro výběr frází. Výsledky pro všechny použité frázové tabulky a jejich modifikace jsou shrnuty do jedné tabulky a porovnány z hlediska statistické významnosti jejich rozdílů. Konečně poslední kapitola obsahuje shrnutí a zhodnocení dosažení vytyčených cílů a možné směry pro další zkoumání.

1.1 Znakovaná řeč

Podle zákona 155/1998 Sb., o komunikačních systémech neslyšících a hluchoslepých osob ve znění zákona 384/2008 Sb. § 3, se komunikačními systémy neslyšících a hluchoslepých osob rozumí český znakový jazyka a komunikační systémy vycházející z českého jazyka. Podle § 4 je český znakový jazyk (ČZJ) vymezen následovně:

- Český znakový jazyk je základním dorozumívacím jazykem těch neslyšících osob v České republice, které jej samy považují za hlavní formu své komunikace.
- Český znakový jazyk je přirozený a plnohodnotný komunikační systém tvořený specifickými vizuálně-pohybovými prostředky, tj. tvary rukou, jejich postavením a pohyby, mimikou, pozicemi hlavy a horní části trupu. Český znakový jazyk má základní atributy jazyka, tj. znakovost, systémovost, dvojí členění, produktivnost, svébytnost a historický rozměr, a je ustálen po stránce lexikální i gramatické.
- Český znakový jazyk může být využíván jako komunikační systém hluchoslepých osob v taktilní formě, která spočívá ve vnímání jeho výrazových prostředků prostřednictvím hmatu.

Důkazy o přirozenosti znakového jazyka lze nalézt v řadě prací, za všechny uveďme práci [Bímová 02], která shrnuje základní vlastnosti přirozeného jazyka definované F. de Saussurem a ukazuje, jak jsou splněny ve znakovém a tedy i českém znakovém jazyce. První prací, která se zabývala znakovým jazykem z lingvistického hlediska a dala podnět k jeho dalšímu zkoumání, byla práce Američana Williama C. Stokoeho: Sign Language Structure, která poprvé vyšla v roce 1960 [Bímová 02].
Podle § 6 o komunikačních systémech neslyšících a hluchoslepých osob vycházejících z českého jazyka pro znakovanou češtinu (ZČ) platí:

- Znakovaná čeština využívá gramatické prostředky češtiny, která je současně hlasitě nebo bezhlasně artikulována. Spolu s jednotlivými českými slovy jsou pohybem a postavením rukou ukazovány jednotlivé znaky, převzaté z českého znakového jazyka. Znakovaná čeština v taktilní formě může být využívána jako komunikační systém hluchoslepých osob, které ovládají český jazyk.

Pro prstovou abecedu platí:

- Prstová abeceda využívá formalizovaných a ustálených postavení prstů a dlaně jedné ruky nebo prstů a dlaní obou rukou k zobrazování jednotlivých písmen české abecedy. Prstová abeceda je využívána k odhláskování cizích slov, odborných termínů, případně dalších pojmů, pro které dosud nejsou ustáleny znaky českého znakového jazyka. Prstová abeceda v taktilní formě může být využívána jako komunikační systém hluchoslepých osob.

Podle § 7 mají neslyšící a hluchoslepé osoby právo na:

- Používání komunikačních systémů neslyšících a hluchoslepých osob.
- Vzdělávání s využitím komunikačních systémů neslyšících a hluchoslepých osob.

- Výuku komunikačních systémů neslyšících a hluchoslepých osob, kterou upravuje jiný právní předpis [1].

Za neslyšící se podle § 2 toho zákona považují osoby, které neslyší od narození, nebo ztratily sluch před rozvinutím mluvené řeči, nebo osoby s úplnou či praktickou hluchotou, které ztratily sluch po rozvinutí mluvené řeči, a osoby těžce nedoslýchavé, u nichž rozsah a charakter sluchového postižení neumožňuje plnohodnotně porozumět mluvené řeči sluchem. Cílem pojmu znakovaná řeč zmíněného v nadpisu této práce je pak v sobě sdružit pojmenování pro český znakový jazyk a znakovanou češtinu, jako jazyků, které ke svému vyjádření používají znaky a jsou tedy znakovány, obdobně jako se řeč využívající mluvení nazývá mluvená řeč. Toto pojmenování si však neklade za cíl být novým termínem v oblasti znakových jazyků, ale slouží hlavně pro potřeby této práce, jako souhrnný termín pro oba již zmíněné jazyky.

1.2 Český znakový jazyk

ČZJ je přirozený, vizuálně-motorický jazyk, který je primárním jazykem používaným při komunikaci mezi neslyšícími, tj. je samotnými neslyšícími preferován. Pro děti neslyšících rodičů je také jazykem mateřským a dochází k jeho osvojení stejně (zhruba ve stejných fázích) jako k osvojení češtiny u slyšících dětí [Bímová 02].

Základní jednotkou znakového jazyka je znak (zhruba odpovídá jednomu slovu (pojmu) v mluveném jazyce, to ale neplatí vždy, jak uvidíme dále). Znak má dvě složky: nemanuální a manuální. Nemanuální složka je vyjádřena mimikou, pohyby a pozicemi hlavy a horní části trupu (tzv. nemanuální nosiče). Manuální složka je vyjádřena tvary, pohyby a pozicemi rukou (tzv. manuální nosiče). Znaky se realizují ve znakovacím prostoru, který je zhruba vymezen rozpaženými lokty, temenem a linií vedenou pod žaludkem. Hlavní rozdíl mezi češtinou a ČZJ je dán tím, že ČZJ je vizuálně-motorický jazyk, tj. tento jazyk není vnímán sluchem ale zrakem a je založen na tvarech, pozicích a pohybu a ne na zvuku. Z toho pramení dvě základní odlišnosti znakového jazyka: simultánnost a existence v trojdimenzionálním prostoru.

1.2.1 Simultánnost a trojdimenzionální prostor

Simultánnost umožňuje některé výrazy produkovat a vnímat současně v jednom okamžiku. To je dáno tím, že některé významy jsou neseny manuálními a některé nemanuálními nosiči, které mohou být produkovány a vnímány současně (např. příslovečné určení způsobu se vyjadřuje většinou tímto způsobem). Simultánně existují i subkomponenty manuální složky znaků, tj. místo artikulace, tvar ruky, orientace dlaně a prstů a pohyb můžeme také produkovat a vnímat současně (např. spojení: „vrátím ti“ lze vyjádřit jedním znakem „JÁ–VRÁTÍM–TI“, které se od „TY–VRÁTÍŠ–MI“ liší pouze počátečním a koncovým místem artikulace a směrem horizontálního pohybu, u slovesa běžet lze jiným tvarem ruky vyjádřit, že „ČLOVĚK–BĚŽÍ“ nebo „KOČKA–BĚŽÍ“). A nakonec lze simultánně také produkovat dva jednoruční znaky (např. ve složeninách), každou rukou jeden.

Trojrozměrný prostor je základnou pro celou řadu gramatických struktur znakového jazyka. V prostoru jsou rozmístěny subjekty komunikace, ať už subjekty přímých účastníků komunikace, nebo subjekty, které jsou předměty sdělování (jejich pozice v prostoru je přitom do značné míry gramatikalizována). Prostor je základnou textové soudržnosti (polem pro odkazování), v opoře o prostor se vyjadřuje řada gramatických kategorií (výrazně např. číslo a

[1] § 16 odst. 7 zákona č. 561/2004 Sb., o předškolním, základním, středním, vyšším odborném a jiném vzdělávání (školský zákon), ve znění zákona č. 384/2008 Sb.

čas), v prostoru (a „přímo“) se vyjadřují časoprostorové vztahy (v češtině vyjadřované např. předložkovými vazbami), v prostoru se ohýbají slovesa, prostor (pohyb v něm) slouží pro vyjadřování věcně obsahových vztahů mezi výpověďmi (např. vztah podmínky). Prostor je také polem pro postupy spojené s výstavbou textu (např. pro tzv. změnu rolí typickou pro vypravování) [Macurová 01a].

1.2.2 Gramatika ČZJ

Od počátků ústavní péče o neslyšící existovaly dva přístupy k jejich vzdělání. První přístup se nazývá francouzská škola (způsob výuky) a je založen na využití znakového jazyka v procesu výuky neslyšících. Jejím zakladatelem byl Francouz Abbé Charles Michel de l'Epée (1712 až 1789), který byl průkopníkem používaní znakového jazyka ve výuce neslyšících. V 18. století se vytvořil i druhý přístup, nazývaný německá škola, jehož zakladatelem byl Němec Samuel Hainicke (1727 až 1790). Jeho metoda je přísně orální. Neslyšící by se měl nejprve naučit mluvit a teprve poté číst a psát (Hainicke totiž zastával názor, že neslyšícím se nesmí nic ulehčit (čtení a psaní), jinak by ztratili zájem o mluvení). Takovým ulehčením bylo podle něj i používání znakového jazyka. Hainicke se tedy snažil o jeho vyloučení z výuky neslyšících a tím rozpoutal spor mezi oralisty a zastánci znakového jazyka ve vzdělávání neslyšících. Tento spor vyvrcholil roku 1880, kdy, na kongresu učitelů neslyšících v Miláně, byla přijata rezoluce, která vyloučila znakový jazyk z výuky neslyšících a označila za jediný přijatelný ryze orální přístup k výuce. V Čechách se k výuce od počátku používal znakový jazyk (Pražský ústav pro hluchoněmé – 1786, 1858 Litoměřice, 1871 České Budějovice, 1881 Hradec Králové, 1913 Plzeň) na Moravě pak orální metoda (1829 Brno, 1844 Mikulov, 1894 Ivančice a Lipník, 1907 Šumperk, 1911 Valašské Meziříčí). V roce 1915 vznikl Zemský spolek pro péči o hluchoněmé, který v roce 1923 prosadil odstranění znakového jazyka z procesu vzdělávání neslyšících u nás. Přes opakované snahy o rehabilitaci znakového jazyka vydržel tento stav až do roku 1989. Výzkum ČZJ a jeho gramatiky u nás tedy začal až po roce 1989, hlavně pak poté, kdy byl v roce 1998 přijat zákon o znakové řeči, který uzákonil právo neslyšících na používání, vzdělaní a výuku ve znakovém jazyce [Hrubý 03].

Výzkum ČZJ vychází a navazuje na výzkum znakových jazyků ve světě, který započal v šedesátých letech minulého století již zmíněnou prací Sign Language Structure Williama C. Stokoeho. Jednou z důležitých věcí, které Stokoe v této práci ukázal a která odstartovala zájem o lingvistický výzkum znakových jazyků, bylo odlišení gesta a znaku. Stokoe ukázal, že na rozdíl od gesta lze znak analyzovat na menší jednotky - komponenty znaku. V jeho koncepci se každý znak skládá ze tří simultánně vystupujících komponentů: umístění v prostoru (TAB), tvaru ruky (DEZ) a pohybu ruky v prostoru (SIG). V každém konkrétním znakovém jazyce jsou tyto komponenty realizovány kvalitativně vymezeným a kvantitativně omezeným souborem prvků, který je srovnatelný se souborem fonémů v mluvených jazycích [Bímová 02]. První výzkumy ČZJ se tedy soustředily na určení jednotlivých komponent znaku (TAB, DEZ a SIG) a vytvoření notace pro zápis jednotlivých znaků (do zápisu znaků ČZJ je zahrnuto: místo, kde se znak artikuluje (TAB); tvar ruky/rukou, která/é artikuluje/í (DEZ); vztah ruky/rukou k tělu: orientace dlaně (ORI 1) a orientace prstů (ORI 2); vztah ruky k ruce: vzájemná poloha rukou (u znaků artikulovaných dvěma rukama) (HA); pohyb/y ruky/rukou (SIG)) [Macurová 96].

1.2.3 Slovesa a jejich typy

Další práce se již zabývali některými gramatickými jevy ČZJ. Práce [Macurová 01b] tak shrnuje dosavadní poznatky o osobních zájmenech (jsou důležitá pro fungování sloves) a slovesech v ČZJ. Tak např. ČZJ nerozlišuje u zájmena třetí osoby rod; ČZJ má více osobních zájmen

než čeština (vedle tvarů pro množné číslo jsou tu tvary i pro číslo dvojné, trojné a čtverné); zájmena ČZJ poskytují navíc informaci o tom, kde je v prostoru situován referent, k němuž zájmena odkazují; jednoznačněji než v češtině jsou vyjádřeny různé druhy plurálu (např. je vždy explicitně vyjádřeno zda jde o plurál inkluzivní („my“ odkazuje k „já + ty“ a zahrnuje tedy do „my“ adresáta) nebo exkluzivní („já + on, resp. ona nebo oni“) a adresáta vylučuje).

Zatímco v české větě se, aby vyjádřila gramatické významy osoby, čísla, času, způsobu, rodu a vidu, proměňují všechna predikátová slovesa, v ČZJ je situace jiná. Pro vyjadřování osoby (a čísla) se proměňují jen některá slovesa - a proměňují se zcela jinak než slovesa v češtině. Vedle nich pak existují i slovesa, která se nemění a musí osobu (a číslo) vyjadřovat prostřednictvím osobního zájmena a jeho užití je tak pro tento typ sloves „povinné“ (podobná situace je např. i v angličtině, kdy např. z pouhého tvaru „go“ není zřejmá osoba, ani číslo - tyto informace nese až „povinně“ užívané zájmeno: „I go“ vs. „they go“). Další slovesné významy (čas, způsob atd.) se pak v českém znakovém jazyce nevyjadřují gramaticky (tj. změnou slovesného tvaru), ale je zapotřebí vyjádřit je lexikálně (přidáním slova s žádoucím lexikálním významem, např. jako u gramatického významu času), popř. lexikálně-gramaticky (jako např. u vidu, resp. aspektu) [Macurová 01b].

Kontextově zapojená slovesa českého znakového jazyka však ve srovnání s češtinou nesou i některé informace „navíc“, takové, které sloveso v češtině nést nemůže: Změnou tvaru artikulující ruky do sebe včleňují (inkorporují) informace o předmětu, který je slovesným dějem zasažen, na který děj přechází nebo kterého se týká: tak např. sloveso „hodit“ nabývá různých podob pro „HODIT–KÁMEN“, „HODIT–OŠTĚP“, „HODIT–ZRNKO–HRÁŠKU“, „HODIT–TVÁRNICI“ atd., „JÍST–PIZZU“ je jiné než „JÍST–JABLKO“, liší se „KOUŘIT–DOUTNÍK“ a „KOUŘIT–CIGARETU“ atd. [Macurová 01b].

Z gramatického hlediska lze slovesa v ČZJ rozdělit do tří skupin. První skupinu tvoří slovesa prostá. To jsou slovesa, která existují pouze v citátovém (slovníkovém) tvaru a v kontextu se tak pro vyjadřování osoby (a čísla) neproměňují. Jejich argumenty je třeba vyjádřit lexikálně a sekvenčně, jménem nebo zájmenem (např. „MÍT RÁD“ ve větě „mám tě ráda“ vyžaduje sled znaků „index–JÁ + MÍT RÁD + index–TY“; „máš mě rád?“ zase sled „index–TY + MÍT RÁD + index–JÁ + nemanuální nosič významu „otázka zjišťovací“) [Macurová 01b].

Druhou skupinu tvoří slovesa shodová (dříve směrová). Jde o slovesa, která maximálně využívají gramatikalizovaného prostoru: své argumenty vyjadřují tato slovesa (v různé míře a různým způsobem) změnou svého tvaru. Vzhledem k tomu, že gramatické významy osoby a čísla subjektu a objektu vyjadřují tato slovesa sama, tak užití jména nebo zájmena spolu se slovesem není „povinné“. Místo toho je, pro vyjádření gramatických významů osoby (a čísla), využita inkorporace místa artikulace a změna směru pohybu v horizontální linii, příp. také ještě změna orientace dlaně. V českém znakovém jazyce se tak např. „JÁ–VRÁTÍM–TI“ liší od „TY–VRÁTÍŠ–MI“ pouze počátečním a koncovým místem artikulace a směrem (horizontálního) pohybu [Macurová 01b].

Ve třetí skupině jsou slovesa prostorová (v odborné literatuře vedle toho figurují ještě termíny slovesa lokalizační, pohybově-lokalizační nebo klasifikátorová). Tato slovesa význam osoby (a čísla) sama o sobě nevyjadřují (jejich argumenty je třeba vyjadřovat jménem nebo zájmenem). Ve větě, v kontextu se nicméně proměňují: podávají informace o umístění (lokalizaci) děje nebo stavu, o pohybu objektu a často vyjadřují (inkorporací změněného tvaru ruky) prostředek. V českém znakovém jazyce je např. jinými tvary ruky artikulováno „HOLIT–SE–BŘITVOU“, „HOLIT–SE–STROJKEM“, „HOLIT–SE–ŽILETKOU“; na jiné lokalizaci než základní „holit se“ je umístěno třeba „HOLIT–SE–V–PODPAŽÍ“ nebo „HOLIT-SI-NOHY“. Změnou lokalizace (místa artikulace) je možné vyjádřit různé objekty lexému „kousnout“ (do nosu, do nohy, do ucha) apod. V této skupině sloves je tak patrný jistý izomorfizmus s reálným

dějem, resp. stavem – s jeho lokalizací, jeho průběhem, jeho směrem (pohyb není na rozdíl od sloves shodových omezen jen na horizontální linii!), jeho začátkem a koncem (ty nejsou omezeny jen na gramatikalizované pozice!) apod. Slovesa tohoto typu jsou tedy nejméně stabilní (mohou být artikulována na různých místech, různými tvary ruky a s různým pohybem) a v maximální možné míře využívají charakteristické rysy znakového jazyka – jeho existenci v prostoru a jeho simultánnost [Macurová 01b].

1.2.4 Tvorba tázacích vět

Práce [Hronová 02] podává přehled o tázacích větách a jejich tvorbě ve znakových jazycích ve světě a v ČZJ. Ve většině jazyků lze rozlišit, podle toho na co se ptáme, dva základní typy otázek: doplňovací a zjišťovací. V případě otázky zjišťovací se ptáme obecně (s elementární odpovědí „Ano“, „Ne“), otázkou doplňovací se pak ptáme specificky (s elementární odpovědí na dotazy typu „Kdo? Co? Kde? Kdy? Který?“ apod.) a doplňujeme si neúplnou znalost věci či informace o světě. Stejné rozlišení otázek lze nalézt i v ČZJ. Mezi prostředky používané ve znakových jazycích pro vyjádření otázky patří: nemanuální složka znakového jazyka a výskyt a umístění tázacích výrazů.

Roli intonace mluvených jazyků zastupují v jazycích znakových výrazné mimické rysy obličeje (nemanuální složka). Specifický výraz obličeje vzniká simultánně s artikulací jednotlivých znaků a slouží k odlišení typu otázky. U otázek zjišťovacích se objevuje nápadně zvednuté obočí, u otázek doplňovacích pak svraštění obočí. Toto rozdělení však není absolutně platné. Klade-li např. mluvčí (označuje osobu používající buď mluvený, nebo znakový jazyk) doplňovací otázku proto, aby si již získanou informaci ověřil, pozorujeme namísto svraštění obočí jeho zvednutí. Toto zjištění vedlo k závěru, že pohyb obočí není nositelem funkce gramatické, nýbrž pragmatické (upozorňuje adresáta, že jde o otázku a že mluvčí očekává odpověď). Kromě nápadného pohybu obočí se při stavbě tázacích vět uplatňuje i oční kontakt (z výzkumu norského znakového jazyka vyplynulo, že u zjišťovacích a doplňovacích otázek se „plný“ oční kontakt vyskytl u všech respondentů, u reprodukovaných otázek (např. „Zeptala se ho, zda má jít k lékaři s ním.“) se oční kontakt neobjevil u žádného z respondentů). Nemanuální složka se u otázek neomezuje jen na obočí, ale promítá se i na další části těla, např. u otázek Ano/Ne mluvčí stojí tak, aby byl nakloněný rameny i hlavou směrem dopředu a bradu nepatrně zdvihne nahoru, aby měl obličej ve vertikální poloze. U doplňovací otázky pak mluvčí nahrbí svá ramena směrem dopředu. Nemanuální složky znakového jazyka tvoří nutnou část každé tázací výpovědi; bez nich by byla výpověď agramatická a stěží srozumitelná [Hronová 02].

Z výzkumu znakových jazyků vyplynulo, že obecně se tázací zájmena mohou vyskytovat v různých pozicích (na začátku, na konci i uprostřed) ve větě, dokonce mohou být v jedné otázce několikrát opakovány. Znakosled otázek tak není fixován jako v případě např. neflektivních jazyků (např. angličtina). Význam otázky se tedy přeskupením znaků nezmění. Podrobným výzkumem několika variant jedné otázky (odlišný znakosled) bylo zjištěno, že neslyšící dávají přednost té variantě otázky, která je z pohledu mluvčího ekonomičtější (v případě, že pro tázací zájmeno v doplňovací otázce existuje více pozic, mluvčí se nakonec vždy řídí artikulací následujícího znaku, tj. vybere pro tázací zájmeno ve větě takovou polohu, která při přechodu na další znak, vyžaduje minimální fonologickou změnu). Tento výběr zajišťuje plynulý přechod z jednoho znaku do druhého. Ve výpovědích neslyšících mluvčích se uplatňují tendence dávat přednost souvislému, plynulému pohybu v jediném směru (podle těchto kriterií hodnotí gramatickou správnost tázacích vět i sami neslyšící). Výzkumem provedeným na materiálu získaném elicitační metodou (tj. data získaná přímo od neslyšících) bylo zjištěno, že umístění tázacích výrazů s funkcí otázky ve výpovědích v ČZJ podléhá podobným pravidlům jako ve znakových

jazycích ve světě. Konkrétně se tázací zájmeno může vyskytovat: na začátku („KDO SPOLU SPOKOJENÝ“ – „S kým jsi byl spokojený?“), na konci („BÁT CO“ — „Čeho se bojíš?“), zhruba uprostřed („TY UDĚLAT JAKÝ DŮVOD“ — „Z jakého důvodu jsi to udělal?“), na začátku i na konci (tzv. „echo“ „KDO LÍBIT + (znak pro vyjádření minulosti) KDO“ — „Komu se to líbilo?“) [Hronová 02].

1.2.5 Vyjadřování času

Práce [Macurová 03] se zabývá otázkou vyjádření času ve znakových jazycích ve světě a na základě toho pak řeší tuto otázku v ČZJ. Čas lze reprezentovat osou, na níž minulost konvenčně „vidíme“ vlevo, budoucnost vpravo. Současnost je reprezentována bodem označovaným obvykle jako nula: - - - - minulost - - - - 0 - - - - budoucnost - - - -. K této časové ose se pak usouvztažňují události, děje, situace, stavy atd., které představují obsah výpovědi. Základním referenčním bodem, vzhledem k němuž se události, děje, situace, stavy apod. umisťují na časovou osu, bývá obvykle současnost, přítomnost. Existují dvě základní možnosti, jak čas vyjádřit: gramaticky nebo lexikálně [Macurová 03].

V případě gramatického vyjádření se čas děje, události, stavu atd. „vyznačuje“ přímo na slovese, specializovaným slovesným tvarem. Tak je tomu např. i v češtině: jeden tvar je pro vyjádření minulosti („pracoval jsem“), jiný pro vyjádření toho co je („pracuji“) a ještě jiný pro to co bude („budu pracovat“) [Macurová 03].

V případě lexikálního vyjádření času se užívají lexikální prostředky tzv. časového poukazu. Jsou to prostředky, které usouvztažňují označovaný čas s okamžikem promluvy. Některé z nich situují událost „před“ čas aktuální promluvy (např. české výrazy jako „v r. 1980, včera, loni, před rokem“ apod.), jiné „po“ tomto čase (např. „příští týden, potom, napřesrok, za 5 let“ apod.). Takovéto lexikální prostředky jsou k dispozici jak v češtině, tak i v ČZJ [Macurová 03].

V jazycích, kde je „vyznačení“ času na slovese závazné, je vyjadřování času záležitostí gramatického systému jazyka, jeho morfologie, je gramatikalizováno – v těchto jazycích tedy existuje tzv. gramatická kategorie času. V jazycích, kde se čas vyjadřuje lexikálně (časovými určeními) a slovesa se pro vyjádření času neproměňují, o gramatické kategorii času mluvit nelze. Mezi takové jazyky patří znakové jazyky i ČZJ. Znakové jazyky tedy nemají gramatickou kategorii času, čas vyjadřují lexikálně. Minulost nebo budoucnost se vyjadřuje spojením lexikálního signálu („MIN“ pro minulost, „BUD“ pro budoucnost) se slovesem (výraz „index–JÁ MIN PRACOVAT“ odpovídá českému: „pracoval/a jsem“; výraz „index–JÁ BUD PRACOVAT“ odpovídá českému: „budu pracovat“). Lexikální signály minulosti a budoucnosti se ale zdaleka nepojí s každým slovesem. Lexikálním signálem času se vytvoří určitý časový rámec, k němuž se vztahují všechna slovesa až do okamžiku, kdy je vytvořen, opět lexikálním signálem času, rámec nový, odlišný od prvního. Přítomný čas přitom nemusí být žádným signálem vyznačován. Pokud jde o postavení lexikálních signálů ve větě, převažuje postavení před slovesem. Objevují se ale i případy, kdy signály sloveso následují, i případy, kdy je signál zdvojen (tj. stojí před i za slovesem). Podmínky, které různé postavení vzhledem ke slovesu určují, není zatím možné jednoznačně určit [Macurová 03].

1.2.6 Specifické znaky

Další práce [Motejzíková 03] a [Vysuček 04] se zabývají různými typy specifických znaků a jejich významem. Specifické znaky jsou znaky, pro které neexistuje v českém jazyce jednoslovný ekvivalent a jejich význam (resp. překlad do češtiny) je silně závislý na konkrétní situaci a kontextu v němž jsou použity (např. speciální znak (SZ), který znamená vyjádření nechuti k vykonání

nějaké činnosti, lze ve větě: „PRÁCE SZ“ přeložit jako: „Do práce se mi fakt nechce“, ve větě: „ÚŘAD ZAŘIZOVAT OBČANSKÝ-PRŮKAZ SZ“ pak jako: „Když si představím, kolik bude kolem té občanky běhání po úřadech, tak bych se na to nejradši vykašlal.“). Tyto znaky tvoří v rámci znakového jazyka zvláštní skupinu znaků, které se vyznačují tím, že v sobě velmi často nesou nějakou pozitivní či negativní emoci, vyjadřují různé stavy lidské psychiky nebo osobní postoj mluvčího k někomu anebo něčemu a jejich hodnocení [Motejzíková 03].

1.3 Automatický překlad

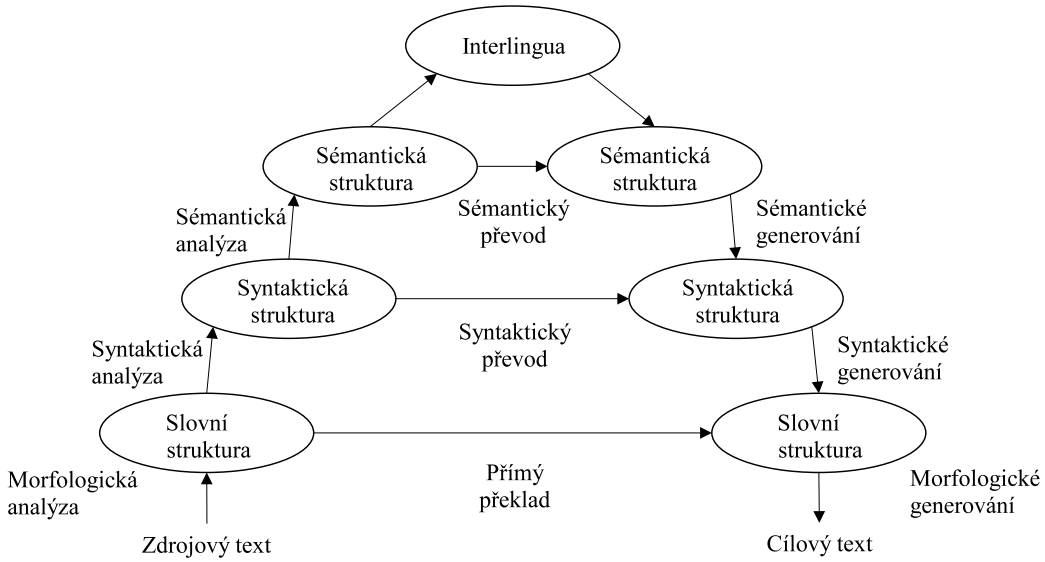
Systémy pro automatický překlad (angl. machine translation (MT)) jsou jednou z disciplín NLP. Tak jako v NLP existují dva základní směry zkoumání přirozeného jazyka – a sice:

- *racionální* - vycházející z předpokladu, že hlavní část lingvistických znalostí je uložena v mozku již při narození a není tedy získána pomocí učení,
- *empirický* - vychází z toho, že při narození nejsou v mozku obsaženy konkrétní lingvistické struktury, ale spíše, že mozek má k dispozici nástroje pro obecnou asociaci, rozpoznávání vzorů a zobecňování; detailní struktury jazyka jsou pak získány aplikací těchto nástrojů na bohatý smyslový vstup,

lze i současné systémy pro automatický překlad rozdělit na dvě skupiny, které odráží oba směry zkoumání přirozeného jazyka.

První skupina, založená na racionálním směru zkoumání, bývá obvykle v literatuře nazývána jako systémy založené na pravidlech (angl. rule-based systems). Tvorba takového systému pak spočívá v návrhu vhodných struktur pro reprezentaci zdrojového a cílového jazyka a v tvorbě pravidel, která slouží k převodu těchto struktur mezi sebou. Tyto struktury a pravidla jsou vytvářeny obvykle lidskými experty (lingvisty), kteří jsou tak hlavním zdrojem znalostí pro vytvářený systém. Podle složitosti použitých struktur lze současné pravidlové systémy rozdělit do tří tříd: systémy přímého překladu (Direct), systémy založené na převodu (Transfer) a nakonec systémy využívající převod do interlingui (mezijazyka) (reprezentace zdrojového textu v interlingui je nezávislá na zdrojovém jazyce, tj. tato reprezentace je shodná pro oba překládané jazyky). Tyto třídy tvoří hierarchické uspořádání, které může být vyjádřeno pomocí tzv. Vauquisova pyramidového diagramu (viz Obrázek 1.1) [Dorr 98].

Jednotlivé vrstvy pyramidy jsou uspořádány podle hloubky analýzy prováděné při překladu mezi jazyky. Na spodku pyramidy je nejjednodušší postup překladu — přímý překlad, který je založen na nejprimitivnější formě převodu slova na slovo (přímý převod mezi povrchovými vrstvami jazyků). V dalších vrstvách pak postupně roste hloubka použité analýzy, překlad je nejprve založen na povrchové (syntaxe) analýze potom na hloubkové (sémantika). Text ve zdrojovém jazyce je vždy nejdříve převeden do příslušné struktury (syntaktické nebo sémantické), na tuto strukturu jsou pak aplikována pravidla, která ji převedou na odpovídající strukturu v cílovém jazyce. Nakonec je na tuto strukturu aplikován modul pravidel, který vytvoří povrchovou formu cílového jazyka odpovídající této struktuře. Na vrcholu pyramidy je pak nejkomplexnější přístup založený na převodu obou jazyků do interlingui, tři funkce potřebné v předešlých architekturách (funkce pro analýzu, převod a vytvoření povrchové formy) jsou nahrazeny dvojicí funkcí, z nichž jedna převede zdrojový text do interlingui a druhá pak z reprezentace v interlingui vytvoří odpovídající povrchovou formu cílového jazyka (odpadá funkce pro převod mezi odpovídajícími si strukturami). Vyšší vrstvy pyramidy přináší pokrytí většího množství rozdílů, které se objevují v obou jazycích, s tím ovšem přichází i větší nároky na množství doménově závislé práce potřebné při vývoji systému.



Obrázek 1.1: Rozdělení pravidlových systémů pro automatický překlad [Dorr 98].

Druhá skupina založená na empirickém směru zkoumání pak bývá obvykle v literatuře nazývána jako systémy založené na datech (angl. data-driven systems). Hlavním zdrojem znalostí pro tyto systémy jsou jedno a dvojjazyčné textové korpusy, o těch pak hovoříme jako o trénovacích datech. Tvorba takového systému pak spočívá v zajištění potřebných dat a návrhu a použití metod pro extrakci vhodných znalostí z těchto dat. Systémy v této skupině lze ještě dále rozdělit na systémy založené na příkladech (angl. example-based (EB) nebo také memory-based systems) a systémy založené na statistickém přístupu (angl. statistical systems). Systémy založené na příkladech využívají při překladu analogií získaných z trénovacích dat. Systémy založené na statistickém přístupu pak definují problém automatického překladu jako problém rozhodnutí, kdy je třeba rozhodnout, který z vytvořitelných překladů zdrojové věty je ten nejlepší. V této práci budeme vycházet především ze statistického přístupu ke konstrukci systému pro automatický překlad. V následující podkapitole tedy blíže popíšeme tyto systémy.

1.4 Statistický přístup k automatickému překladu

Jako první navrhl už v roce 1955 použití statistických metod a myšlenek z oblasti teorie informace v automatickém překlad W. Weawer ve své práci „Machine Translation of Languages: fourteen essays“. Navzdory prvním úspěchům se problém ukázal být obtížnější, než se čekalo. To dokládá např. zpráva výboru ALPAC[2] z roku 1966, která konstatovala, že deset let výzkumu na tomto poli nepřineslo očekávané výsledky. Prostředky do této oblasti výzkumu tak byly radikálně omezeny. Díky rostoucí výpočetní síle počítačů a dostupnosti paralelních korpusů došlo však na konci osmdesátých let k obnovení výzkumu v této oblasti [Zens 08]. O to se zasloužily především práce vědců z firmy IBM [Brown 88, Brown 90, Brown 93]. Toto obnovení úzce souvisí také s oživením empirického směru zkoumání přirozeného jazyka, ke kterému došlo právě na konci osmdesátých let.

[2] http://en.wikipedia.org/wiki/ALPAC

1.4.1 Model zdrojového kanálu

Práce [Brown 88, Brown 90, Brown 93] definují základní přístup ke statistickému automatickému překladu (angl. statistical machine translation (SMT)). Předpokládejme, že máme větu $\mathbf{s}$ ve zdrojovém jazyce, tu můžeme přeložit na větu $\mathbf{t}$ v cílovém jazyce mnoha různými způsoby. V případě statistického překladu předpokládáme, že každá cílová věta $\mathbf{t}$ je možný překlad dané zdrojové věty $\mathbf{s}$. Každému páru $(\mathbf{s}, \mathbf{t})$ přiřadíme číslo $Pr(\mathbf{s}|\mathbf{t})$, které vyjadřuje pravděpodobnost toho, že překladatel přeloží $\mathbf{t}$ jako $\mathbf{s}$. Dále také předpokládáme, že když rodilý mluvčí tvořil větu $\mathbf{s}$, měl zároveň v mysli větu $\mathbf{t}$, kterou tak přeložil na $\mathbf{s}$. Úkolem našeho překladového sytému je tedy k dané větě $\mathbf{s}$ nalézt větu $\mathbf{t}$, kterou měl rodilý mluvčí v mysli, když vytvářel větu $\mathbf{s}$. Abychom minimalizovali možnost chyby, vybereme ze všech možných vět $\mathbf{t}$ takovou větu $\hat{\mathbf{t}}$, která maximalizuje pravděpodobnost $Pr(\mathbf{t}|\mathbf{s})$. Použitím Bayesova teorému dostaneme [Brown 93]:

$$Pr(\mathbf{t}|\mathbf{s}) = \frac{Pr(\mathbf{s}|\mathbf{t}) \cdot Pr(\mathbf{t})}{Pr(\mathbf{s})}. \tag{1.1}$$

Protože jmenovatel je nezávislý na $\mathbf{t}$, dostaneme pro nalezení $\hat{\mathbf{t}}$ následující vzorec:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}}\, Pr(\mathbf{s}|\mathbf{t}) \cdot Pr(\mathbf{t}). \tag{1.2}$$

Dostali jsme tak základní rovnici automatického překladu. Tato rovnice je známa i z jiných oblastí automatického zpracování přirozeného jazyka, např. rozpoznávání řeči, rozpoznávání naskenovaného textu, značkování textu, atd. Tato rovnice se nazývá *model zdrojového kanálu* (angl. source-channel model). Pravděpodobnost $Pr(\mathbf{s}|\mathbf{t})$ nazýváme *překladový model*, tento model popisuje, jak spolu souvisí jednotlivé dvojice zdrojových a cílových vět. Pravděpodobnost $Pr(\mathbf{t})$ nazýváme *jazykový model*, tento model vyjadřuje správnost přeložených vět $\mathbf{t}$ z hlediska cílového jazyka. Problém modelování pravděpodobnosti $Pr(\mathbf{t}|\mathbf{s})$ jsme tak rozdělili na dva nezávislé modely, které mohou být vytvářeny nezávisle na sobě. To je v případě automatického překladu přínosné, neboť, zatímco pro trénování překladového modelu potřebujeme paralelní texty, které je většinou obtížné a nákladné získat ve větším množství, pro trénování jazykového modelu lze použít libovolné texty v cílovém jazyce. Alternativní možností je pak modelovat pravděpodobnost $Pr(\mathbf{t}|\mathbf{s})$ přímo.

1.4.2 Log-lineární model

Použití log-lineárního modelu (nebo také modelu maximální entropie, angl. maximum entropy model) pro automatický překlad bylo poprvé navrženo v pracích [Papineni 98, Och 02b]. Pravděpodobnost $Pr(\mathbf{t}|\mathbf{s})$ je dána takto:

$$Pr(\mathbf{t}|\mathbf{s}) = p_{\lambda_1^M}(\mathbf{t}|\mathbf{s}) \tag{1.3}$$

$$p_{\lambda_1^M}(\mathbf{t}|\mathbf{s}) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}, \mathbf{s}))}{\sum_{\mathbf{t}'} \exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}', \mathbf{s}))}. \tag{1.4}$$

Zde máme modely (nebo také *rysy*, angl. features) $h_m(\mathbf{t}, \mathbf{s})$, které modelují vztah mezi zdrojovým a cílovým jazykem, a váhy λ_m těchto modelů. Opět jako v případě zdrojového kanálu můžeme při hledání nejlepšího překladu $\hat{\mathbf{t}}$ zanedbat jmenovatele v předešlém výraze a dosta-

neme lineární kombinaci individuálních modelů $h(\cdot,\cdot)$:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}}\{Pr(\mathbf{t}|\mathbf{s})\} \tag{1.5}$$

$$= \underset{\mathbf{t}}{\operatorname{argmax}} \left\{ \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t},\mathbf{s}))}{\sum_{\mathbf{t}'} \exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}',\mathbf{s}))} \right\} \tag{1.6}$$

$$= \underset{\mathbf{t}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{t},\mathbf{s}) \right\}. \tag{1.7}$$

Tento přístup je zobecněním modelu zdrojového kanálu [Zens 08]. Jestliže totiž použijeme jako individuální modely $h(\cdot,\cdot)$ logaritmus pravděpodobnosti $Pr(\mathbf{s}|\mathbf{t})$ a $Pr(\mathbf{t})$, tj.

$$h_1(\mathbf{t},\mathbf{s}) = \log Pr(\mathbf{s},\mathbf{t}) \tag{1.8}$$

$$h_2(\mathbf{t}) = \log Pr(\mathbf{t}), \tag{1.9}$$

kde $\lambda_1 = \lambda_2 = 1$, dostaneme model zdrojového kanálu [Och 02b]. Výhodou tohoto přístupu je pak snadná integrace dalších modelů $h(\cdot,\cdot)$ do výsledného systému. Váhy λ_1^M modelu můžeme trénovat podle kritéria maximální třídní aposteriorní pravděpodobnosti (nebo také kritéria maximální vzájemné informace, angl. maximal mutual information (MMI) criterion) pomocí GIS (angl. generalized iterative scaling) algoritmu, pro váhy pak platí [Och 02b]:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{n=1}^{N} \log p_{\lambda_1^M}(\mathbf{t}_n|\mathbf{s}_n) \right\}, \tag{1.10}$$

kde N je počet dvojic v trénovacím korpusu. Nebo je můžeme trénovat s ohledem na kvalitu výsledného překladu měřenou nějakým chybovým kritériem, tento přístup se nazývá trénování minimální chyby (angl. minimum error rate (MER) training (MERT)) a váhy pak určíme podle následujícího vztahu [Och 03a]:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} E(r_s, \hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M)) \right\} \tag{1.11}$$

$$= \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} \sum_{k=1}^{K} E(\mathbf{r}_s, \mathbf{t}_{s,k}) \delta(\hat{\mathbf{t}}(s_s, \lambda_1^M), \mathbf{t}_{s,k}) \right\} \tag{1.12}$$

$$\hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M) = \underset{\mathbf{t}\in C_s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{t},\mathbf{s}_s) \right\} \tag{1.13}$$

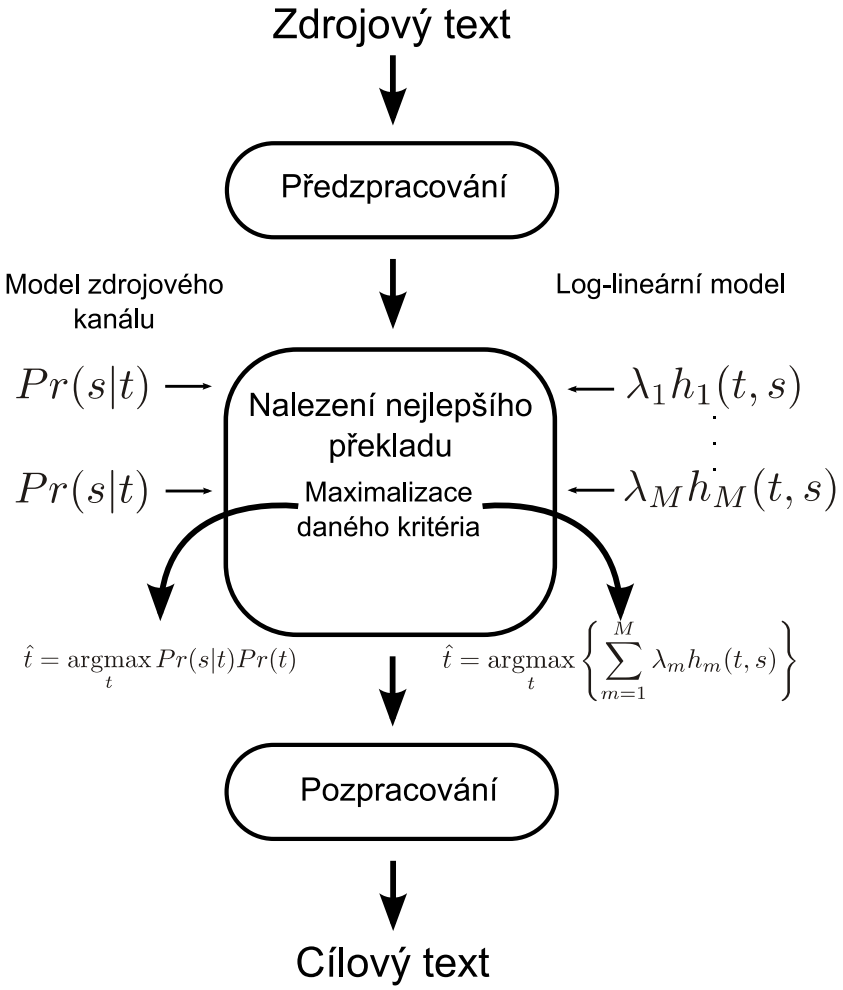
$$\delta(\hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M), \mathbf{t}_{s,k}) = \begin{cases} 1 & \text{jestliže } \hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M) = \mathbf{t}_{s,k} \\ 0 & \text{jinak,} \end{cases}$$

kde S je počet dvojic v množině dat určených k optimalizaci vah, E je chybové kritérium, které minimalizujeme, $\mathbf{r}_s$ je referenční překlad zdrojové věty $\mathbf{s}_s$ a $C_s = \{\mathbf{t}_{s,1}, ..., \mathbf{t}_{s,K}\}$ je množina K různých překladů každé zdrojové věty $\mathbf{s}_s$. Oba výše zmíněné přístupy k automatickému překladu jsou shrnuty na Obrázku 1.2.

1.5 Slovní a frázový překlad

1.5.1 Slovní překlad

Základní přístup ke statistickému automatickému překladu (viz práce [Brown 93]) předpokládá, že máme zdrojovou větu $\mathbf{s} = s_1^J = s_1...s_j...s_J$ a její překlad $\mathbf{t} = t_1^I = t_1...t_i...t_I$. K



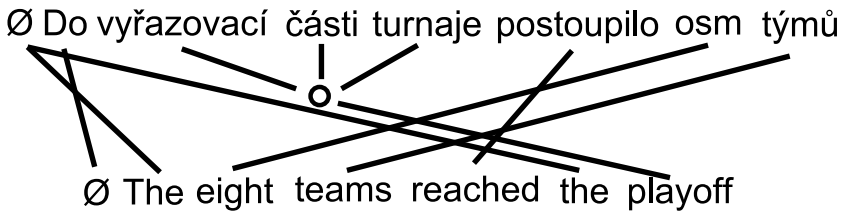
Obrázek 1.2: Schéma systému pro statistický automatický překlad.

tomu abychom pomocí rovnice 1.2 našli nejlepší překlad $\hat{\mathbf{t}}$ věty $\mathbf{s}$ potřebujeme znát *překladové pravděpodobnosti* $Pr(\mathbf{s}|\mathbf{t})$ pro všechny věty. Tyto pravděpodobnosti je v reálu ovšem nemožné získat, protože bychom v podstatě potřebovali nekonečný paralelní korpus, který by obsahoval přinejmenším aspoň jeden výskyt každé možné dvojice vět $(\mathbf{s}, \mathbf{t})$. Abychom se tomuto problému vyhnuli a mohli získat překladové pravděpodobnosti použitelné pro nalezení nejlepšího překladu $\hat{\mathbf{t}}$, zavedeme do překladového modelu skrytou proměnnou $\mathbf{a}$ – *přiřazení* (angl. alignment). Překladový model pak vypadá takto:

$$Pr(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}). \tag{1.14}$$

Proměnná $\mathbf{a}$ určuje, jak spolu souvisí jednotlivá zdrojová slova s_j a cílová slova t_i, tj. určuje přiřazení mezi pozicemi slov ve zdrojové větě a pozicemi slov v cílové větě: $\mathbf{a} = a_1^J = a_1 ... a_j ... a_J, a_j \in \{0, ..., I\}$. Nyní se omezíme jen na ta přiřazení, kde je každému zdrojovému slovu s_j přiřazeno právě jedno nebo žádné (pro to je zde $a_j = 0$) cílové slovo t_i (odtud slovní překlad, neboť překladové pravděpodobnosti modelují vztah mezi dvěma slovy). Příklad takového přiřazení můžeme vidět na Obrázku 1.3.

Je-li tedy zdrojové slovo na pozici j přiřazeno cílovému slovu na pozici i, tak $a_j = i$. Bez



Obrázek 1.3: Příklad přiřazení mezi slovy při překladu z češtiny do angličtiny.

ztráty obecnosti můžeme nyní zapsat:

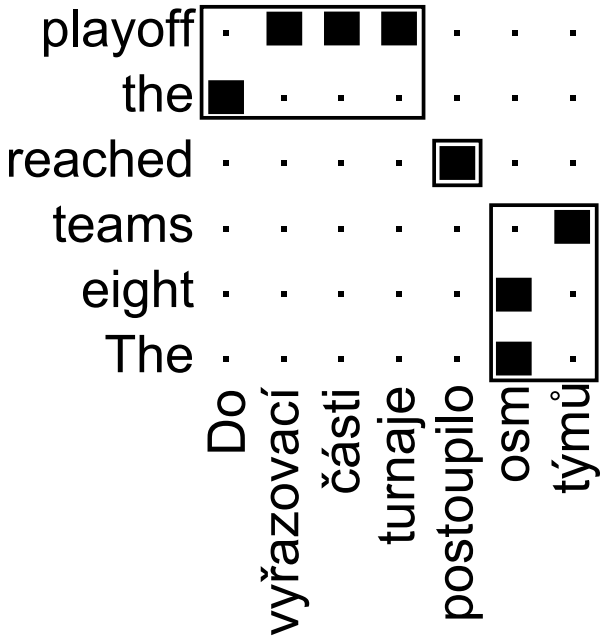
$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) = Pr(J|\mathbf{t}) \cdot \prod_{j=1}^{J} Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}) \cdot Pr(s_j|a_1^j, s_1^{j-1}, J, \mathbf{t}), \tag{1.15}$$

kde $Pr(J|\mathbf{t})$ představuje pravděpodobnost délky J zdrojové věty $\mathbf{s}$ za podmínky cílové věty $\mathbf{t}$, $Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t})$ je pravděpodobnost $j-$tého přiřazení a_j za podmínky posloupnosti předešlých přiřazení a_1^{j-1}, posloupnosti zdrojových slov s_1^{j-1}, délky zdrojové věty J a cílové věty $\mathbf{t}$ a $Pr(s_j|a_1^j, s_1^{j-1}, J, \mathbf{t})$ pak je pravděpodobnost $j-$tého zdrojového slova s_j za podmínky posloupnosti přiřazení a_1^j a dále stejně jako u předchozí pravděpodobnosti. Toto je jen jedna z možností, jak může být pravděpodobnost $Pr(\mathbf{s}, \mathbf{a}|\mathbf{t})$ zapsána jako součin řady podmíněných pravděpodobností. Rovnice 1.15 je přesným vyjádřením pravděpodobnosti $Pr(\mathbf{s}, \mathbf{a}|\mathbf{t})$ bez aproximací. Pravá strana rovnice však takto obsahuje příliš mnoho nezávislých parametrů, které by bylo třeba získat z trénovacích dat. V práci [Brown 93] bylo proto navrženo pět modelů (v literatuře jsou nazývány jako IBM nebo Model 1 – 5), které představují různé aproximace rovnice 1.15. Modely jsou řazeny vzestupně od jednodušších, které obsahují méně parametrů a jsou jednoduše trénovatelné, až po složité modely s více parametry a složitějším trénováním (při trénování složitějších modelů je třeba obvykle použít aproximace, neboť nelze nalézt optimální řešení v přijatelném čase). Vyšší modely berou do úvahy více parametrů z rovnice 1.15 a měli by tak lépe popisovat skutečný překladový model $Pr(\mathbf{s}, \mathbf{a}|\mathbf{t})$, jež se snažíme určit. Při trénování překladového modelu se počítá s tím, že vstupem vyššího modelu je výstup modelu nižšího. Díky tomu lze zjistit i hodnoty vyšších, výpočetně náročných modelů. Podrobněji o těchto modelech viz práce [Brown 93] a také dále v této práci Kapitola 4.1.

1.5.2 Frázový překlad

Později se v SMT prosadil přístup založený na překladu celých frází a ne jen jednotlivých slov. Frází se v této souvislosti rozumí neprázdná, souvislá řada slov. Většina dnešních frázových systémů je odvozena z přístupu založeného na *přiřazovacích vzorech* (angl. alignment templates). Tento přístup se poprvé objevil v práci [Och 99], kde je přiřazovací vzor definován jako trojice, která popisuje přiřazení mezi zdrojovou a cílovou frází. Fráze jsou přitom definované na slovních třídách. V pracích [Och 99, Och 02a] je ukázáno, že tento přístup založený na přiřazovacích vzorech překonává do té doby používaný přístup založený na slovním překladu. V současné době se používají také fráze, které jsou definovány přímo na slovech [Zens 02, Koehn 03, Zens 08].

Základní myšlenkou frázového překladu je rozdělit zdrojovou větu na fráze, tyto fráze přeložit a nakonec vhodným uspořádáním těchto překladů sestavit cílovou větu. Příklad takového rozdělení je na Obrázku 1.4. Formálně definujeme rozdělení daného páru vět (s_1^J, t_1^I) na K



Obrázek 1.4: Příklad rozdělení vět na fráze při překladu z češtiny do angličtiny.

překladových párů takto:

$$k \rightarrow a_k := (i_k; b_k, j_k),\ pro\ k = 1...K, \tag{1.16}$$

kde i_k označuje konec k-té cílové fráze a dvojice (b_k, j_k) začátek a konec zdrojové fráze, která je přiřazena k-té cílové frázi. Rozdělení na fráze je omezeno tak, že každé slovo ve zdrojové a cílové větě je pokryto právě jednou frází. V rozdělení tedy nejsou žádné mezery a také zde není žádný překryv mezi frázemi, tedy:

$$\bigcup_{k=1}^{K} [(b_k, j_k)] = \{1, ..., J\} \tag{1.17}$$

$$(b_k, j_k) \cap (b_{k'}, j_{k'}) = \emptyset \quad \forall k \neq k'. \tag{1.18}$$

Pro daný pár vět (s_1^J, t_1^I) a dané rozdělení $\mathbf{a} = a_1^K$ definujeme dvojjazyčné fráze takto:

$$\tilde{t}_k := t_{i_{k-1}+1}...t_{i_k} \tag{1.19}$$

$$\tilde{s}_k := s_{b_k}...s_{j_k}. \tag{1.20}$$

Délka fráze $\tilde{s}$ je pak definována jako $|\tilde{s}|$. Rozdělení a_1^K obsahuje také úplnou informaci o přeuspořádání cílových překladů tak, abychom dostali požadovanou cílovou větu. Do překladového modelu zahrneme rozdělení a_1^K opět jako skrytou proměnnou:

$$Pr(\mathbf{t}|\mathbf{s}) = \sum_{\mathbf{a}} Pr(\mathbf{t}, \mathbf{a}|\mathbf{s}) \tag{1.21}$$

$$= \sum_{\mathbf{a}} \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}, \mathbf{a}; \mathbf{s}))}{\sum_{\mathbf{t}', \mathbf{a}'} \exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}', \mathbf{a}'; \mathbf{s}))} \tag{1.22}$$

$$\approx \max_{\mathbf{a}} \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}, \mathbf{a}; \mathbf{s}))}{\sum_{\mathbf{t}', \mathbf{a}'} \exp(\sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}', \mathbf{a}'; \mathbf{s}))}. \tag{1.23}$$

Z teoretického hlediska je správné sčítat přes všechny možné segmentace zdrojové a cílové věty. V praxi se ovšem místo této sumy používá maximální aproximace (viz Rovnice 1.23), kde je součet nahrazen rozdělením s největší pravděpodobností, tzv. *Viterbiho přiřazením* (angl. Viterbi alignment). Při hledání nejlepšího překladu můžeme opět zanedbat jmenovatel, který je nezávislý na cílové větě $\mathbf{t}$ a dostaneme tak:

$$\hat{\mathbf{t}} = \underset{\mathbf{t}}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}, \mathbf{a}; \mathbf{s}) \right\}. \tag{1.24}$$

Výsledkem maximální aproximace je, že modely $h(.)$ nezávisí jen na větném páru $(\mathbf{s}, \mathbf{t})$, ale také na segmentaci $\mathbf{a}$, tj. máme modely $h(\mathbf{t}, \mathbf{a}; \mathbf{s})$. Během hledání nejlepšího překladu je vždy překladová hypotéza vytvářena frázi po frázi. Tento postup můžeme interpretovat jako řadu K rozhodnutí, kdy v každém kroku musíme rozhodnout o trojici $(\tilde{t}_k, b_k, j_k)$. V kroku k tedy přeložíme zdrojové pozice $b_k, ..., j_k$, tedy zdrojovou frázi s_k, jako cílovou frázi $\tilde{t}_k$ [Zens 08].

1.6 Syntaktický překlad

Od slov a frází se můžeme posunout ještě dále ke složitějším větným celkům, které jsou tvořeny syntaktickými a sémantickými strukturami (slouží jako formální reprezentace mínění dané věty). To odpovídá postupu do vyšších pater ve Vauquisově pyramidě. Automatický překlad založený na slovech a frázích tak můžeme z tohoto pohledu vnímat spíše jako přímý překlad. V případě syntaktického překladu je naším cílem využít při překladu informace obsažené v syntaktických a sémantických strukturách přiřazených zdrojové a případně i cílové větě. Hlavní motivací pro použití syntaktických překladových modelů je zajistit překlad a přeuspořádání frází a slov v cílové větě odpovídající lingvistickým znalostem daného jazyka.

Pro popis syntaktické struktury věty lze použít nějakou gramatiku, která popisuje obecný postup pro odvození přípustných kombinací slov daného jazyka tvořících věty. Syntaktická struktura dané věty je tak popsána pomocí syntaktického stromu, jehož uzly odpovídají jednotlivým použitým přepisovacím pravidlům dané gramatiky. V případě automatického překladu jsou pak využívány synchronní gramatiky, jejichž přepisovací pravidla obsahují přepisy pro oba jazyky. Při konstrukci stromu odpovídajícímu zdrojové větě je tak zároveň vytvářen strom odpovídající cílové věty (tu lze pak vygenerovat na základě tohoto stromu). Tento postup lze použít jen v případě, že máme k dispozici dvojice odpovídajících si stromů, tj. je k dispozici strom pro zdrojovou i cílovou větu. O korpusu obsahujícím tyto stromové páry pak hovoříme jako o treebanku. V případě češtiny je k dispozici Prague Dependency Treebank 2.0 (PDT, [Hajič 06]) a Prague Czech-English Dependency Treebank (PCEDT, [Čmejrek 04]) pro překlad mezi češtinou a angličtinou. Každá věta v PDT (PCEDT) má přiřazeny tři anotační vrstvy. První vrstva je morfologická a dále následují analytická (odpovídající povrchové syntaxi) a tektogramatická vrstva (odpovídající hloubkové syntaxi). Analytická a tektogramatická anotace je zachycena pomocí závislostního stromu. Z prací využívajících pro překlad synchronní gramatiky zmiňme jednu z posledních [Bojar 08], která se zabývá překladem z angličtiny do češtiny založeném na převodu hloubkové syntaxe. Pro překlad je využita synchronní gramatika se záměnou stromů (angl. synchronous tree substitution grammar) představená v práci [Hajič 02] a formalizovaná v pracích [Eisner 03, Čmejrek 06]. Pokud jsou k dispozici stromy jen pro jednu překladovou stranu, lze je využít tak, že je vytvořena synchronní gramatika, odrážející známou syntaxi. V případě druhé překladové strany je pak dovoleno libovolné přeuspořádání daných symbolů. V tomto případě pak hovoříme o překladu založeném na syntaxi (angl. syntax-based translation [Lopez 08]). Tento přístup byl použit např v pracích [Wu 98, Yamada 01]. Pokud

není k dispozici žádná lingvistická informace, lze využít postup, kdy jsou přepisovací pravidla dané struktury odvozena přímo z paralelního korpusu, viz práce [Chiang 05, Chiang 07]. Jestliže jsou pravidla konstruována nad frázemi, hovoříme pak o hierarchickém frázovém překladu (angl. hierarchical phrase-based translation [Lopez 08]).

1.7 Přehled systémů pro překlad do znakového jazyka

V této části popíšeme existující systémy pro automatický překlad znakových jazyků. Budou blíže popsány a zhodnoceny překladové systémy pro různé národní znakové jazyky. Systémy budou zhodnoceny především z hlediska použitého přístupu k automatickému překladu a také z hlediska použitých zdrojů dat. Popisované systémy můžeme rozdělit, podle již dříve zmíněného přístupu k automatickému překladu, na pravidlové a datové systémy. Většina prací, které se zabývají automatickým překladem znakového jazyka, je založena na pravidlovém přístupu k řešení problému automatického překladu. To je dáno tím, že neexistuje oficiální psaná forma žádného znakového jazyka a je tak tedy obtížné získat hlavní zdroj znalostí pro systémy založené na datovém přístupu – a sice paralelní korpus. S rostoucím úspěchem statistických překladových systémů pro mluvené jazyky se však objevily snahy použít tento přístup i pro překlad znakového jazyka. V druhé části této kapitoly tak bude představeno také několik systémů pro statistický překlad znakového jazyka. V případě použití tohoto přístupu se tak museli autoři, kromě problémů obvyklých pro statistické překladové systémy, vypořádat i s potřebou paralelního korpusu. Většina prací řeší tento problém návrhem vlastního zápisu znakového jazyka a vytvořením potřebného korpusu. Tyto korpusy tak zatím mají v porovnání s korpusy mluvených jazyků zanedbatelnou velikost. Tato velikost také v řadě případů limituje přesnost překladů získaných těmito systémy.

1.7.1 Pravidlové systémy

Práce [Huenerfauth 03] přináší přehled a detailní srovnání čtyř systémů určených pro překlad z angličtiny do znakového jazyka. Všechny čtyři systémy jsou založeny na pravidlovém přístupu k překladu a používají rozdílné lingvistické nástroje a postupy pro řešení problémů spojených s překladem do znakového jazyka.

Prvním systémem je VISICAST Translator, který vznikl v rámci projektu VISICAST Evropské Unie. Tento systém je určený pro překlad angličtiny do Britského znakového jazyka[3] (angl. British Sign Language (BSL)), v rámci projektu se také počítalo s překladem do Německého a Holandského znakového jazyka (DGS a NGT) ([Marshall 02, Safar 02, Marshall 01]). Zatím však byl realizován jen překlad do BSL. Vstupní text je nejprve analyzován CMU Link parserem [Sleator 91], takto získaná syntaktická struktura věty je pak pomocí pravidel vytvořených v jazyce Prolog převedena do sémantické reprezentace založené na discourse representation structures (DRS) [Bos 94]. Pro nalezení správného překladu je tato DRS reprezentace (odpovídající angličtině) převedena na jinou DRS reprezentaci, která odpovídá reprezentaci překládané věty ve zvoleném znakovém jazyce. Poté jsou na převedenou reprezentaci aplikována pravidla head-driven phrase structure (HDPS) gramatiky [Pollard 94], která slouží k odvození gramaticky správně utvořené věty ve zvoleném znakovém jazyce. Takto odvozená věta je již zapsána pomocí textové reprezentace znaků vhodné pro animaci (v tomto projektu je použita XML verze notace HamNoSys (Hamburg notation system) [Prillwitz 89] nazývaná sign gesture markup language (SiGML) [Elliott 00]).

[3] http://en.wikipedia.org/wiki/British_Sign_Language

Druhým systémem je systém ZARDOZ [Veale 98], který je založen na převodu vstupního textu na množinu ručně vytvořených konceptů, které slouží jako interlingua. Tento systém je určen pro překlad z angličtiny do různých znakových jazyků (Britského znakového jazyka, Amerického[4] (angl. American) (ASL), Irského[5] (angl. Irish) (ISL) a Japonského[6] (angl. Japanese) (JSL) znakového jazyka). Celý systém je založen na využití samostatných úkolově-orientovaných znalostních démonů, kteří spolu komunikují pomocí tabule. Vstupní text je nejprve morfologický zpracován, aby byla nalezena složená slova. Pak je provedena idiomatická redukce a text je rozparsován pomocí unifikační gramatiky (angl. unification grammar), která vytvoří hloubkovou syntakticko/sémantickou reprezentaci textu. Z této reprezentace je pomocí schematizace, která odstraní metaforické a metonymické struktury typické pro zdrojový jazyk, vytvořena jazykově nezávislá mezijazyková reprezentace. Při generování překladu ve znakovém jazyce je nejprve provedena anaforická rezoluce a pak, pomocí prostorových závislostních grafů (popisujících syntaxi znakového jazyka) a funkcí pro mapování znaků (převod konceptů na znaky), vytvořena posloupnost znaků. Nakonec je tato posloupnost znaků převedena na program v DCL (angl. doll control language) jazyce, který slouží pro animaci loutky na obrazovce počítače. Jestliže pro vstupní text neexistuje odpovídající koncept, je přeložen po jednotlivých slovech.

Třetím systémem je ASL Workbench [Speers 01], který je určen pro překlad angličtiny do Amerického znakového jazyka. Tento systém je založen na využití lexikálně-funkční gramatiky (angl. lexical-functional grammar (LFG), viz [Kaplan 82]). Text je pomocí LFG převeden na funkční strukturu (anglická f-struktura), ta je potom pomocí ručně vytvořených pravidel převedena na funkční strukturu odpovídající ASL (ASL f-struktura). Nakonec jsou opět použita pravidla LFG pro vytvoření posloupnosti znaků ASL. Pro překlad anglických slov a frází do ASL je použit speciální převodní slovník. Kdykoliv systém zaznamená problémy při překladu, je uživatel vyzván k jejich řešení. Systém také při překladu vytváří jednoduchý model anglické promluvy (skládá se ze seznamu elementů promluvy a jejich prostorového umístění, jestliže je specifikováno), všechna referenční rozlišení (tj. nalezení odkazovaných entit) však musí být provedena uživatelem. Fonologický model ASL v tomto systému je založen na Movement-Hold fonologii [Liddell 89].

Posledním systémem zmíněným v této práci je TEAM projekt [Zhao 00], který vznikl na univerzitě v Pensylvánii jako systém pro překlad z angličtiny opět do ASL. Tento systém využívá přístup založený na lexicalized tree adjoining grammar [Joshi 87] pravidlech k vytvoření lexikalizované syntaktické struktury ASL během vytváření závislostního stromu odpovídajícího překládanému anglickému textu. Z takto získané syntaktické struktury ASL je pak vytvořena odpovídající promluva v ASL.

Práce [Huenerfauth 04a, Huenerfauth 04b, Huenerfauth 06] pak přináší vlastní řešení problému návrhu systému pro překlad angličtiny do ASL. Navrhovaný systém používá pro překlad všechny tři vrstvy překladové pyramidy. Převodní vrstva je určena pro překlad většiny anglických vět. Věty, které při překladu do ASL obsahují tzv. „*klasifikátorové přísudky*“ (angl. classifier predicates) jsou překládány pomocí převodu do interlinguy. Zbylé věty, které nelze přeložit pomocí 2. ani 3. převodní vrstvy, se překládají přímo. V případě překladu pomocí převodu se počítá s využitím klasických metod MT určených pro překlad mluvených jazyků. V případě klasifikátorových přísudků tento postup nelze využít, neboť u nich dochází k využití prostoru před mluvčím a k odkazování na věci, které byly do tohoto prostoru umístěny. Proto byl pro tyto věty zvolen překlad založený na převodu do interlinguy, v kterém lze tato pro-

[4] http://en.wikipedia.org/wiki/American_Sign_Language
[5] http://en.wikipedia.org/wiki/Irish_Sign_Language
[6] http://en.wikipedia.org/wiki/Japanese_Sign_Language

storová uspořádání vyjádřit nezávisle na jazyku. Jako interlingua je použit symbolický popis 3D virtuální scény založený na parametrizovaných popisech akcí (angl. parameterized action representation (PAR) [Badler 00]). PAR lze považovat za animačně/lingvistická primitiva pro strukturovaný popis pohybů ve 3D scéně (jde vlastně o struktury s množstvím parametrů, které mohou nabývat různých hodnot podle toho, o jaký pohyb jde a kdo ho koná). Vstupní text je tedy nejprve převeden do PAR struktur odpovídajících anglické větě. Na základě toho je vytvořen model 3D scény. Tento model je pak převeden do PAR struktur odpovídajících klasifikátorovým přísudkům (překládají se jen věty obsahující klas. přísudky). Z těchto struktur je nakonec obdobným procesem jako v případě animace 3D scény vytvořena animace znaků.

Práce [Suszczanska 02] představuje systém TGT určený pro překlad z polštiny do polského znakového jazyka. Překlad je založen na převodu mezi sémantickou strukturou polské věty a jí odpovídající sémantickou reprezentací ve znakovém jazyce. Věta je nejprve syntakticky analyzována pomocí gramatiky *syntaktických skupin* (angl. syntactic groups (SG) [Suszczańska 99]). SG gramatika je určena pro analýzu jazyka s volným pořádkem slov. Syntaktická skupina je nějaká množina slov, které se objeví ve větě (nemusí spolu sousedit). Syntaktická reprezentace věty popisuje i vztahy mezi členy syntaktických skupin na všech úrovních. Vytvořená syntaktická struktura pak může být reprezentována grafem, kde jednotlivé uzly odpovídají jednotlivým syntaktickým skupinám a hrany mezi nimi zachycují jejich vzájemné syntaktické vztahy. Kořenem tohoto grafu je slovesná skupina. Z této syntaktické struktury je pak vytvořena přísudková reprezentace. Jde o graf, který má stejnou strukturu jako ten syntaktický, jednotlivým uzlům je ovšem přiřazena nějaká sémantická role (jsou použity jen tři role: Akce, Agent, Objekt). Z této sémantické reprezentace je pak pomocí generativní gramatiky vytvořena sémantická reprezentace odpovídající znakové větě (předpokládá se, že obě věty mají shodnou přísudkovou reprezentaci, je ovšem třeba doplnit některé chybějící věci, které jsou v polské větě vyjádřeny implicitně (např. osoba)). Z doplněné přísudkové reprezentace odpovídající znakovému jazyku už je pak vytvořena posloupnost znaků vhodná pro animaci.

1.7.2 Systémy založené na datech

První prací, která se zabývá použitím SMT systému pro překlad znakového jazyka, je práce [Bauer 99]. V této práci je navržen systém pro překlad z Německého znakového jazyka (angl. German Sign Language[7] (DGS)) do němčiny. Jedná se tak o opačný směr než v předchozích pracích. Systém se skládá ze dvou částí. První částí je modul pro rozpoznávání znaků DGS a druhou pak systém pro překlad rozpoznaných znaků do němčiny. Vzhledem k době návrhu se jedná o slovní překladový systém založený na modelu zdrojového kanálu. V práci jsou však uvedeny experimentální výsledky jen pro modul rozpoznávání, který dosahuje pro slovník velikosti 100 znaků více než 90 % úspěšnosti rozpoznávání. Pro modul překladu se pak předpokládá obdobná více než 90 % úspěšnost překladu.

Práce [Bungeroth 04] tak jako první přináší výsledky použití SMT systému pro překlad znakového jazyka. Jde o překlad z němčiny opět do DGS. V článku je porovnán překladový systém založený na přiřazovacích vzorech se systémem založeným na slovním překladu. Tak, jako v případě mluvených jazyků, je i zde úspěšnost překladů u systému založeného na přiřazovacích vzorech výrazně vyšší než u slovního překladu. V práci je také představen první paralelní korpus znakového jazyka. Tento korpus obsahuje 1399 páru vět, což je v porovnání s korpusy mluvených jazyků, které obsahují minimálně desítky tisíc párů vět, zanedbatelná velikost, tomu tak také odpovídají dosažené výsledky přesnosti překladu. Pro zápis znakového jazyka byla v tomto korpusu použita notace založená na glosách (angl. gloss notation). Tato

[7] http://en.wikipedia.org/wiki/German_Sign_Language

notace se používá pro transkripci video nahrávek znakového jazyka. Glosy tvoří sémantickou reprezentaci znakového jazyka, jedna glosa tak popisuje význam znaku spolu s nemanuální složkou znaku. Díky této notaci však obsahoval korpus mnoho slov, která se vyskytla jen jednou, tzv. *singletonů*. Takovýto korpus je nevhodný pro trénování SMT systému, proto byl vytvořen menší korpus (cca 200 párů vět), který obsahoval malé množství singletonů a byl použit pro natrénování a otestování vytvořeného systému.

Práce [Morrissey 05] je založena na využití metod učení z korpusu. Konkrétně je v této práci představen systém pro překlad z angličtiny do Nizozemského znakového jazyka (Sign Language of the Netherlands[8] (NGT)) založený na učení z příkladů. K učení je použit anotovaný korpus NGT, který vznikl v rámci projektu ECHO[9] (cílem ECHO projektu je vytvořit plně anotované korpusy různých znakových jazyků). Anotace zahrnuje časové vymezení odpovídajících si částí paralelních textů (text a znakový jazyk), dále popis artikulace jednotlivých rukou a také popis nemanuálních složek (ústa, obočí, otevření očí, směr pohledu, atd.), které přísluší k jednotlivým znakům. K nalezení překladu je použita metoda značkovací hypotézy (angl. marker hypothesis), která předpokládá, že syntaktické struktury jsou v povrchové vrstvě jazyka vyznačeny uzavřenou množinou specifických lexémů a morfémů. Pomocí této uzavřené množiny slov (jedná se hlavně o slova z uzavřených tříd, tj. zájmena, číslovky, spojky, předložky a částice) tedy mohou být oba paralelní texty rozděleny na odpovídající si části. Tato segmentace je ovšem použita jen pro anglický text, neboť ve znakovém jazyce je množina značkovacích slov velmi redukována. Pro segmentaci textu ve znakovém jazyce byly proto použity časové značky pro vymezení odpovídajících si částí textů (to vedlo k rozdělení části korpusu ve znakovém jazyce na konceptuální části). Ze segmentů anglického textu a jim odpovídajících konceptuálních částí jsou pak vytvořeny překladové dvojice, které pak slouží jako šablony pro překlad.

Práce [Morrissey 07, Morrissey 08] pak popisují vylepšený EBMT systém a jeho kombinaci se SMT systémem při překladu do znakového jazyka. Konkrétně jde o EBMT systém MaTrEx (více viz [Stroppa 06]) a SMT systém vyvinutý na RWTH Aachen University (více viz [Matusov 06]). V článku [Morrissey 07] je popsána kombinace těchto systémů pro překlad z angličtiny do Irského znakového jazyka a z němčiny do Německého znakového jazyka. Tato kombinace má příznivý účinek na úspěšnost automatického překladu v případě obou znakových jazyků. Pro natrénování obou systémů byl použit ATIS (Air Travel Information Service, více viz [Hemphill 90]) korpus, který byl přeložen do ISL a DGS [Bungeroth 08]. Jedná se o korpus, který obsahuje informace z oblasti letecké dopravy a obsahuje 595 anglických vět. Práce [Morrissey 08] se pak vzhledem ke své povaze zabývá podrobným popisem tvorby EBMT systému pro překlad znakového jazyka. Je podrobně popsána tvorba paralelního korpusu znakového jazyka a navržen a otestován prototyp EBMT systému. Dále jsou pak popsány experimenty se systémem MaTrEx a způsob evaluace výstupu systému pro automatickou syntézu znakového jazyka.

V práci [Stein 06] je představen frázový systém pro překlad z němčiny do DGS. Tento systém je dále rozšířen o před- a pozpracování založené na morfo-syntaktické analýze němčiny. Předzpracování zahrnuje zkrácení německých slov na jejich kořen (angl. stemming) dále rozdělení složených slov a smazání slov, která se v DGS nepoužívají (jde především o členy a některé spojky). Pozpracování se pak soustředí na správný překlad prostorových výrazů. Systém je natrénován a otestován na novém korpusu z oblasti předpovědi počasí - Phoenix korpus, který obsahuje 2 468 párů vět.

[8] http://en.wikipedia.org/wiki/Dutch_Sign_Language
[9] http://www.let.kun.nl/sign-lang/echo/

Kapitola 2

Cíle dizertační práce

Hlavním cílem této dizertační práce je návrh a realizace automatického systému pro obousměrný překlad mezi češtinou a znakovanou řečí. Jak bylo řečeno v předcházející kapitole, pojem znakovaná řeč zahrnuje dva jazyky - český znakový jazyk (ČZJ) a znakovanou češtinu (ZČ). Z předešlé kapitoly také dále vyplývají dva základní přístupy ke konstrukci automatického systému pro překlad – pravidlový a datový. Proveďme nyní diskuzi o použitelnosti jednotlivých přístupů pro řešení našeho zadání. Pokud bychom chtěli použít pravidlový systém pro překlad znakované řeči, museli bychom vytvořit překladová pravidla jak pro ZČ, tak i pro ČZJ. Vzhledem k tomu, že ZČ používá gramatická pravidla češtiny mohli bychom v tomto případě vystačit se systémem přímého překladu (jak však ukázaly předběžné testy, není tento přístup dostatečný (viz práce [Kanis 06]), resp. lze použitím jiných přístupů dosáhnout výrazně lepších výsledků). V případě ČZJ by však bylo třeba do překladu zapojit i vyšší vrstvy jazyka (syntaktickou a sémantickou). To ovšem předpokládá, např. pro zapojení syntaktické vrstvy jazyka, existenci uceleného gramatického popisu daného jazyka. Pro ČZJ však zatím žádný takovýto popis neexistuje, v předešlé kapitole jsou uvedeny jen poznámky ke gramatice ČZJ získané během jeho zatím poměrně krátkého lingvistického průzkumu (lingvistický výzkum ČZJ inicioval v roce 1993 Institut pro neslyšící v Berouně a v současné době dále probíhá především na Univerzitě Karlově v rámci oboru Čeština v komunikaci neslyšících). Další nevýhodou pravidlového systému je také jeho obtížná adaptace pro překlad dalších jazyků. Kdybychom chtěli vytvořený systém použít pro překlad dalších, ať už znakových nebo mluvených jazyků, museli bychom vytvořit překladová pravidla i pro tyto jazyky. Z těchto důvodů se tedy jako vhodnější jeví datový přístup, neboť vytvořený systém bude použitelný jak pro znakovanou řeč tak i další znakové i mluvené jazyky.

Abychom ovšem mohli tento přístup použít, potřebujeme paralelní korpus znakované řeči. Paralelní korpus je totiž klíčovou součástí při tvorbě systému založeného na datech, neboť slouží jako hlavní zdroj znalostí o překládaných jazycích. Získání takového korpusu je i v případě mluvených jazyků obtížné, neboť tvorba paralelního korpusu je časově i finančně náročná. V případě znakované řeči je pak další komplikací také neexistence oficiální psané formy ani ČZJ, ani ZČ. Nelze tak korpus vytvořit např. z překladů stejného, volně přístupného textu, jako je to běžné u mluvených jazyků. Např. pro dvojici angličtina čeština byl takto vytvořen volně přístupný korpus CzEng (Czech-English Parallel Corpus), který obsahuje volně přístupné paralelní texty (více viz [Bojar 06]). Součástí práce tedy musí být i vytvoření paralelního korpusu znakované řeči, který pak bude použit při konstrukci systému pro obousměrný automatický překlad mezi češtinou a znakovanou řečí. Abychom tento korpus mohli vytvořit, je třeba navrhnout psanou formu znakované řeči, tedy ČZJ a ZČ. Jak ukazují poznatky o znakových jazycích obecně i předešlé práce zabývající se automatickým překladem znako-

vých jazyků zmíněné na konci předešlé kapitoly, není vytvoření psané formy znakového jazyka triviální záležitostí. Žádný z navržených způsobů zápisu použitých v popsaných systémech, zatím uspokojivě nevyřešil všechny problémy spojené se zápisem znakového jazyka. Většina těchto problémů je spojena s existencí znakového jazyka v prostoru a obtížností zaznamenat tuto prostorovou orientaci pomocí textu. Překlad do ČZJ je pak dále také komplikován jeho současným stavem, kdy neexistuje všeobecně používaná a všeobecně platná jazyková varieta v rámci jednoho jazykového společenství, obvykle označovaná jako spisovný jazyk. Je tedy obtížné zajistit kvalifikovaný překlad zvoleného textu do ČZJ, neboť i sami neslyšící dávají při komunikaci se slyšícími přednost ZČ, která je jimi považována za prestižnější formu znakované řeči než ČZJ (navíc rodilým mluvčím ČZJ se může stát jen neslyšící dítě neslyšících rodičů, pouze 5-10 % neslyšících rodičů má však také neslyšící dítě [Macurová 98]). Tento stav je daný hlavně pozicí ČZJ před rokem 1989, kdy bylo zakázáno používání tohoto jazyka ve vzdělávacím procesu neslyšících, kde se převážně uplatňoval mluvený jazyk a ZČ (odtud také nejspíše pramení upřednostňování ZČ před ČZJ). ČZJ tak byl pouze neoficiálním jazykem používaným samotnými neslyšícími. S těmito a dalšími problémy se potýkají i samotní lingvisté při výzkumu ČZJ (více viz např. práce [Macurová 01a]). Hlavně z těchto důvodů jsme se rozhodli v rámci této práce vytvořit paralelní korpus ZČ. Díky použitému přístupu k tvorbě systému však bude vytvořený systém jednoduše použitelný, v případě existence paralelního korpusu, i pro překlad ČZJ.

Shrňme nyní tedy cíle této dizertační práce. Hlavním cílem je vytvoření jazykově nezávislého automatického systému pro překlad a jeho použití pro obousměrný překlad mezi češtinou a znakovanou češtinou. Z nabízených možností jsme se rozhodli pro statistický přístup k tvorbě překladového systému, základní překladovou jednotkou bude fráze. Frázové systémy jsou v současné době nejpoužívanějším typem SMT systémů a dosahují špičkových výsledků z hlediska přesnosti a rychlosti při překladu různorodých jazykových párů. S ohledem na návrh a realizaci překladového systému můžeme nyní definovat tyto dílčí cíle:

- Vytvoření paralelního korpusu čeština – znakovaná čeština. Bude vytvořena psaná forma ZČ a vybrán a přeložen vhodný český text. Při překladu bude anotátory vyznačeno přiřazení mezi překládanými frázemi.
- Návrh vlastní automatické metody výběru frází z paralelního korpusu a vyhodnocení její výkonnosti při překladu ZČ.
- Návrh vlastního dekodéru pro frázový překlad. Při návrhu bude kladen důraz na použití dekodéru v reálných aplikacích. Výkon vytvořeného dekodéru bude porovnán s volně dostupným frázovým dekodérem, který představuje současný standard mezi frázovými dekodéry.
- Porovnání úspěšnosti překladu s ručně vytvořenými frázemi s úspěšností překladu s frázemi získanými automaticky a to standardní metodou a metodou navrženou v této práci.
- Otestování výkonu navrženého překladového systému při překladu z češtiny do znakované češtiny.
- Návrh úprav základního systému a jejich otestování při obousměrném překladu z češtiny do znakované češtiny.

Kapitola 3

Czech – Signed Czech (CSC) paralelní korpus

V této kapitole popíšeme tvorbu paralelního korpusu pro jazykovou dvojici: čeština – znakovaná čeština. Bude popsán již existující český korpus, který byl zvolen jako základ pro vytvoření paralelního korpusu a navržena psaná forma ZČ, která bude následně použita k překladu tohoto korpusu. Dále bude popsán proces překladu a použité nástroje a techniky. Nakonec bude porovnána shoda mezi jednotlivými překladateli a na jejím základě bude určena horní hranice úspěšnosti automatického překladového systému vytvořeného na základě tohoto nového paralelního korpusu.

Jako paralelní korpus označujeme sbírku textů ve zdrojovém jazyce a jim odpovídajících překladů v cílovém jazyce. V korpusu může být i více překladů téže zdrojové věty. To je výhodné, neboť jednotlivé překlady stejné věty se mohou značně lišit. Je-li v korpusu více překladů ke každé zdrojové větě, může se systém naučit více variant překladu a také pak může dosáhnout vyšší úspěšnosti při překladu testovacího textu. Rozmanitost překladů stejné zdrojové věty je dána nejednoznačností přirozeného jazyka, kdy lze řadu věcí vyjádřit několika různými způsoby. Výsledný překlad je také ovlivněn osobností a náladou překladatele, tj. různí lidé nebo ten samý člověk v jiném čase mohou stejnou větu přeložit jinak a vždy z hlediska obsahu a stylu správně. Aby se omezil vliv překladatele, bylo by tedy lepší použít pro celý korpus jen jednoho překladatele. To ovšem není možné, neboť by pak tvorba korpusu zabrala příliš mnoho času. Z tohoto důvodu se tedy obvykle, pokud je to možné, používá při testování více různých překladů stejného textu. Korpus také dále obsahuje informaci o přiřazení mezi jednotlivými větami a jejich překlady. Toto přiřazení může být vytvořeno ručně, pokud je např. daný zdrojový text překládán pro potřeby korpusu (víme pak přesně který překlad odpovídá které zdrojové větě), nebo automaticky (viz např. práce [Gale 93]), pokud do korpusu zařadíme celý text a jeho překlad (např. text knihy nebo nějakého dokumentu v obou jazycích).

Jak již bylo řečeno, je paralelní korpus klíčovou součástí při konstrukci každého SMT systému. Jeho rozsah přímo ovlivňuje úspěšnost překladu vytvořeného systému, protože pokud bude v tomto korpusu málo vět, nebude z něj možné získat použitelné znalosti (v případě frázového systému jde především o pravděpodobnosti překladu mezi jednotlivými frázemi a také o pravděpodobnosti pro jazykový model). Protože jsme pro překlad ZČ nuceni si vytvořit vlastní korpus, je lepší zvolit nějaké omezené téma a tím získat dostatečně rozsáhlý korpus i při menším počtu obsažených vět. Jako základ našeho CSC korpusu jsme tak zvolili existující Human–Human Train Timetable (HHTT) dialogový korpus [Jurčíček 05, Jurčíček 07], který obsahuje přepisy telefonních dotazů do informačního centra vlakových jízdních řádů.

3.1 Human–Human Train Timetable (HHTT) dialogový korpus

HHTT korpus byl vytvořen v rámci práce na automatickém dialogovém systému ([Jelínek 03, Jelínek 04, Jurčíček 05, Jurčíček 07]) a obsahuje záznam telefonické komunikace mezi uživatelem a operátorem informačního centra vlakových jízdních řádů. V korpusu je uložen záznam rozhovoru mezi uživatelem a operátorem, ortografická a normalizovaná transkripce tohoto rozhovoru a navíc také normalizovaná transkripce doplněná o *pojmenované entity* (angl. named entities, jde např. o jména osob, zastávek atd.) a značky dialogových aktů, které představují abstraktní sémantickou anotaci promluvy. Rozhovory byly zaznamenávány průběžně od dubna do září v roce 2000. Bylo nahráno celkem 6 548 rozhovorů a 6 353 jich bylo transkribováno. Celý korpus tedy obsahuje 106 hodin spontánní řeči, která je nahrána mono a vzorkována na 8kHz pomocí A–Law komprese. Průměrná délka rozhovoru je 60 sekund. Rozhovory byly rozděleny na kola podle změny řečníka, v korpusu je tak celkem 81 543 kol. Z pohledu na slova obsahuje korpus více než 603 000 tokenů, které tvoří slovník 12 000 unikátních slov (slovník operátora je menší a obsahuje 5 839 položek, slovník uživatele pak 9 485 položek).

Při přepisu rozhovorů do textu byla použita ortografická transkripce [Psutka 04], protože je nejvhodnější pro přepis spontánní češtiny. Ta totiž obsahuje slova a obraty, která nejsou ve spisovné psané ani mluvené češtině. Normalizovaná transkripce byla vytvořena automaticky pomocí slovníku obsahujícího pouze spisovná česká slova. Jako entity vhodné k pojmenování byly vybrány jména osob, stanic, oblastí a vlaků. Pojmenované entity byly při přepisu označeny ručně anotátory a poté zkontrolovány pomocí slovníku, např. jména stanic byla zkontrolována pomocí slovníku všech stanic z internetové aplikace IDOS[1], která umožňuje vyhledávání vlakových spojení. Popišme nyní podrobněji schéma značek pro dialogové akty.

3.2 Značkovací schéma dialogových aktů

V HHTT korpusu je použito značkovací schéma inspirované DAMSL ((dialogue act markup in several layer [Allen 97]) schématem a primárně založené na DATE (dialogue act tagging for evaluation [Walker 01]) schématu. Toto schéma používá třídimenzionální anotaci – (1) DOMAIN (oblast), (2) SPEECH-ACT (řečový akt), (3) SEMANTIC (sémantika). V korpusu jsou takto anotovány vždy celé promluvy operátora i uživatele na rozdíl od DATE anotace, která byla navržena jen pro evaluaci dialogových systémů a obsahuje tak jen anotaci výstupů dialogového systému.

3.2.1 DOMAIN dimenze

Tato dimenze zařadí každou promluvu do jedné ze tří možných konverzačních akcí: **Task** (úkol), **Communication** (komunikace), **Frame** (rámec).

- *Task* – označuje promluvu, která souvisí se zodpovězením uživatelského dotazu, jedná se obvykle o dotaz nebo odpověď na dotaz o jízdním řádu nějakého spoje.
- *Communication* – pak označuje promluvu, která slouží k řízení dialogu a poskytuje zpětnou vazbu o tom co bylo rozuměno, často se tak zde objevuje opakování, např. času odjezdu spoje apod.

[1] http://jizdnirady.idnes.cz/vlakyautobusy/spojeni/

- *Frame* – je pro promluvu, která popisuje stav dialogu, jde např. o instrukce nebo omluvy.

3.2.2 SPEECH-ACT dimenze

SPEECH-ACT dimenze označuje komunikační cíl promluvy bez ohledu na její formu. Tato dimenze tedy může rozlišit promluvy, které mají stejné hodnoty sémantické dimenze (např. promluva se sémantickou hodnotou DEPARTURE(TIME, FROM(STATION)) může být na úrovni SPEECH-ACT dimenze rozlišena na *request_info* (požadovaná informace) nebo *present_info* (předložená informace)). V této dimenzi je definováno 15 následujících značek: acknowledgment, apology, closing, explicit_confirmation, implicit_confirmation, instruction, offer, opening, present_info, request_info, speech_repair, status_report, thanking, verify, verify_neg.

- *Acknowledgment* (přijetí)– označuje promluvu, která přijímá nebo zamítá nějakou dříve zmíněnou informaci.
- *Apology* (omluva) – označuje promluvu, která obsahuje omluvu uživateli.
- *Closing* (uzavření) – označuje promluvu, která ukončuje rozhovor mezi operátorem a uživatelem.
- *Explicit_confirmation* (přímé potvrzení) – používá se pro promluvu, která zmiňuje dřívější informaci a vyžaduje přímé potvrzení této informace, je tedy většinou vyžadována odpověď.
- *Implicit_confirmation* (nepřímé potvrzení) – používá se také pro promluvu, která obsahuje dřívější informaci. Obvykle se jedná o opakování této informace a odpověď se očekává jen v případě, že je tato informace mylná.
- *Instruction* (instrukce) – je pro promluvu, která obsahuje nějaké instrukce, např. co a jak udělat, co říci, atd.
- *Offer* (nabídka) – představuje promluvu, která uživateli nabízí nějakou další možnost, která nebyla uživatelem přímo požadována.
- *Opening* (otevření) – tato promluva začíná dialog.
- *Present_info* (předkládaná informace) – patří k promluvě, která uvádí nějakou novou informaci. Je to tak odpověď na promluvu označenou SPEECH-ACT značkou request_info.
- *Request_info* (požadovaná informace) – patří k promluvě, která požaduje nějakou informaci. Obvykle tak obsahuje dotazovaný objekt a případně jeho atributy.
- *Speech_repair* (opravná řeč) – označuje opakování a neplynulosti v řeči, které se obvykle objevují při opravě dříve řečeného.
- *Status_report* (situační zpráva) – označuje promluvy, které nesouvisí s tématem dialogu.
- *Thanking* (poděkování) – promluva, která obsahuje poděkování.
- *Verify* (ověření) – reprezentuje promluvu, která ověřuje dříve zmíněnou informaci. Ověření by se vždy mělo vztahovat k hlavnímu tématu rozhovoru.
- *Verify_neg* (ověření negací) – opět ověřuje dříve zmíněnou informaci, která ale reprezentuje významový opak použitého konceptu, tj. význam potvrzující a zamítající odpovědi je prohozen (ne znamená ano).

Mluvčí	DOMAIN + SPEECH-ACT SEMANTIC	Promluva
operátor	frame + opening GREETING	informace prosím
uživatel	frame + opening GREETING	dobrý den
	task + request_info DEPARTURE(TIME, TRAIN_TYPE, TO(STATION))	já mám prosbu jakpak jedou dneska osobní vlaky ňák dopoledne do starýho_plzence
operátor	frame + status_report OTHER_INFO	no tak tam už moc na výběr nemáte
	task + present_info TIME, TIME	teďka jede v osm šestnáct jestli stihnete potom až v jedenáct deset
uživatel	comm + implicit_conf TIME	až v jedenáct deset
	task + acknowledgment ACCEPT(TIME, FROM(STATION))	a to by tak ňák stačilo těch jedenáct deset z hlavního
	task + verify TRAIN_TYPE	jo a dá se tam vzít kočárek

Tabulka 3.1: Ukázka z HHTT korpusu.

3.2.3 SEMANTIC dimenze

Úkolem SEMANTIC dimenze je zachytit v každé promluvě relevantní informace vztahující se k hlavnímu cíli dialogu. Protože se pohybujeme v oblasti odpovědí na dotazy o vlakových spojích, je hlavním cílem zachytit informace potřebné pro zodpovězení dotazu. Tato sémantická anotace pak slouží k natrénování sémantického parseru (více o konstrukci sémantického parseru viz [Jurčíček 07, Jurčíček 08]). Aby se zjednodušilo trénování tohoto parseru, předpokládá se, že uspořádání abstraktních sémantických konceptů v anotaci je stejné jako uspořádání odpovídajících slov v promluvě. Cílem parseru je přiřadit sémantickou anotaci každé uživatelově a operátorově promluvě, z této anotace by pak mělo být možné jednoduchým algoritmem zjistit potřebné informace. V této dimenzi je definováno nejvíce značek, je jich 33 a jsou to: ACCEPT, AMOUNT, AREA, ARRIVAL, BACK, DELAY, DEPARTURE, DISCONNECT, DISTANCE, DURATION, FROM, GREETING, LENGTH, MAYBE, NEXT, NUMBER, OTHER_INFO, PERSON, PLATFORM, PREVIOUS, PRICE, REF, REJECT, REPEAT, STATION, SYSTEM_FEATURE, THROUGH, TIME, TO, TRAIN_TYPE, TRANSFER, WAIT, WHAT_TIME. Nyní blíže popíšeme ty nejdůležitější a nejpoužívanější značky.

- *ARRIVAL* (příjezd) – reprezentuje promluvu, která se ptá nebo zodpovídá dotaz o příjezdu vlaku.

- *ACCEPT* (přijetí) – představuje pozitivní odpověď, souhlas nebo přijetí předešlého faktu.
- *AMOUNT* (částka) – označuje konkrétní částku peněz.
- *BACK* (zpátky) – indikuje požadavek na zpáteční cestu.
- *DEPARTURE* (odjezd) – reprezentuje otázku nebo odpověď o odjezdu nějakého vlaku.
- *FROM* (z) – označuje odjezdovou stanici.
- *GREETING* (pozdrav) – znamená pozdrav a používá se jen se značkou SPEECH-ACT opening.
- *NEXT* (další) – představuje promluvu, která požaduje informaci o dalším vlaku. Používá se jen se značkou SPEECH-ACT request_info nikdy s present_info.
- *NUMBER* – označuje nějakou číselnou hodnotu, která ale není vyjádřením času. Často se používá pro určení pořadí požadavků v dotazu nebo čísla nástupiště a podobně.
- *OTHER_INFO* – používá se pro promluvy, které nespadají do žádného jiného konceptu.
- *PERSON* (osoba) – označuje jméno nebo identifikaci osoby. Používá se jen se značkou SPEECH-ACT opening.
- *PRICE* (cena) – znamená cenu jízdenky nebo dalších služeb.
- *PREVIOUS* (předchozí) – představuje promluvu, která požaduje informaci o dřívějším vlaku. Používá se jen se značkou SPEECH-ACT request_info nikdy s present_info.
- *REF* – představuje odkaz na již zmíněný objekt.
- *REJECT* (zamítnutí) – představuje zamítnutí nějaké informace.
- *STATION* (stanice) – znamená jméno stanice a to jak oficiální, tak i lokální nebo hovorové.
- *TIME* (čas) – označuje čas nebo datum.
- *TO* (do) – označuje cílovou stanici, používá se také ve smyslu směru jízdy vlaku.
- *TRAIN_TYPE* (typ vlaku) – určuje typ vlaku (osobní vlak, rychlík, atd.).
- *TRANSFER* (přestup) – označuje přestup mezi vlaky.

Pomocí dialogových značek popsaných výše bylo anotováno celkem 1 109 z 6 353 transkribovaných dialogů (ukázka anotace viz Tabulka 3.1). To představuje 12 395 kol, která byla při anotaci dále rozdělena na 17 900 dialogových aktů, které tvoří 118 000 tokenů [Jurčíček 07]. Jako základ našeho paralelního korpusu použijeme právě tyto anotované dialogy. Tato volba má řadu výhod. Shrňme si nyní ty nejdůležitější:

1. HHTT korpus představuje uzavřené a dobře definované téma.
2. HHTT korpus (resp. námi zvolená část) obsahuje množství anotačních vrstev, které přináší řadu informací použitelných pro další zpracování. Je zde transkripce a normalizovaná transkripce řeči, vyznačení pojmenovaných entit a anotace pomocí dialogových značek, která představuje sémantický popis jednotlivých promluv.

3. HHTT korpus je záznamem telefonních rozhovorů a jeho použitím pro vytvoření paralelního korpusu se tak oblast telefonní komunikace otevírá i neslyšícím. Jde také o záznam skutečných rozhovorů, takže dosažené výsledky by měly být použitelné i v reálných aplikacích.

4. Po přidání překladu do ZČ bude možné použít jediný korpus pro vytvoření systému pro překlad z mluvené do znakované řeči (rozpoznávač řeči na straně slyšícího a překladový systém a avatar na straně neslyšícího) a vyřešit tak problém komunikace ve směru slyšící – neslyšící v dialogu mezi slyšícími a neslyšícími. Zároveň lze tento korpus využít i pro vytvoření systému pro řízení automatického dialogu mezi počítačem a neslyšícím na dané téma.

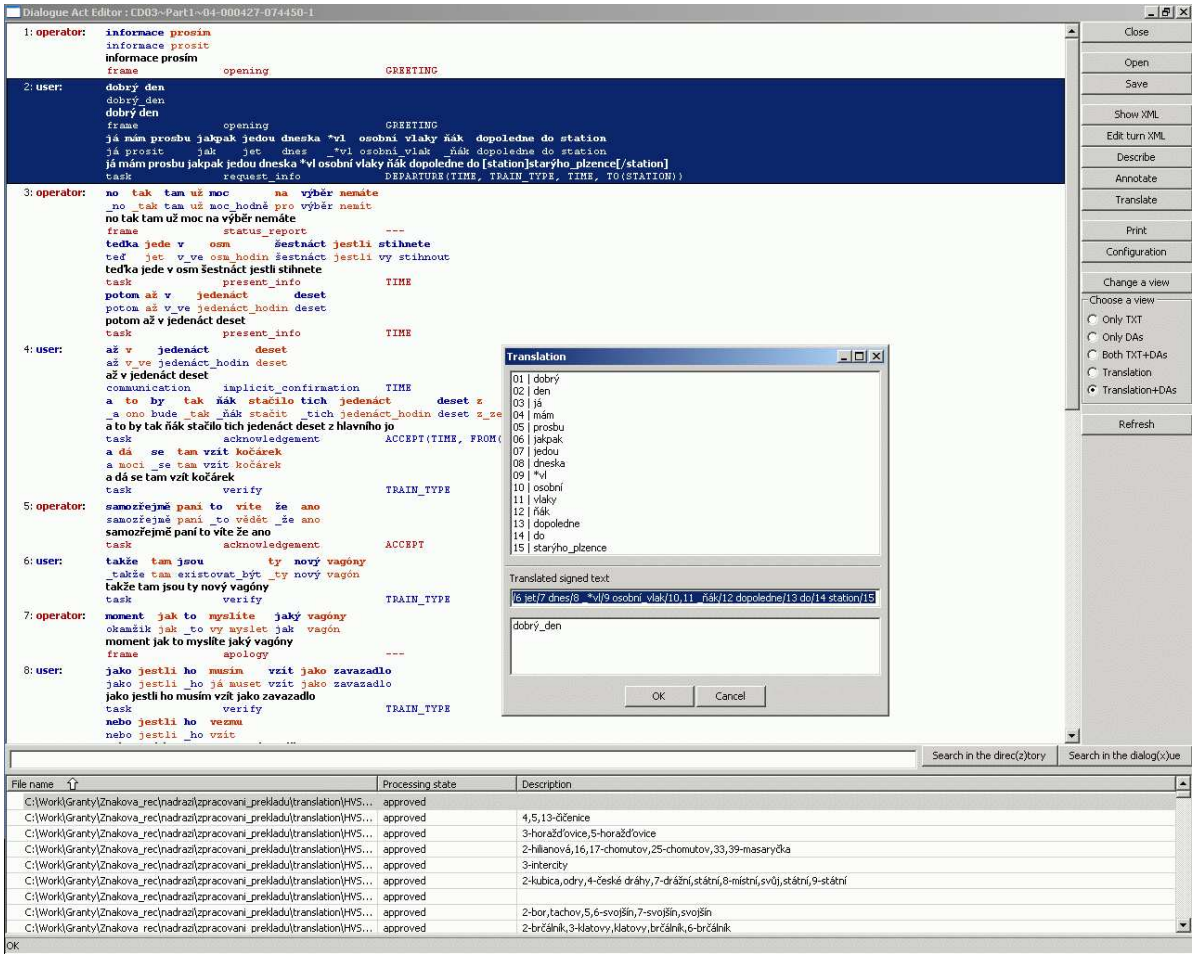
3.3 Překlad do znakované češtiny

Náš CSC korpus jsme se tedy rozhodli vytvořit překladem všech dialogů z HHTT korpusu, které byly plně anotovány, tj. obsahují anotaci pomocí dialogových značek popsaných v předcházející části. Abychom to mohli udělat, bylo nejprve třeba vytvořit psanou formu znakované češtiny. Jak bylo již řečeno v první kapitole, neexistuje žádná oficiální psaná forma žádného znakového jazyka, to samé platí i pro ZČ a je tedy třeba ji pro naše potřeby vytvořit. ZČ používá znaky ČZJ a gramatická pravidla češtiny. Její psanou formu tak lze jednoduše vytvořit zapsáním znaků ČZJ v pořadí odpovídajícím českým slovům v překládané větě. Každý znak ČZJ je třeba pro tento účel reprezentovat jedinečným řetězcem, tyto řetězce můžou např. odpovídat sémantickému obsahu znaku, jako je tomu v případě psané formy znakového jazyka založené na glosách. Abychom v našem případě zajistily konzistentnost překladů (tj. jednotné a jedinečné pojmenování znaků), používali všichni překladatelé při překladu stejný slovník. Pro překlad mohli tedy používat jen znaky z tohoto slovníku. Jako základ tohoto slovníku jsme použili textovou verzi největšího dostupného slovníku ČZJ (viz [Langer 04], tento slovník obsahuje 3 063 znaků). Slovník jsme upravili tak, aby byl pro každý znak jedinečný název, přidali jsme upřesňující popis u znaků, které to vyžadovaly a také jsme přidaly dva speciální znaky: znak „_“, který znamená prázdný překlad a znak „spelling“, který se používá pro slova, jež se hláskují pomocí prstové abecedy. Dále jsme také do tohoto slovníku přidali další znaky, které byly třeba pro překlad textů v korpusu. Tyto znaky byly buď převzaty z dalších slovníků nebo byly zjištěny přímo dotazem ve Spolku neslyšících Plzeň, celková velikost použitého slovníku byla tedy 3 185 znaků. Tím byla zajištěna skutečná existence všech znaků (tj. známe jejich prostorovou podobu), které byly použity při překladu tak, aby výsledný překladový systém byl použitelný pro automatickou syntézu znakované řeči (při ní je třeba znát prostorovou podobu všech syntetizovaných znaků).

3.3.1 DAE editor

Abychom urychlili proces překladu a zajistili konzistentnost překladů v korpusu (tj. použití stejné množiny znaků a vyznačení přiřazení mezi slovy a jejich překlady), využili jsme pro tvorbu překladů anotační nástroj, který byl vytvořen pro anotaci HHTT korpusu. Tento nástroj se nazývá Dialogue Act Editor[2] (DAE) a byl vytvořen na Katedře Kybernetiky FAV ZČU. Tento nástroj byl navržen tak, aby poskytoval řadu funkčností potřebných pro tvorbu kvalitních anotací, např. konfigurovatelnou množinu potřebných značek, anotační robustnost, validaci každého možného vstupu, různé možnosti zobrazení anotovaného dialogu a další. Při

[2] volně ke stažení na adrese: http://code.google.com/p/dialogue-act-editor/

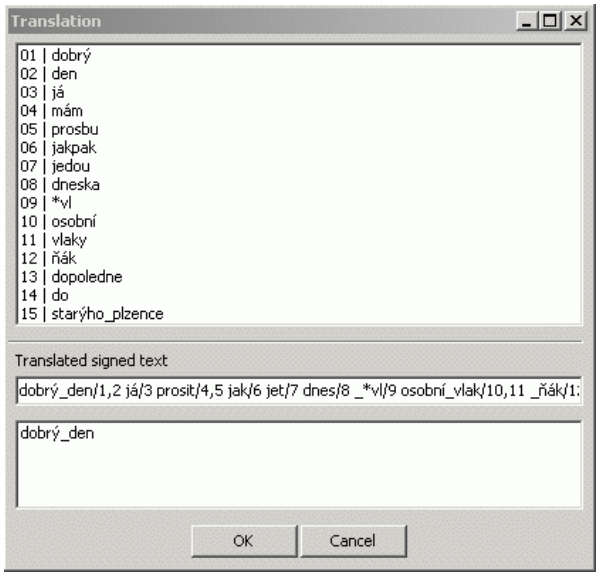


Obrázek 3.1: Základní okno DAE editoru s otevřeným dialogem pro překlad do ZČ.

návrhu editoru byla požadována nezávislost na operačním systému a snadná údržba. Proto byl pro implementaci zvolen programovací jazyk Python[3] spolu s GUI nástrojem wxPython[4], který zpřístupňuje mezi platformovou GUI knihovnu wxWidgets[5]. Abychom mohli tento nástroj použít pro tvorbu překladů, rozšířili jsme DAE editor o dialog umožňující přidat překlad do ZČ ke každé promluvě (toto rozšíření je již obsaženo v současně dostupné verzi, viz poznámka 2 pod čarou). Na Obrázku 3.1 je zobrazeno základní okno DAE editoru spolu s otevřeným dialogem pro zadání překladu zvolené promluvy.

Na Obrázku 3.2 je pak bližší pohled na samotný dialog pro překlad do ZČ. V horním okně jsou na řádcích jednotlivé očíslované tokeny promluvy, každý token má přiřazeno unikátní číslo. Pod tímto oknem je řádek, do kterého se zapisuje překlad promluvy do ZČ. Překlad lze vytvořit jen ze znaků, které jsou v používaném slovníku. Když tedy začne překladatel zapisovat do řádku jméno znaku, objeví se mu v dolním okně všechny znaky z používaného slovníku začínající napsaným řetězcem. Překladatel pak může dopsat jméno znaku nebo ho vybrat z nabídnutého seznamu. Uložit lze pak ale jen překlady, které obsahují jen názvy znaků uložené v použitém slovníku (tím je zajištěna konzistentnost vytvořených překladů z hlediska použitých názvů znaků). Každému znaku je také dále v tomto řádku vyznačeno přiřazení k

[3] http://www.python.org/
[4] http://wxpython.org/
[5] http://www.wxwidgets.org/



Obrázek 3.2: Dialog pro překlad vybrané promluvy do ZČ.

tokenu promluvy, jehož je překladem (čísla za lomítkem, která odpovídají očíslování tokenů v horním okně). Jednomu znaku lze přiřadit více tokenů (jejich čísla jsou oddělena čárkou) nebo lze více znaků přiřadit jednomu tokenu (stejné číslo za lomítkem). Z takto označených překladů pak lze jednoduše vytvořit tabulku frází použitelnou v automatickém překladovém systému. Při uložení vytvořeného překladu se kontroluje, zda je každému znaku přiřazen aspoň jeden token, tím je zajištěno, že překladatel musí při překladu vzít do úvahy všechna slova překládané promluvy (pokud se dané slovo nepřekládá, lze využít speciální znak „_“, viz výše). Tímto způsobem bylo přeloženo všech 1 109 dialogů z HHTT korpusu, které byly kompletně anotovány dialogovými značkami popsanými výše. To představuje paralelní korpus s 15 772 páry vět, které tvoří na straně češtiny 107 953 tokenů a na straně ZČ pak 107 663 tokenů. Slovník na straně češtiny je podle očekávání větší a obsahuje 4 081 slov, slovník ZČ pak jen 2 366 znaků. V Tabulce 3.2 je ukázka části vytvořeného korpusu spolu s anotací pomocí dialogových značek.

3.3.2 Spolehlivost překladů

Na tvorbě korpusu se podílely celkem čtyři různí překladatelé, dva zkušení překladatelé, kteří vystudovali obor speciální pedagogika se zaměřením na sluchově postižené a udržují pravidelný kontakt s neslyšícími a dva nezkušení, kteří absolvovali dvousemestrový kurz znakového jazyka na ZČU. Pro omezení možných chyb v překladu, byly vytvořené překlady vždy ještě zkontrolovány jiným překladatelem ze skupiny. Abychom zjistili vliv konkrétního překladatele na překlad, nechali jsme všechny čtyři překladatele přeložit stejnou množinu 50 dialogů (to představuje sadu 665 párů vět). To nám umožní porovnat shodu mezi jednotlivými překladateli. Také můžeme dále tyto překlady využít pro vyhodnocení kvality vytvořeného překladového systému (obvykle se totiž výsledky překladového systému, pokud je to možné, porovnávají s více různými překlady testovacího textu tak, aby se omezil vliv překladatele a nejednoznačnosti přirozeného jazyka).

Standardním postupem pro porovnání shody mezi anotátory, kteří provádí klasifikaci, je kappa statistika κ (první novodobou zmínku lze nalézt v práci [Cohen 60], její použití pro klasifikační úlohu pak v práci [Carletta 96]), která měří párovou shodu mezi anotátory, opravenou

Mluvčí	DOMAIN + SPEECH-ACT SEMANTIC	Promluva *Překlad v ZČ*
operátor	frame + opening GREETING	informace prosím *informace/1 _/2*
uživatel	frame + opening GREETING	dobrý den *dobrý_den/1,2*
	task + request_info DEPARTURE(TIME, TRAIN_TYPE, TO(STATION))	já mám prosbu jakpak jedou dneska osobní vlaky ňák dopoledne do starýho_plzence *já/3 potřebovat/4,5 kdy/6 jet/7 dnes/8 osobní_vlak/9,10 _/11 dopoledne/12 do/13 starý/14 plzeň/14 malý_věc/14*
operátor	frame + status_report OTHER_INFO	no tak tam už moc na výběr nemáte *_/1 _/2 _/3 už/4 moc_hodně/5 _/6 výběr/7 ne/8*
	task + present_info TIME, TIME	teďka jede v osm šestnáct jestli stihnete potom až v jedenáct deset *teď/9 jet/10 v_ve/11 osm_hodin/12 šestnáct/13 jestli/14 stihnout/15 potom/16 až/17 v_ve/18 jedenáct_hodin/19 deset/20*
uživatel	comm + implicit_conf TIME	až v jedenáct deset *až/1 v_ve/2 jedenáct_hodin/3 deset/4*
	task + acknowledgment ACCEPT(TIME, FROM(STATION))	a to by tak ňák stačilo těch jedenáct deset z hlavního *_/5 _/6 _/7 _/8 _/9 stačit/10 _/11 jedenáct_hodin/12 deset/13 z_ze/14 důležitý/15*
	task + verify TRAIN_TYPE	jo a dá se tam vzít kočárek *_/16 _/17 moci/18 _/19 tam/20 vzít/21 _/22*

Tabulka 3.2: Ukázka z CSC korpusu.

o očekávanou shodu:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \tag{3.1}$$

kde $P(A)$ odpovídá podílu případů skutečně pozorované shody anotátorů k počtu všech možných shod a $P(E)$ pak očekávanému podílu shod anotátorů k počtu všech možných shod (tj. jde o shodu, kterou bychom docílili náhodnou klasifikací, kdyby klasifikátor vybíral všechny možné třídy se stejnou pravděpodobností). $P(E)$ tak reprezentuje složitost klasifikační úlohy (čím je více tříd do kterých se klasifikuje, tím je menší šance, že při náhodném výběru dojde ke shodě). Důležité hodnoty κ jsou 0, když zde není jiná shoda než náhodná a 1, jestliže dojde k úplné shodě. V případě automatického překladu můžeme $P(E)$ zanedbat, neboť pravděpodobnost, že náhodně vybereme správný překlad slova nebo dokonce celé věty je rovna téměř nule (při velikosti slovníku 3 185 znaků a maximální délce věty např. 20 slov je počet možných vět roven 3185^{20}, ve skutečnosti je samozřejmě počet vět uvažovaných každým překladatelem

Číslo překladatele	1	2	3	4	Průměr
1	–	90,37	89,45	79,03	86,28
2	90,42	–	88,88	77,88	85,73
3	89,76	89,14	–	82,85	87,25
4	79,85	78,67	83,52	–	79,92
Průměr	86,68	86,06	87,28	79,92	84,99

Tabulka 3.3: Vyhodnocení shody mezi překladateli pomocí BLEU kritéria.

mnohem nižší, stále však dostatečně velký na to, abychom pravděpodobnost $P(E)$ mohli bez obav zanedbat). Za tohoto předpokladu je tedy v našem případě κ vždy rovna $P(A)$. Pro výpočet shody mezi dvěma překlady lze použít celou řadu kritérií, která se používají pro hodnocení úspěšnosti překladu automatických překladových systémů. Pro účely porovnání shody mezi překladateli jsme z této množiny vybrali čtyři následující kritéria:

- BLEU (angl. biLingual evaluation understudy) - Toto kritérium bylo navrženo v práci [Papineni 01] a je v současnosti nejpoužívanějším kritériem pro výpočet přesnosti překladu. Počítá modifikovanou n-gramovou přesnost vytvořeného překladu s ohledem na referenční překlad.
- WER (angl. word error rate) - Toto kritérium je převzato z oblasti automatického rozpoznávání řeči (angl. automatic speech recognition (ASR)) a je definováno jako Levenstheinova editační vzdálenost mezi vytvořeným překladem a referenčním překladem (jde o poměr počtu všech smazaných, nahrazených a vložených slov ve vytvořeném překladu k počtu všech slov v referenčním překladu)
- SER (angl. sentence error rate) - Toto kritérium se nabízí jako první a počítá jednoduše poměr správně přeložených vět k počtu všech překládaných vět. V oblasti SMT se obvykle nepoužívá, protože při obecném překladu je velmi malá šance na shodu celého referenčního překladu s hodnoceným překladem, v našem případě ho však lze použít dobře, neboť vzhledem k povaze korpusu (omezené téma, krátké věty, omezená množina slov) je pravděpodobnost shody značná.
- PER (angl. position-independent word error rate) - Toto kritérium vzniklo oslabením WER kritéria, kdy počítáme jen poměr počtu správně přeložených slov k počtu všech překládaných slov bez ohledu na jejich pořadí.

V případě BLEU kritéria jsou vyšší hodnoty lepší, v případě zbylých kritérií je tomu naopak. V následujících tabulkách jsou výsledky porovnání překladů mezi jednotlivými překladateli pro jednotlivá kritéria. V Tabulce 3.3 jsou výsledky pro BLEU kritérium. V Tabulce 3.4 jsou výsledky pro WER kritérium. V Tabulce 3.5 jsou výsledky pro SER kritérium. A konečně v Tabulce 3.6 jsou výsledky pro PER kritérium.

Překladatelé byli porovnáni vždy každý s každým a to tak, že jednou byl každý překlad použit jako referenční a jednou jako výstup hodnoceného překladového systému. Každý překladatel je označen číslem, 1 a 2 označuje zkušené překladatele, 3 a 4 pak nezkušené. V řádku vždy daný text slouží jako referenční, ve sloupci pak jako hodnocený.

Z tabulek je patrné, že nejvyšší průměrná shoda je u třetího překladatele, to je celkem překvapivé, neboť bychom očekávali nejlepší shodu pro překladatele 1 nebo 2. Tento výsledek je

Číslo překladatele	1	2	3	4	Průměr
1	–	5,93	5,6	9,54	7,02
2	5,9	–	6,34	10,73	7,66
3	5,59	6,36	–	7,6	6,52
4	9,3	10,53	7,42	–	9,08
Průměr	6,93	7,61	6,45	9,29	7,57

Tabulka 3.4: Vyhodnocení shody mezi překladateli pomocí WER kritéria.

Číslo překladatele	1	2	3	4	Průměr
1	–	25,11	25,86	42,86	31,28
2	25,11	–	27,37	44,36	32,28
3	25,86	27,37	–	34,89	29,37
4	42,86	44,36	34,89	–	40,7
Průměr	31,28	32,28	29,37	40,7	33,41

Tabulka 3.5: Vyhodnocení shody mezi překladateli pomocí SER kritéria.

Číslo překladatele	1	2	3	4	Průměr
1	–	3,63	4,27	7,97	5,29
2	4,97	–	5,17	9,44	6,53
3	4,64	4,13	–	6,34	5,04
4	7,77	7,92	5,83	–	7,17
Průměr	5,79	5,23	5,09	7,92	6,01

Tabulka 3.6: Vyhodnocení shody mezi překladateli pomocí PER kritéria.

dán hlavně tím, že shoda mezi překladatelem 3 a 4 je významně větší než shoda mezi překladatelem 4 a překladateli 1 nebo 2. Tabulka 3.5 je jako jediná symetrická, u zbylých záleží na tom, kdy je porovnávaný text referenční a kdy hodnocený, což je dáno definicí použitých kritérií. Pokud však spočteme průměr z průměrných hodnot pro každého překladatele, dostaneme jak pro řádky tak pro sloupce stejné číslo. Pokud tedy pohlížíme na korpus jako na celek, nezáleží na tom, kterou část použijeme pro trénování a kterou pro testování (resp. je třeba testovací část vybrat tak, aby v ní byly odpovídajícím způsobem zastoupeny překlady od všech překladatelů). Pokud porovnáme shodu mezi překladateli v jednotlivých skupinách, tak je vidět, že průměrná shoda mezi zkušenými překladateli je vždy zhruba o několik procentních bodů vyšší než mezi nezkušenými (od 1,6 pro PER až po 9,78 pro SER). Z pohledu na jednotlivé překladatele je nejmenší shoda mezi překladateli 2 a 4 a největší pak mezi překladateli 1 a 2. Z průměru průměrných shod také můžeme odhadnout, kde asi leží maximální hodnoty kritérií, kterých může dosáhnout systém natrénovaný z tohoto korpusu. V případě BLEU kritéria se lze ideálně dostat na hodnotu 85, u WER kritéria na hodnotu 7,5, u SER kritéria na hodnotu 33,5

Číslo překladatele	Počet dialogů	Počet kol	Počet CZ tokenů	Počet SCZ tokenů
1	330(29,8%)	4 591(29,1%)	31 795(29,5%)	31 437(29,2%)
2	250(22,5%)	3 554(22,5%)	24 988(23,1%)	24 952(23,2%)
3	429(38,7%)	6 066(38,5%)	41 010(38%)	40 871(38%)
4	100(9%)	1 561(9,9%)	10 160(9,4%)	10 261(9,6%)
Celkem	1 109	15 772	107 953	107 521

Tabulka 3.7: Podíl jednotlivých překladatelů na CSC korpusu.

a konečně u PER kritéria na hodnotu 6 procentních bodů. Na závěr ještě uveďme v Tabulce 3.7 podíl jednotlivých překladatelů na vytvořeném korpusu.

Kapitola 4

Výběr frází

V této kapitole nejprve popíšeme stávající metody pro výběr frází a pak vlastní metodu pro výběr frází založenou na principu minimální ztráty. Nejprve budou uvedeny metody založené na slovním přiřazení, dále metody založené na frázovém přiřazení a nakonec několik metod, které používají různé přístupy k tvorbě frázové tabulky. V případě vlastní metody pro výběr frází bude popsána základní metoda a dále její úpravy, které směřují ke zlepšení výsledků překladu a zmenšení získané tabulky.

Tabulka frází je jednou z klíčových komponent frázového překladového systému. Jsou v ní uloženy informace o zdrojových frázích a jim příslušejících překladech - cílových frázích. Díky použití frází je zde také uložena informace o lokálním uspořádání cílových slov při překladu zdrojové fráze. Tím, že danou zdrojovou frázi přeložíme pomocí příslušné cílové fráze, zjistíme zároveň, jak mají být tato cílová slova uspořádána ve výsledném překladu (pořadí těchto slov je jednoduše zachyceno v přiřazené cílové frázi, která je přidána k výslednému překladu). V kapitole 1.5.2 jsme frázi definovali jako neprázdnou souvislou řadu slov. Pro daný pár vět (s_1^J, t_1^I) a dané rozdělení $\mathbf{a} = a_1^K$, kde každé $a_k = (i_k; b_k, j_k)$ je trojice, která představuje konec i_k k-té cílové fráze $\tilde{t}_k$ a začátek a konec (b_k, j_k) zdrojové fráze $\tilde{s}_k$, definujeme dvojjazyčné fráze takto:

$$\tilde{t}_k := t_{i_{k-1}+1} ... t_{i_k} \tag{4.1}$$

$$\tilde{s}_k := s_{b_k} ... s_{j_k}. \tag{4.2}$$

Při tvorbě frázové tabulky je naším cílem získat z paralelního korpusu odpovídající si páry frází, tzv. *překladové páry*. Základní metody pro výběr frází jsou založeny na slovním přiřazení. V následující části tedy popíšeme metody používané pro modelování slovního přiřazení mezi zdrojovým a cílovým textem.

4.1 Slovní přiřazení

V kapitole 1.5.1 o slovním překladu jsme zavedli pojem slovní přiřazení a ukázali jeho zapojení jako skryté proměnné do rovnice 1.14 pro překladový model:

$$Pr(\mathbf{s}|\mathbf{t}) = \sum_{\mathbf{a}} Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}),$$

kterou lze dále rozepsat jako rovnici 1.15:

$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) = Pr(J|\mathbf{t}) \cdot \prod_{j=1}^{J} Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}) \cdot Pr(s_j|a_1^j, s_1^{j-1}, J, \mathbf{t}),$$

kde $\mathbf{s} = s_1^J = s_1...s_j...s_J$, $\mathbf{t} = t_1^I = t_1...t_i...t_I$ a $\mathbf{a} = a_1^J = a_1...a_j...a_J, a_j \in \{0, ..., I\}$. Jak již bylo řečeno v kapitole 1.5.1, rovnice 1.15 je přesným vyjádřením pravděpodobnosti $Pr(\mathbf{s}, \mathbf{a}|\mathbf{t})$. Pravá strana rovnice však obsahuje příliš mnoho nezávislých parametrů, které by bylo třeba získat z trénovacích dat. V základní práci o SMT [Brown 93] bylo proto navrženo pět modelů (v literatuře jsou nazývány jako IBM nebo Model 1 – 5), které představují různé aproximace rovnice 1.15. Modely jsou řazeny vzestupně od jednodušších, které obsahují méně parametrů a jsou jednoduše trénovatelné, až po složité modely s více parametry a složitějším trénováním (při trénování modelu je třeba použít aproximace, neboť nelze nalézt optimální řešení v přijatelném čase). Popišme si nyní blíže tyto modely.

Prvním modelem je Model 1, který je definován následujícím vztahem:

$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) = \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} p_t(s_j|t_{a_j}), \tag{4.3}$$

kde $\epsilon \equiv Pr(J|\mathbf{t})$ (předpokládáme, že $Pr(J|\mathbf{t})$ je nezávislé na $\mathbf{t}$ a J), $Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}) \equiv (I+1)^{-1}$ (neboť $Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t})$ závisí jen na I) a překladová pravděpodobnost $p_t(s_j|t_{a_j}) \equiv Pr(s_j|a_1^j, s_1^{j-1}, J, \mathbf{t})$ ($(Pr(s_j|a_1^j, s_1^{j-1}, J, \mathbf{t}))$ závisí jen na s_j a t_{a_j}). Slovní přiřazení získáme, jestliže budou určeny hodnoty a_j pro j od 1 do J, z nichž každá může nabývat hodnoty od 0 do I. Tedy:

$$Pr(\mathbf{s}|\mathbf{t}) = \frac{\epsilon}{(I+1)^J} \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \prod_{j=1}^{J} p_t(s_j|t_{a_j}). \tag{4.4}$$

Naším cílem je nalézt překladové pravděpodobnosti maximalizující $Pr(\mathbf{s}|\mathbf{t})$ za omezující podmínky, že pro každé t platí:

$$\sum_s p_t(s|t) = 1. \tag{4.5}$$

Podle standardního postupu pro maximalizaci za omezujících podmínek zavedeme Lagrangeovy multiplikátory λ_t a budeme hledat maximum následující pomocné funkce:

$$h(p_t, \lambda) \equiv \frac{\epsilon}{(I+1)^J} \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \prod_{j=1}^{J} p_t(s_j|t_{a_j}) - \sum_t \lambda_t \cdot (\sum_s p_t(s|t) - 1). \tag{4.6}$$

Extrém této funkce dostaneme, jestliže parciální derivace výrazu h podle proměnných p_t a λ položíme rovny nule. Jestliže položíme rovnu nule parciální derivaci podle λ, dostaneme opět rovnici omezujících podmínek pro překladové pravděpodobnosti p_t. Položíme-li rovnu nule parciální derivaci podle p_t dostaneme rovnici:

$$p_t(s|t) = \lambda_t^{-1} \frac{\epsilon}{(I+1)^J} \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \sum_{j=1}^{J} \delta(s, s_j) \cdot \delta(t, t_{a_j}) \prod_{k=1}^{J} p_t(s_k|t_{a_k}), \tag{4.7}$$

kde δ je Kroneckerova delta funkce, která nabývá jedničky, jestliže jsou oba argumenty shodné, jinak je nulová. Tato rovnice tak představuje iterativní postup pro nalezení překladových pravděpodobností. Pro daný počáteční odhad hodnot překladových pravděpodobností můžeme spočítat pravou stranu Rovnice 4.7 a dostat tak nový odhad hledaných hodnot, který opět můžeme použít pro výpočet nových hodnot překladových pravděpodobností. Tento postup je jednou z instancí obecného iterativního algoritmu pro výpočet parametrů pravděpodobnostního modelu závislého na nepozorovatelné, skryté proměnné a nazývá se EM (angl. expectation maximization) algoritmus (použití tohoto algoritmu pro odhad pravděpodobností statistického modelu a důkaz o jeho konvergenci viz práce [Baum 72] a [Dempster 77]). Obecně se skládá ze

dvou kroků. Prvním krokem je očekávání (angl. expectation), kdy jsou na základě současných hodnot hledaných parametrů spočteny očekávané hodnoty pravděpodobností pro množinu trénovacích dat. Druhým krokem je pak maximalizace (angl. maximization), kdy jsou na základě očekávaných hodnot pravděpodobností získaných v předešlém kroku, spočteny nové hodnoty hledaných parametrů. Tento postup se opakuje, dokud není dosaženo ustálené hodnoty parametrů. Odhad hodnot konverguje k jedinému lokálnímu maximu bez ohledu na jejich počáteční hodnotu (důkaz a bližší podrobnosti o výpočtu parametrů Modelu 1 viz [Brown 93]).

Dalším modelem je Model 2, který zohledňuje přiřazení mezi jednotlivými zdrojovými a cílovými slovy. Model 2 vychází ze stejných předpokladů jako Model 1 a přidává navíc přiřazovací pravděpodobnosti $a(a_j|j, J, I)$, když předpokládáme, že $Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t})$ závisí na j, a_j a J a I tedy:

$$a(a_j|j, J, I) \equiv Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}). \tag{4.8}$$

Pro přiřazovací pravděpodobnosti platí následující omezující podmínka:

$$\sum_{i=0}^{I} a(i|j, J, I) = 1 \tag{4.9}$$

pro každou trojici (j, J, I). Dosadíme-li pravděpodobnosti $a(a_j|j, J, I)$ do Rovnice 4.4 dostaneme:

$$Pr(\mathbf{s}|\mathbf{t}) = \epsilon \cdot \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \prod_{j=1}^{J} p_t(s_j|t_{a_j}) \cdot a(a_j|j, J, I). \tag{4.10}$$

Stejně jako v případě Modelu 1 zavedeme při hledání maxima této funkce Lagrangeovy multiplikátory a budeme hledat maximum pomocné funkce:

$$\begin{aligned} h(p_t, a, \lambda, \mu) \equiv \epsilon \cdot \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \prod_{j=1}^{J} p_t(s_j|t_{a_j}) \cdot a(a_j|j, J, I) \\ - \sum_t \lambda_t \cdot (\sum_s p_t(s|t) - 1) - \sum_j \mu_{jJI} \cdot (\sum_i a(i|j, J, I) - 1). \end{aligned} \tag{4.11}$$

Položením parciálních derivací výrazu h podle jednotlivých proměnných rovným nule opět dostaneme iterativní postup pro výpočet odhadovaných hodnot parametrů modelu. Oproti Modelu 1 je třeba navíc počítat hodnoty přiřazovacích pravděpodobností (podrobný popis pro výpočet parametrů Modelu 2 viz [Brown 93]).

Modely 3, 4 a 5 popsané dále vychází z rovnice:

$$\begin{aligned} Pr(\tau, \pi|\mathbf{t}) = & \prod Pr(\phi_i|\phi_1^{i-1}, \mathbf{t}) \cdot Pr(\phi_0|\phi_1^I, \mathbf{t} \times \\ & \prod_{i=0}^{I} \prod_{k=1}^{\phi_i} Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{t}) \times \\ & \prod_{i=1}^{I} \prod_{k=1}^{\phi_i} Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{t}) \times \\ & \prod_{k=1}^{\phi_0} Pr(\pi_{0k}|\pi_{01}^{k-1}, \pi_1^I, \tau_0^I, \phi_0^I, \mathbf{t}), \end{aligned} \tag{4.12}$$

kde proměnná ϕ_i představuje *fertilitu* každého cílového slova t (nebo-li popisuje počet zdrojových slov, který může být přiřazen k danému t), τ_i pak představuje různé seznamy zdrojových slov přiřazených cílovému slovu t a nakonec π_i popisuje uspořádání zdrojových slov v seznamu

τ_i tak, abychom dostali zdrojovou větu **s**. Znalost τ a π jednoznačně určuje zdrojovou větu a přiřazení mezi ní a cílovou větou. Obecně různé páry τ, π mohou určovat ten samý pár **s**, **a**, což lze vyjádřit rovnicí:

$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) = \sum_{(\tau,\pi)\in<\mathbf{s},\mathbf{a}>} Pr(\tau, \pi|\mathbf{t}), \tag{4.13}$$

kde $< \mathbf{s}, \mathbf{a} >$ představuje množinu dvojic τ, π, které vedou na stejný pár **s**, **a**. Rovnice 4.13 tedy představuje alternativní vyjádření Rovnice 1.15.

Model 3 vychází z Rovnice 4.12 a je založen na předpokladu, že pro i mezi 1 a I závisí $Pr(\phi_i|\phi_1^{i-1}, \mathbf{t})$ jen na ϕ_i a t_i a dále že, pro všechna i závisí $Pr(\tau_{ik}|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{t})$ jen na τ_{ik} a t_i a nakonec že, pro i mezi 1 a I závisí $Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{t})$ jen na π_{ik}, i, J a I. Parametry Modelu 3 jsou tedy množina fertilních pravděpodobností $n(\phi|t_i) \equiv Pr(\phi|\phi_1^{i-1}, \mathbf{t})$, množina překladových pravděpodobností $p_t(s|t_i) \equiv Pr(\tau_{ik} = s|\tau_{i1}^{k-1}, \tau_0^{i-1}, \phi_0^I, \mathbf{t})$ a množina distorzních pravděpodobností $d(j|i, J, I) \equiv Pr(\pi_{ik}|\pi_{i1}^{k-1}, \pi_1^{i-1}, \tau_0^I, \phi_0^I, \mathbf{t})$. Distorzní a fertilní pravděpodobnosti pro t_0 jsou pak definovány tak, že příspěvek distorzních pravděpodobností všech slov v τ_0 je roven $1/\phi_0!$ a ϕ_0 závisí jen na I, tedy:

$$Pr(\phi_0|\phi_1^I, \mathbf{t}) = \begin{pmatrix} \phi_1 + ... + \phi_I \\ \phi_0 \end{pmatrix} p_0^{\phi_1+...+\phi_I-\phi_0} \cdot p_1^{\phi_0}, \tag{4.14}$$

kde p_0 a p_1 jsou pomocné proměnné, které sčítají do jedničky (předpokládáme, že každé slovo z τ_1^I vyžaduje navíc extra slovo s pravděpodobností p_1, které je přiřazeno k prázdné pozici t_0). Pravděpodobnost, že přesně ϕ_0 slov z τ_1^I bude potřebovat extra slovo je vyjádřena právě předchozí rovnicí 4.14. Stejně jako v případě Modelu 1 a 2 je přiřazení mezi zdrojovou a cílovou větou definováno, jestliže jsou určeny všechna a_j. Fertility ϕ_0 až ϕ_I jsou funkcí a_j, neboť ϕ_i se rovná počtu j pro která je a_j rovno i. Model 3 je tedy definován vztahem:

$$Pr(\mathbf{s}|\mathbf{t}) = \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) \tag{4.15}$$

$$= \sum_{a_1=0}^{I} ... \sum_{a_J=0}^{I} \begin{pmatrix} J - \phi_0 \\ \phi_0 \end{pmatrix} p_0^{J-2\phi_0} \cdot p_1^{\phi_0} \prod_{i=1}^{I} \phi_i! \cdot n(\phi_i|t_i) \times \prod_{j=1}^{J} p_t(s_j|t_{a_j}) \cdot d(j|a_j, J, I) \tag{4.16}$$

s omezujícími podmínkami:

$$\sum_s p_t(s|t) = 1 \quad \sum_j d(j|i, J, I) = 1 \quad \sum_\phi n(\phi|t) = 1 \quad p_0 + p_1 = 1.$$

Hledáme tedy maximum následující pomocné funkce:

$$\begin{aligned} h(p_t, d, n, p, \lambda, \mu, \nu, \epsilon) =& Pr(\mathbf{s}|\mathbf{t}) - \sum_t \lambda_t \cdot (\sum_s p_t(s|t) - 1) - \sum_i \mu_{iJI} \cdot (\sum_j d(j|i, J, I) - 1) \\ & - \sum_t \nu_t \cdot (\sum_\phi n(\phi|t) - 1) - \epsilon \cdot (p_0 + p_1 - 1). \end{aligned} \tag{4.17}$$

Položením parciálních derivací výrazu h podle jednotlivých proměnných rovným nule opět dostaneme iterativní postup pro výpočet odhadovaných hodnot parametrů modelu. Na rozdíl od Modelu 1 a 2 je nutné použít při výpočtu parametrů aproximaci sumy přes všechna možná přiřazení, neboť není znám, na rozdíl od Modelu 1 a 2, efektivní algoritmus pro výpočet

této sumy. V praxi se tato suma nahrazuje množinou velmi pravděpodobných přiřazení (ta je tvořena nejpravděpodobnějším přiřazením, které můžeme nalézt a dále množinou přiřazení, která vzniknou malou změnou tohoto přiřazení). Nejpravděpodobnější přiřazení se nazývá Viterbiho přiřazení. Podrobný popis pro výpočet parametrů Modelu 3 a aproximace sumy všech přiřazení viz [Brown 93].

Model 4 je úpravou Modelu 3 založenou na rozdělení distorzní pravděpodobnosti na dva další parametry, jeden pro umístění hlavních slov (toto rozdělení vychází z myšlenky, že jestliže překládáme nějakou frázi, tak nejdříve umístíme hlavní slovo fráze a pak zbytek slov) a jeden pro umístění zbývajících slov. Nové slovo je závislé na minule přiřazeném slově a na slovních třídách okolních slov. Výpočet je pak obdobný jako v případě Modelu 3. Model 5 pak řeší nedostatečnost Modelu 3 a 4 vhodnou úpravou přiřazovacích pravděpodobností. Tato nedostatečnost vzniká tím, že oba modely rezervují část pravděpodobnosti i pro případy, které nemohou nastat (to je tím, že současné přiřazení je nezávislé na minulém přiřazení a mohou tak vznikat tzv. zobecněné řetězce, které mají některé pozice obsazené více slovy zatímco jiné pozice jsou neobsazené). Kompletní popis všech modelů a experimentů s nimi lze nalézt ve zmíněné práci [Brown 93].

Dalším používaným přiřazovacím modelem je HMM (skrytý Markovův model, angl. hidden Markov model) model, který je popsán v práci [Vogel 96]. Ten opět vychází z rovnice 1.15 a je dán následujícím vztahem:

$$Pr(\mathbf{s}|\mathbf{t}) = p(J|I) \cdot \sum_{a_1^J} \prod_{j=1}^{J} a(a_j|a_{j-1}, I) \cdot p_t(s_j|t_{a_j}), \tag{4.18}$$

kde jsou přiřazovací a překladová pravděpodobnost definovány jako:

$$Pr(a_j|a_1^{j-1}, s_1^{j-1}, J, \mathbf{t}) \equiv a(a_j|a_{j-1}, I) \tag{4.19}$$

$$Pr(s_j|a_1^{j}, s_1^{j-1}, J, \mathbf{t}) \equiv p_t(s_j|t_{a_j}). \tag{4.20}$$

Dále předpokládáme, že přiřazovací pravděpodobnosti $a(a_j|a_{j-1}, I) = a(i|i', I)$ závisí jen na relativní změně přiřazení, tj. na $a_j - a_{j-1}$. Jestliže použijeme množinu nezáporných parametrů $\{r(a_j - a_{j-1})\}$, můžeme přiřazovací pravděpodobnosti zapsat ve tvaru:

$$a(a_j|a_{j-1}, I) = \frac{r(a_j - a_{j-1})}{\sum_{l=1}^{I} r(a_l - a_{j-1})}. \tag{4.21}$$

Tato definice nám zaručí, že pro každou slovní pozici i', $i' = 1, ..., I$ splňují přiřazovací pravděpodobnosti normalizační podmínku. Stejně jako v případě Modelu 3 a vyšších je při výpočtu parametrů modelu použita aproximace sumy přiřazení pomocí nejpravděpodobnějšího přiřazení.

Práce [Och 03b] pak přináší srovnání výše uvedených modelů (Model 1 – 5, HMM model), několika heuristických modelů (ty jsou založené na funkci, která počítá podobnost mezi dvěma jazyky) a nově navrženého modelu, který se nazývá Model 6. Model 6 je definován jako log-lineární kombinace předchozích modelů, obecně tedy:

$$Pr(\mathbf{s}, \mathbf{a}|\mathbf{t}) = \frac{\prod_{k=1}^{K} Pr_k(\mathbf{s}, \mathbf{a}|\mathbf{t})^{\alpha_k}}{\sum_{\mathbf{s}', \mathbf{a}'} \prod_{k=1}^{K} Pr_k(\mathbf{s}', \mathbf{a}'|\mathbf{t})^{\alpha_k}}, \tag{4.22}$$

kde $k = 1, ..., K$ označuje použité modely. V práci je konkrétně použita kombinace HMM modelu a Modelu 4. Modely jsou srovnávány na různých úlohách a z různých hledisek jako

je doba trénování, přesnost přiřazení a počet výsledných parametrů. Pro výpočet přesnosti přiřazení je v této práci použito kritérium míry chyb přiřazení (angl. alignment error rate (AER)), které počítá rozdíl mezi referenčním přiřazením vytvořeným člověkem a hodnoceným přiřazením A, tedy:

$$AER(S, P, A) = 1 - \frac{|A \cap S| + |A \cup P|}{|A| + |S|}, \tag{4.23}$$

kde S je množina jistých přiřazení zvolených anotátorem a P pak množina možných přiřazení. Při trénování modelů jsou použita také různá trénovací schémata, kdy jsou hodnoty parametrů nižších modelů použity pro inicializaci výpočtu parametrů vyšších modelů (platí toto uspořádání: Model 1 < Model 2 = HMM model < Model 3 < Model 4 < Model 5 < Model 6, při trénování vyššího modelu, lze vynechat některý nižší model, např. Model 6 se trénuje rovnou z parametrů Modelu 4). Z porovnání výsledků modelů vyplývá, že statistické modely jsou vždy lepší než heuristické. Nejlepší výsledky z hlediska přesnosti přiřazení pak byly dosaženy při použití Modelu 6.

Dalším přístupem použitelným pro modelování slovního přiřazení jsou diskriminativní modely. V roce 2005 se objevila nezávisle na sobě řada prací [Liu 05, Taskar 05, Moore 05, Ittycheriah 05, Fraser 05], které popisují použití diskriminativního modelu pro slovní přiřazení v různých úlohách. V těchto pracích jsou použity různé kombinace modelů a rysů v diskriminativním modelu a různé algoritmy pro trénování diskriminativního modelu. V práci [Liu 05] je použit log-lineární model se třemi rysy (hodnoty IBM modelu 3, POS (angl. part of speech) značky a položky dvojjazyčného slovníku), který je natrénován pomocí GIS algoritmu. V práci [Moore 05] jsou pak představeny tři nové rysy pro popis slovního přiřazení a je popsáno trénování log-lineárního modelu pomocí modifikované verze učení pomocí zprůměrovaných perceptronů [Collins 02]. Odlišný přístup je zvolen v práci [Taskar 05], kde je na problém slovního přiřazení pohlíženo jako na problém maximálního váženého bipartitního párování (jedná se o problém z teorie grafů, kdy je naším cílem nalézt maximálně ohodnocené spojení všech uzlů bipartitního[1] grafu za daných omezení). To vede na diskriminativní model, který je trénován pomocí algoritmu založeného na výpočtu minimálního toku ze zdroje. Práce [Liang 06, Lacoste-Julien 06, Moore 06] pak dále rozvíjí diskriminativní přistup k problému slovního přiřazení. Jedná se především o použití dalších modelů a rysů pro popis přiřazení a modifikaci nebo použití nových trénovacích algoritmů a upravených omezení kladených na použité modely. Výsledkem jsou nejlepší dosažené výsledky pro slovní přiřazení v konkrétních úlohách (poznamenejme, že nejlepších výsledků je vždy dosaženo použitím hodnot IBM modelu (obvykle 4) jako jednoho z rysů v diskriminativním modelu).

Problémem slovního přiřazení se zabývá i řada dalších prací, za všechny jmenujme jednu z posledních. Jde o práci [Zens 08], která přináší další zlepšení standardních IBM přiřazovacích modelů a překonává tak výsledky publikované v práci [Och 03b]. Tohoto zlepšení je dosaženo použitím dvou technik, které se snaží řešit hlavní omezení vyplývající z definice standardních IBM modelů a sice, že každé zdrojové slovo může být přiřazeno nejvýše jednomu cílovému slovu. Za prvé je použito simultánní trénování v obou směrech překladu, jsou tak využity znalosti získané z obou směrů. Je použita lineární i log-lineární kombinace těchto znalostí. Za druhé je použit algoritmus pro přiřazení mezi slovy, který neklade žádná omezení na toto přiřazení. Tento algoritmus řeší problém přiřazení jako problém nalezení hranového pokrytí s minimální náklady (jako náklady jsou použity parametry IBM modelů) v bipartitním grafu. Výhodou tohoto přístupu je také to, že může být efektivně nalezeno globální optimální řešení tohoto problému na rozdíl např. od Modelu 4, kde není znám efektivní algoritmus pro nale-

[1] Pojmem bipartitní graf se v teorii grafů označuje takový graf, jehož množinu vrcholů je možné rozdělit na dvě disjunktní množiny tak, že žádné dva vrcholy ze stejné množiny nejsou spojeny hranou: zdroj www.wikipedia.cz

zení globálního optima. Stejně jako v pracích [Ayan 06, Vilar 06, Fraser 07] je i v této práci konstatováno, že zlepšení přesnosti přiřazení měřené pomocí AER kritéria nemusí vést ke zlepšení přesnosti (což v tomto případě také nevedlo) překladu systému využívajícího frázovou tabulku vytvořenou na základě tohoto přiřazení. Jak konstatují všechny tři zmíněné práce, hlavní příčinnou je nízká korelace mezi AER a BLEU kritériem používaným pro hodnocení kvality vytvořených překladů. Proto jsou v těchto pracích navrženy i další kritéria pro hodnocení kvality přiřazení, žádné z nich však neřeší problém korelace dostatečně. Jako hlavní závěr tak lze z těchto poznatků vyvodit, že je důležité porovnat kvalitu jednotlivých přiřazovacích modelů prostřednictvím jejich zařazení do identického překladového systému a vyhodnocením kvality takto získaného překladu např. pomocí BLEU kritéria.

4.1.1 Výběr frází ze slovního přiřazení

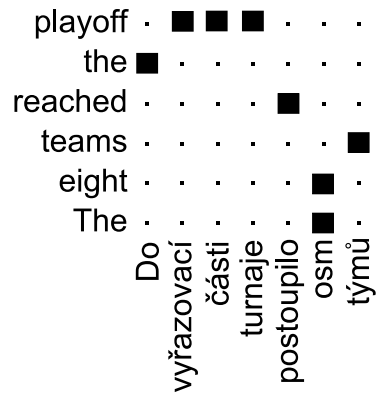
Jestliže máme vytvořeno slovní přiřazení mezi slovy ve zdrojové a cílové větě, můžeme z něj vybrat fráze, které uložíme do frázové tabulky pro pozdější použití ve vytvořeném překladovém systému. Jako základní metoda pro výběr frází se používá postup popsaný v práci [Och 02a] nebo [Zens 08] (poprvé byl tento postup navržen v práci [Och 99]). Nejprve jsou natrénovány slovní přiřazovací modely a spočteno Viterbiho přiřazení trénovacího korpusu pro oba směry překladu. Poté jsou obě získaná přiřazení zkombinována pomocí vhodných heuristik tak, abychom dostali symetrické přiřazení. Cílem symetrického přiřazení je dostat přiřazení, které umožňuje přiřazení jeden k více mezi slovy a tím zlepšit kvalitu výsledného přiřazení. Jestliže tedy máme dány dvě přiřazení $A_1 = \{(a_j, j)|a_j > 0\}$ a $A_2 = \{(i, b_i)|b_i > 0\}$, lze z nich symetrické přiřazení A vytvořit použitím následujících heuristik navržených v práci [Och 02a]:

- Průnik: $A = A_1 \cap A2$.
- Sjednocení: $A = A_1 \cup A2$.
- Propracovanější metoda: Nejprve provedeme průnik jednotlivých přiřazení $A = A_1 \cap A2$. Pak takto získané přiřazení A dále rozšiřujeme o přiřazení (i, j), které se objevuje jen v přiřazení A_1 nebo v přiřazení A_2 jestliže buď s_j nebo t_i má přiřazení v A, nebo platí obě následující podmínky:
  - přiřazení (i, j) má horizontálního souseda $(i - 1, j)$, $(i + 1, j)$ nebo vertikálního souseda $(i, j - 1)$, $(i, j + 1)$, který je již v A.
  - množina $A \cup \{(i, j)\}$ neobsahuje přiřazení s oběma horizontálními a vertikálními sousedy.

Množina dvojjazyčných frází $\mathcal{BP}$ větného páru (s_1^J, t_1^I) mezi jehož slovy platí přiřazení $A \subseteq J \times I$ je definována kritériem [Zens 08]:

$$\begin{aligned}\mathcal{BP}(s_1^J, t_1^I, A) = \Big\{(s_{j_1}^{j_2}, t_{i_1}^{i_2}) :&\forall (j,i) \in A : j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i_2 \\ &\wedge \exists (j,i) \in A : j_1 \leq j \leq j_2 \wedge i_1 \leq i \leq i_2\Big\}.\end{aligned} \tag{4.24}$$

To znamená, že dvě fráze jsou považovány za vzájemné překlady jen v případě, že slova těchto frází jsou přiřazena jen mezi sebou a ne žádnému vnějšímu slovu, fráze dále musí být také spojité. Na Obrázku 4.1 je v levé části symetrické přiřazení mezi dvojicí vět a v pravé pak seznam frází extrahovaných z tohoto větného páru podle uvedeného kritéria (jsou vybrány fráze až do délky sedm).



Zdrojové fráze	Cílové fráze
Do	the
Do vyřazovací části turnaje	the playoff
Do vyřazovací části turnaje postoupilo	reached the playoff
Do vyřazovací části turnaje postoupilo osm týmů	The eight teams reached the playoff
vyřazovací části turnaje	playoff
vyřazovací části turnaje postoupilo	playoff reached
vyřazovací části turnaje postoupilo osm týmů	eight teams reached the playoff
postoupilo	reached
postoupilo osm týmů	The eight teams reached
osm	The eight
osm týmů	The eight teams
týmů	teams

Obrázek 4.1: Symetrické slovní přiřazení páru vět a fráze extrahované z tohoto přiřazení.

Další prací, která se zabývá výběrem frází ze slovního přiřazení, je práce [Venugopal 03]. Nejprve jsou určeny všechny zdrojové n-gramy do dané délky n. Na základě vytvořeného slovního přiřazení je pak ke každému zdrojovému n-gramu vybrána množina možných cílových překladů. Při výběru cílových překladů délky 1 až I je použito posuvné okénko s posunem 0 až $I - 1$. Výsledná množina překladů je ohodnocena a prořezána tak, aby byla zohledněna relativní důvěra jednotlivých kandidátů a vyloučeny falešné překlady, které vznikají díky použití posuvného okénka. Důvěra v každého kandidáta je odhadnuta pomocí hodnot trojice modelů: odhad na základě nejpravděpodobnějšího slovního přiřazení, odhad pomocí slovního překladu a jazykově specifický model. Při prořezání množiny kandidátů je použito kritérium maximální separace. Kandidátské překlady jsou seřazeny podle dosaženého skóre a na základě kritéria maximální separace, které maximalizuje rozdíl středních hodnot rozdělení, je zvolen bod, kde budou odděleny korektní a nekorektní překlady.

Práce [Tillmann 03] také vychází ze slovního přiřazení. Je použit HMM model pro oba směry překladu a průnikem takto vytvořených jednosměrných přiřazení je získáno výsledné symetrické přiřazení. Toto symetrické přiřazení pak slouží pro výběr bloků, z kterých jsou vytvářeny páry vět v trénovacím korpusu (předpokládáme, že daný pár vět můžeme vytvořit pomocí sekvence bloků, kde blok je tvořen zdrojovou frází a jí odpovídajícím překladem). Při výběru bloků se uplatňují i body přiřazení, které vzniknou spojením jednosměrných přiřazení. Takto získané bloky (fráze) jsou pak použity v překladovém systému.

4.2 Frázové přiřazení

Stejně jako lze vytvořit přiřazení mezi slovy, můžeme vytvořit přiřazení rovnou mezi frázemi zdrojové a cílové věty, a tím získat fráze vhodné pro použití v překladovém systému. Tento postup byl poprvé popsán v práci [Marcu 02]. Podobně jako v práci [Brown 93] je zde definován model pro přiřazení mezi zdrojovými a cílovými frázemi. Tento model popisuje, jak mohou být zdrojová a cílová věta vytvořeny současně, tj. tento model popisuje sdruženou pravděpodobnost

jednotlivých překladových párů.

V práci jsou představeny dva modely. Prvním je Model 1, který je založen na předpokladu, že každý větný pár v korpusu byl vytvořen následujícím stochastickým procesem:

1. Vytvoř množinu konceptů C.
2. Pro každý koncept $c_i \in C$, vytvoř pár frází $(\tilde{s}_i, \tilde{t}_i)$ v souladu s rozdělením $p_t(\tilde{s}_i, \tilde{t}_i)$, kde $\tilde{s}_i$ a $\tilde{t}_i$ každý obsahuje aspoň jedno slovo.
3. Uspořádej vytvořené zdrojové a cílové fráze tak, aby vytvořily lineární posloupnosti frází, které odpovídají párům vět v korpusu.

Pro jednoduchost předpokládáme, že množina konceptů a uspořádání vytvořených frází jsou modelovány uniformním rozdělením. Nepředpokládáme, že c_i je skrytá proměnná, která vytváří pár $(\tilde{s}_i, \tilde{t}_i)$, ale raději, že $c_i = (\tilde{s}_i, \tilde{t}_i)$. Za těchto předpokladů je pravděpodobnost vytvoření páru vět $(\mathbf{s}, \mathbf{t})$ použitím konceptů $c_i \in C$ dána součinem všech frázových překladových pravděpodobností $\prod_{c \in C} p_t(\tilde{s}_i, \tilde{t}_i)$ tedy:

$$Pr(\mathbf{s}, \mathbf{t}) = \sum_{C \in \mathcal{C} | L(\mathbf{s}, \mathbf{t}, C)} \prod_{c_i \in C} p_t(\tilde{s}_i, \tilde{t}_i), \tag{4.25}$$

kde $L(\mathbf{s}, \mathbf{t}, C)$ znamená, že množina konceptů C může být linearizována na pár vět $(\mathbf{s}, \mathbf{t})$, tj. věty $\mathbf{s}$ a $\mathbf{t}$ mohou být vytvořeny permutací frází $\tilde{s}_i$ a $\tilde{t}_i$, které charakterizují všechny koncepty $c_i \in C$ a $C \in \mathcal{C}$ je pak množina všech konceptů C takových, že mohou být linearizovány na $(\mathbf{s}, \mathbf{t})$.

Protože Model 1 je nevhodný pro překlad neviděných vět, neboť neklade žádná omezení na uspořádání frází spojených s danými koncepty, byl mírně modifikován generativní proces Modelu 1 tak, aby bylo vzato v úvahu také uspořádání frází. Model 2 je tak založen na tomto generativním procesu:

1. Vytvoř množinu konceptů C.
2. Inicializuj $\mathbf{s}$ a $\mathbf{t}$ jako prázdné posloupnosti ϵ.
3. Náhodně vyber koncept $c_i \in C$ a vytvoř pár frází $(\tilde{s}_i, \tilde{t}_i)$ v souladu s rozdělením $p_t(\tilde{s}_i, \tilde{t}_i)$, kde $\tilde{s}_i$ a $\tilde{t}_i$ každý obsahuje aspoň jedno slovo. Vyjmi c_i z C.
4. Přidej frázi $\tilde{s}_i$ na konec $\mathbf{s}$, k je počáteční pozice $\tilde{s}_i$ v $\mathbf{s}$.
5. Přidej frázi $\tilde{t}_i$ na pozici l v $\mathbf{t}$ tak, že žádná jiná fráze není na pozici mezi l a $l + |\tilde{t}_i|$, kde $|\tilde{t}_i|$ je délka fráze $\tilde{t}_i$. Tím jsme vytvořili přiřazení mezi frázemi $\tilde{s}_i$ a $\tilde{t}_i$ s pravděpodobností:

$$\prod_{p=k}^{k+|\tilde{s}_i|} d(p, (l + |\tilde{t}_i|)/2).$$

6. Opakuj krok 3 až 5 dokud není C prázdné.

Pravděpodobnost vytvoření páru vět $(\mathbf{s}, \mathbf{t})$ je pak dána výrazem:

$$Pr(\mathbf{s}, \mathbf{t}) = \sum_{C \in \mathcal{C} | L(\mathbf{s}, \mathbf{t}, C)} \prod_{c_i \in C} p_t(\tilde{s}_i, \tilde{t}_i) \times \prod_{k=1}^{|\tilde{t}_i|} d(pos(\tilde{s}_i^k), pos_{cm}(\tilde{t}_i)), \tag{4.26}$$

kde $pos(\tilde{s}_i^k)$ znamená pozici slova k fráze $\tilde{s}_i$ ve větě $\mathbf{s}$ a $pos_{cm}(\tilde{t}_i)$ pozici středu fráze $\tilde{t}_i$ ve větě $\mathbf{t}$. Pro trénování těchto modelů je opět použit EM algoritmus. Stejně jako v případě IBM modelu 3 je i trénování těchto modelů výpočetně velmi náročné (exponenciální počet přiřazení, která mohou vytvořit pár vět $(\mathbf{s}, \mathbf{t})$), je tedy třeba použít nějaká omezení a zjednodušení, abychom mohli spočítat parametry těchto modelů. Za prvé je množina frází, které jsou uvažovány při trénování, omezena jen na často se vyskytující n-gramy (konkrétně byly v článku uvažovány jen n-gramy, které se v trénovacím korpusu vyskytly aspoň pět krát). Při výpočtu EM algoritmu jsou pak opět, jako v případě IBM modelu 3, uvažovány jen velmi pravděpodobná přiřazení (tj. Viterbiho přiřazení a množina přiřazení, které vzniknou změnou tohoto přiřazení). Pro inicializaci překladových pravděpodobností p_t jsou použita Sterlingova čísla druhého řádu, která jsou aproximací skutečných pravděpodobností. Na konci trénování dostaneme rozdělení sdružené pravděpodobnosti p_t, které můžeme marginalizovat, abychom dostali podmíněné pravděpodobnosti $p_t(\tilde{s}_i|\tilde{t}_i)$ a $d(pos[\mathbf{s}]|pos[\mathbf{t}])$. Pro podrobnější popis viz [Marcu 02].

Práce [Birch 06] se dále zabývá trénováním předešlých frázových přiřazovacích modelů především z hlediska výpočetní náročnosti. V práci je použito omezení množiny prohledávaných frázových přiřazení pomocí vysoce pravděpodobných slovních přiřazení definovaných pro danou větu. Při prohledávání jsou tak použity jen fráze, které jsou konzistentní s těmito slovními přiřazeními. Tato slovní přiřazení jsou získána ze symetrického přiřazení, které vznikne průnikem jednosměrných Viterbiho přiřazení.

Další prací, která se také zabývá frázovým přiřazením je práce [Denero 06]. Autoři se zabývají tím, proč fráze získané ze slovního přiřazení heuristickou metodou popsanou výše (viz část 4.1.1) dosahují lepších výsledků než fráze získané pomocí modelů pro frázové přiřazení. Podle autorů je to tím, že zatímco v případě modelů pro slovní přiřazení vede zavedení reestimace parametrů k dobrým výsledkům, v případě frázového přiřazení to neplatí. Reestimace totiž zavádí do procesu učení soupeření a v případě slovního přiřazení tak podporuje výběr jednoho nejlepšího překladu a jemu odpovídajícího přiřazení, další přiřazení se tak již dál neuvažují. Tento efekt však v případě frázového přiřazení nepřevažuje, protože frázové přiřazovací modely, jako např. model navržený v práci [Marcu 02], obsahují prvek rozdělení (jde o rozdělení zdrojové a cílové věty na fráze). Takže zatímco je oprávněné předpokládat, že jestliže je jedno přiřazení správné, ostatní jsou chybná, je situace v případě rozdělení (segmentace) složitější. Jestliže totiž nějaká segmentace obsahuje jinou segmentaci, mohou být obě stejně platné. Ve většině případů však dochází k soupeření těchto platných segmentací, což vede k přeučení na trénovacích datech a přílišné determinizaci odhadnutých frázových překladů. V práci je dále definován frázový přiřazovací model odpovídající IBM modelu 3 pro slovní přiřazení (viz část 4.1). Tento model je založen na následujícím generativním postupu:

1. Začni s cílovou větou $\mathbf{t}$.
2. Rozděl $\mathbf{t}$ na posloupnost K frází $\tilde{t}_1^K$, které pokryjí celou větu.
3. Pro každou frázi $\tilde{t}_k \in \tilde{t}_1^K$ vyber odpovídající pozici j v zdrojové větě a stanov přiřazení $a_j = i$, potom vytvoř právě jednu frázi $\tilde{s}_k$ z $\tilde{t}_k$.
4. Posloupnost $\tilde{s}_k$ uspořádaná podle $\mathbf{a} = a_1^K$ představuje zdrojovou větu $\mathbf{s}$.

Pravděpodobnostní model tohoto postupu je:

$$Pr(\mathbf{s}|\mathbf{t}) = \sum_{\tilde{t}_1^K, \tilde{s}_1^K, \mathbf{a}} Pr(\mathbf{s}, \tilde{t}_1^K, \tilde{s}_1^K, \mathbf{a}|\mathbf{t}) = \sum_{\tilde{t}_1^K, \tilde{s}_1^K, \mathbf{a}} \sigma(\tilde{t}_1^K|\mathbf{t})) \prod_{\tilde{t}_k \in \tilde{t}_1^K} \phi_T(\tilde{s}_k|\tilde{t}_k) \cdot d(a_j = i|\mathbf{t}), \quad (4.27)$$

kde σ je model rozdělení (předpokládáme uniformní rozdělení pro všechny možné segmentace věty), ϕ_T je překladová pravděpodobnost a d je distorzní pravděpodobnost založená na absolutní větné pozici. K trénování je opět použit EM algoritmus, aby byl však tento algoritmus použitelný, je třeba provést určité aproximace při jeho výpočtu. Pro tento účel je využito symetrické slovní přiřazení (na fráze je kladeno stejné omezení jako v případě výběru frází ze slovního přiřazení (viz část 4.1.1)), které umožní snížit počet uvažovaných frází a také spočítat aproximaci sumy všech možných přiřazení.

4.3 Další metody pro výběr frází

V této části představíme několik prací, které popisují různé další metody pro výběr frází. Každá z těchto metod volí vlastní přístup odlišný od ostatních i od postupů uvedených výše.

První prací je práce [Zhang 03], která k výběru frází používá bodovou vzájemnou informaci mezi zdrojovými a cílovými slovy. Na začátku extrakce je pro daný pár vět $(\mathbf{s}, \mathbf{t})$ vytvořena dvoudimenzionální matice $R_{J\times I}$, kde každý prvek odpovídá bodové vzájemné informaci mezi slovy s_j a t_i tedy:

$$R[j,i] = \log_2 \frac{P(s_j, t_i)}{P(s_j) \cdot P(t_i)}.$$

Všechny prvky matice jsou na začátku označeny jako „volné“. Nyní mezi všemi „volnými“ prvky nalezni ten s nejvyšší bodovou vzájemnou informací. Tento prvek dále rozšiř na co největší obdélník (frázi), přitom musí být splněna následující dvojice podmínek: Za prvé hodnota prvků, které budou použity k rozšíření, musí být podobná hodnotě dříve zvoleného prvku, tj. poměr hodnot těchto prvků musí být větší než daný práh. Za druhé do rozšíření nesmí být zahrnut prvek, jehož hodnota je vyšší než hodnota prvního vybraného prvku. Všechny prvky v obdélníku nyní označ jako „zablokované“. Jestliže jsou zde stále nějaké „volné“ prvky opakuj opět výběr „volného“ prvku s největší hodnotou. Pokud již nezbývají žádné „volné“ prvky, algoritmus končí a na výstupu jsou nalezené překladové páry.

V práci [Lavecchia 08] je k výběru frází také využita vzájemná informace mezi zdrojovou a cílovou frází a sice ve formě mezijazykových spouštěčů. Mezijazykové spouštěče jsou inspirovány konceptem spouštěčů pro statistické jazykové modelování ([Tillmann 97a]), kde spouštěč (angl. trigger) je složen ze slova a jemu odpovídajícímu nejlepšímu spuštěnému (předpovězenému) slovu z hlediska jejich vzájemné informace. Mezijazykové spouštěče jsou pak složeny ze zdrojového spouštějícího slova (resp. posloupnosti slov) a spuštěného cílového slova (resp. posloupnosti slov). V práci jsou použity jak spouštěče 1-k-1 tak i N-k-M. Aby mohli být mezijazykové spouštěče použity pro automatický překlad, je každému spouštěči přiřazena následující pravděpodobnost:

$$\forall s, t_i \in Trig(s) \quad P(t_i|s) = \frac{MI(t_i, s)}{\sum_{t\in Trig(s)} MI(t, s)}, \tag{4.28}$$

kde $MI(t,s)$ je jako obvykle definovaná vzájemná informace mezi cílovou a zdrojovou frází a $Trig(s)$ je množina cílových slov spuštěná zdrojovým slovem s. Při výběru frází je nejprve zdrojová část korpusu přepsána ve formě frází, získaných metodou pro výběr relevantních frází [Zitouni 03]. Každé takto získané zdrojové frázi je pak přiřazena množina k nejlepších mezijazykových spouštěčů. Nakonec jsou pomocí algoritmu pro simulované žíhaní (angl. simmulated annealing) vybrány nejlepší překladové páry. Nejprve se začne se systémem, který obsahuje jen spouštěče 1-k-1 a pak jsou postupně náhodně přidávány spouštěče N-k-M, pokud dojde přidáním ke zlepšení přesnosti překladu, je přidaný spouštěč zahrnut do výsledné frázové tabulky.

Práce [Moore 07] popisuje postup pro výběr frází, který nepoužívá segmentaci zdrojové a cílové věty. Snaží se tak vyhnout problémům, které způsobuje zahrnutí segmentace do trénování modelu frázového přiřazení, jak je uvedeno v práci [Denero 06]. V práci je tak navržen iterativně trénovatelný model, který je založen na dvou stochastických procesech, výběru a přiřazení, následovně:

1. Pro každý pár vět, který má vytvořeno slovní přiřazení, identifikuj všechny možné fráze, které jsou konzistentní s kritériem pro výběr frází ze slovního přiřazení (viz část 4.1.1).
2. Každé zdrojové frázi, která je zahrnuta v nějakém překladovém páru, vyber náhodně jednu z cílových frází.
3. Každé cílové frázi, která je zahrnuta v nějakém překladovém páru, vyber náhodně jednu ze zdrojových frází.
4. Zdrojová fráze je přiřazena cílové tehdy a jen tehdy, jestliže je jejich výběr vzájemný.

Pro trénování tohoto modelu byl opět použit EM algoritmus. Nejprve jsou uniformě inicializovány překladové pravděpodobnosti. Z nich jsou pak spočteny výběrové pravděpodobnosti, na jejichž základě je spočten nový odhad překladových pravděpodobností, který může být použit pro další iteraci algoritmu. Tento postup je opakován, dokud není dosaženo požadovaného výsledku.

V práci [Deng 08] je na problém výběru frází pohlíženo jako na problém vyhledávání informací, který je řešen pomocí log-lineárního modelu. Tento přístup tak umožňuje použít kombinaci různých rysů a modelů (v článku jsou představeny rysy založené na slovním přiřazení, aposteriorní pravděpodobnosti i na jedno a dvojjazyčných informačních metrikách (ty měří např. vzájemnou souvislost mezi frázemi a kvalitu jednotlivých frází)). Je představen obecný postup pro výběr frází, který může být optimalizován společně s překladovým systémem tak, aby byla maximalizována přesnost výsledného systému. Při tomto postupu jsou nejprve natrénovány IBM model 1 a HMM model slovního přiřazení (ty jsou dále využity při výpočtu použitých rysů a modelů). Pro každý větný pár jsou určeny všechny možné kandidátské fráze pro obě strany překladu a pro všechny možné překladové páry jsou pak spočteny hodnoty všech použitých rysů a modelů. Dále je vypočteno celkové skóre každého uvažovaného páru a je určen kandidát s nejvyšším skóre. Všechny páry pro které je rozdíl mezi jejich hodnotou skóre a nejvyšším skóre větší než daný práh jsou zařazeny do frázové tabulky. Nakonec jsou diskriminativně určeny váhy pro výpočet celkového skóre překladového páru a práh pro zařazení páru do tabulky. Tento postup se opakuje, dokud nejsou nalezeny optimální hodnoty použitých vah a prahu.

Nakonec zmiňme práci [Zettlemoyer 07], která se na rozdíl od předešlých prací nezabývá tvorbou frázové tabulky, ale úpravou stávající tak, aby byla menší a poskytovala lepší výsledky. Na začátku je frázová tabulka, která obsahuje fráze získané pomocí postupu pro výběr frází konzistentních se slovním přiřazením (viz část 4.1.1). Výsledkem je množství frází, které mají velký překryv (viz Obrázek 4.1) a rozdílnou překladovou kvalitu. Cílem je tedy z těchto frází vybrat nepřekrývající se vysoce kvalitní fráze. Za tímto účelem je každé frázi přiřazeno skóre, které se skládá z překladových pravděpodobností pro oba směry překladu, lexikální pravděpodobnosti a jim odpovídajícím váhám. Hodnoty těchto pravděpodobností jsou získány z plné frázové tabulky, hodnoty vah pak pomocí MERT tréninku s plnou frázovou tabulkou. Při výběru frází jsou použity dvě hlediska: za prvé jsou vybírány jen fráze, které mají vysoké skóre a za druhé se při výběru uplatňuje omezení o nadbytečnosti, tj. předpokládáme, že pro daný větný pár má každá zdrojová i cílová fráze nejvýše jeden překlad. Algoritmus pro výběr

překladových páru pak pracuje následovně. Pro každý větný pár v trénovacím korpusu je pro každou zdrojovou i cílovou frázi určen překladový pár s nejvyšším skóre obsahující tuto zdrojovou nebo cílovou frázi. Každý takto označený překladový pár je vybrán a je vytvořena nová frázová tabulka (jsou přepočteny překladové a lexikální pravděpodobnosti).

4.4 Výběr frází založený na principu minimální ztráty

V této části bude popsána nově navržená metoda pro výběr frází. Stejně jako metoda popsaná v práci [Deng 08] nebo metoda popsaná v práci [Zettlemoyer 07] je tato metoda založena na výběru překladových párů. Použitím tohoto přístupu se vyhneme problémům, které nastávají při použití metod založených na modelování slovního nebo frázového přiřazení. Jde především o chyby v přiřazení a výpočetní problémy, jako např. potřeba aproximace množiny všech přiřazení pro výpočet složitějších přiřazovacích modelů, a také o problémy se skrytou proměnnou pro rozdělení zdrojové a cílové věty v případě modelů pro frázového přiřazení (jak je uvedeno v práci [Denero 06] je přítomnost této skryté proměnné v modelu příčinnou nižší přesnosti frází získaných z modelu frázového přiřazení oproti frázím získaným heuristicky ze slovního přiřazení). Pro výpočet skóre každého překladového páru je stejně jako v práci [Deng 08] použit diskriminativní model, který umožňuje prostřednictvím rysů a jejich vah jednoduchou a účelnou kombinaci informací z různých zdrojů. Hlavní rozdíl zde představené metody, oproti metodě popsané v práci [Deng 08], je ve způsobu výběru frází s nejvyšším skóre založeném na principu minimální ztráty, oproti metodě popsané v práci [Zettlemoyer 07] je pak mimo způsobu výběru rozdíl hlavně ve výchozím bodě, kdy se začíná se všemi možnými frázemi a ne s již vytvořenou frázovou tabulkou. Poznamenejme, že zde představená metoda byla vyvinuta nezávisle na metodách představených v pracích [Zettlemoyer 07, Deng 08].

Naším úkolem při výběru frází je každé zdrojové frázi $\tilde{s}$ nalézt její překlad, tj. odpovídající cílovou frázi $\tilde{t}$. Předpokládáme, že na začátku máme dvojjazyčný korpus, který obsahuje odpovídající si páry vět. Začneme se zdrojovou větou $\mathbf{s} = w_1, ..., w_J$ a cílovou větou $\mathbf{t} = w_1, ..., w_I$ a vytvoříme pro každou z nich balík β všech možných frází až do dané **délky** l:

$$\beta\{\mathbf{s}\} = \{\tilde{s}_m\}_{m=1}^{l} \quad \{\tilde{s}_m\} = \{w_n, ..., w_{n+m-1}\}_{n=1}^{J-m+1} \tag{4.29}$$

$$\beta\{\mathbf{t}\} = \{\tilde{t}_m\}_{m=1}^{l} \quad \{\tilde{t}_m\} = \{w_n, ..., w_{n+m-1}\}_{n=1}^{I-m+1}. \tag{4.30}$$

Tímto postupem ale samozřejmě dostaneme obrovské množství frází, které by bylo třeba dále zkoumat. Abychom zmenšili počet uvažovaných frází a tím i počet možných překladových párů, jsou pro další zpracovaní v případě zdrojových frází delších než jedna uvažovány jen fráze, které se v korpusu vyskytli minimálně tolikrát jako je daný práh τ (stejně jako v práci [Marcu 02] uvažujeme jen fráze, které se vyskytly alespoň pětkrát). Cílové fráze jsou naproti tomu zachovány všechny bez ohledu na četnost jejich výskytu v korpusu. Každá cílová fráze je nyní považována za možný překlad každé vybrané zdrojové fráze tedy:

$$\forall \tilde{s} \in \beta\{\mathbf{s}\} : |\tilde{s}| = 1 \vee N(\tilde{s}) \geq \tau : \tilde{s} \rightarrow \beta\{\mathbf{t}\}, \tag{4.31}$$

kde $N(\tilde{s})$ je počet výskytů fráze $\tilde{s}$ v trénovacím korpusu. Nyní pro každý možný překladový pár $(\tilde{s}, \tilde{t}) : \tilde{t} \in T(\tilde{s}), T(\tilde{s}) = \{\tilde{t}\} : \tilde{s} \rightarrow \tilde{t}$ spočteme odpovídající skóre c. To je obecně dáno log-lineárním modelem:

$$c(\tilde{s}, \tilde{t}) = \sum_{k=1}^{K} \lambda_k h_k(\tilde{s}, \tilde{t}), \tag{4.32}$$

kde $h_k(\tilde{s}, \tilde{t}), k = 1, 2, ..., K$ je množina K rysů, které popisují vztah mezi dvojicí frází $(\tilde{s}, \tilde{t})$. Výsledná skóre $\mathbf{c} = \{c\}$ jsou uložena do hash tabulky, kde klíčem je vždy zdrojová fráze $\tilde{s}$

Algoritmus výběru frází založený na principu minimální ztráty

for všechny *páry vět (**s**,**t**)* **do**
  Vytvoř balík frází pro každou stranu překladu
  for všechny frázové páry $(\tilde{s}, \tilde{t})$ **do**
    Spočti hodnoty všech rysů $h_k(\tilde{s}, \tilde{t})$
    Vypočti a ulož do hash tabulky hodnotu celkového skóre $c(\tilde{s}, \tilde{t}) = \sum_{k=1}^{K} \lambda_k h_k(\tilde{s}, \tilde{t})$
  end for
end for
for všechny *páry vět (**s**,**t**)* **do**
  Vytvoř balík frází pro každou stranu překladu
  for všechny fráze $\tilde{s} \in \beta(\boldsymbol{s})$ **do**
    for všechny překlady $\tilde{t} \in \beta(\boldsymbol{t})$ **do**
      Spočti a ulož překladovou ztrátu $L(\tilde{s}, \tilde{t}) = \frac{\sum_{\tilde{s}_i \in \beta(\boldsymbol{s}), \tilde{s}_i \neq \tilde{s}} c(\tilde{s}_i, \tilde{t})}{c(\tilde{s}, \tilde{t})}$
    end for
    Uspořádej všechny překlady podle velikosti překladové ztráty a vyber překlad s nejnižší ztrátou $\tilde{t_G} = argmin_{\tilde{t}} L(\tilde{s}, \tilde{t})$
    Vybraný překlad započítej do výsledné tabulky
  end for
end for

Obrázek 4.2: Algoritmus pro výběr frází založený na principu minimální ztráty.

a jako data jsou uloženy všechny její možné překlady $\tilde{t} \in T(\tilde{s})$ spolu s odpovídajícím skóre $c(\tilde{s}, \tilde{t})$. Takto projdeme celý trénovací korpus a uložíme do tabulky skóre pro všechny možné překladové páry z celého korpusu. Optimální hodnoty vah λ_k můžeme nalézt pomocí MER trénování z hlediska zvoleného kritéria.

Dalším krokem je určení jen „dobrých“ překladů $\tilde{t}_G$ z množiny všech možných překladů $T(\tilde{s})$ pro každou zdrojovou frázi $\tilde{s}$, tj. dostaneme množinu frází $T_G(\tilde{s}) = \{\tilde{t}_G\} : \tilde{s} \rightarrow \tilde{t}_G$. Existuje několik možností jak tyto „dobré“ překlady vybrat. Jako první se nabízí postup, kdy pro každou zdrojovou frázi uspořádáme všechny její možné překlady podle velikosti výsledného skóre c a za „dobré“ prohlásíme obecně N překladů s nejvyšším skóre. Další možností, která je použita v práci [Deng 08], je pro každý větný pár uspořádat všechny možné překladové páry podle velikosti výsledného skóre c, nalézt pár s nejvyšším skóre c_m a pak vybrat všechny překladové páry, pro něž je rozdíl mezi jejich skóre a maximálním skóre větší něž daný práh: $c_m - c \geq \tau_S$. Novou možností navrženou v této práci je pak výběr založený na principu minimální ztráty. V tomto případě postupujeme při výběru následovně. Znovu procházíme celý korpus a pro každý větný pár vytvoříme opět balík všech frází do dané délky l pro obě věty (v případě zdrojových frází delších než jedna uvažujeme jen fráze, které se vyskytly minimálně tolikrát jako je daný práh τ). Pak pro každou frázi $\tilde{s} \in \beta(\mathbf{s})$ a každý její možný

překlad $\tilde{t} \in T(\tilde{s}) = \beta(\mathbf{t})$ spočteme překladovou ztrátu L_T. Překladová ztráta L_T zdrojové fráze $\tilde{s}$ a jejího potenciálního překladu $\tilde{t}$ je definována jako:

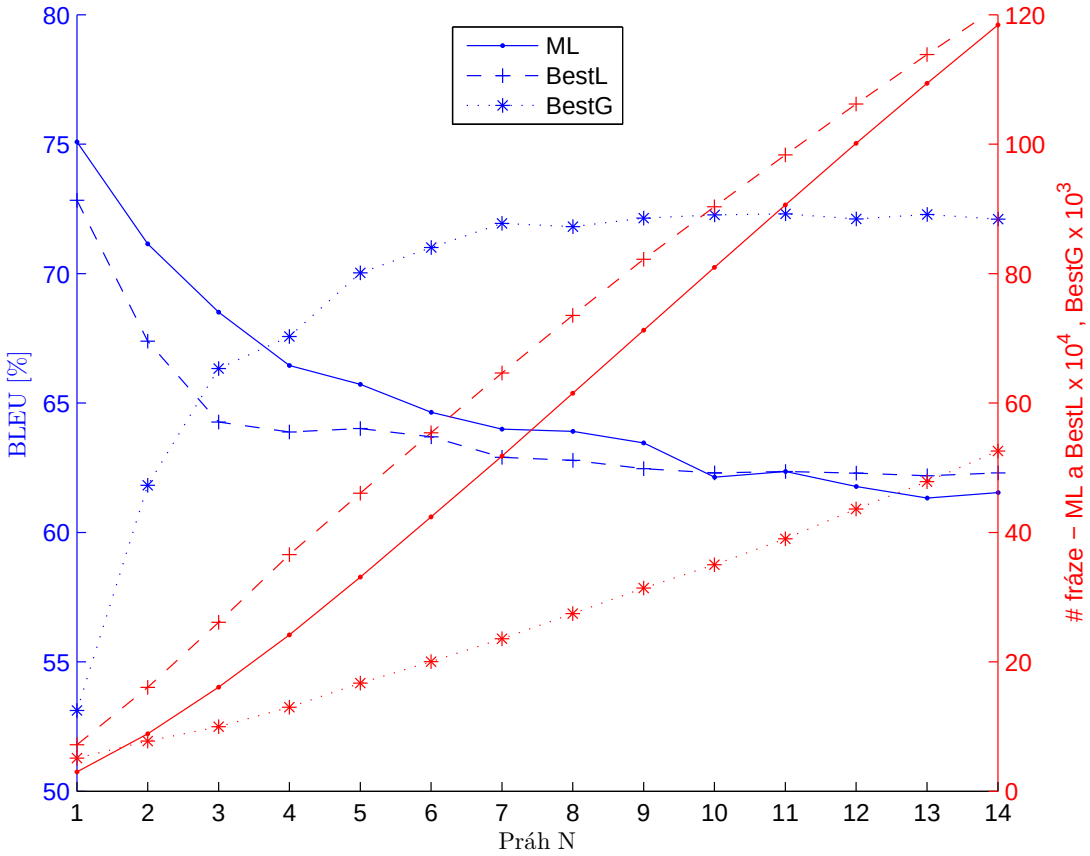
$$L_T(\tilde{s}, \tilde{t}) = \frac{\sum_{\tilde{s}_i \in \beta(\mathbf{s}), \tilde{s}_i \neq \tilde{s}} c(\tilde{s}_i, \tilde{t})}{c(\tilde{s}, \tilde{t})}. \tag{4.33}$$

Tedy počítáme, jak velké množství pravděpodobnosti ztratíme pro zbytek zdrojových frází z balíku $\beta(\mathbf{s})$, jestliže přeložíme $\tilde{s}$ jako $\tilde{t}$. Pro každou frázi $\tilde{s}$ pak uložíme všechny překladové ztráty $L_T(\tilde{s}, \tilde{t})$ pro všechny možné překlady $\tilde{t} \in \beta(\mathbf{t})$. Jako „dobrý“ překlad $\tilde{t}_G$ fráze $\tilde{s}$ pak označíme ten překlad $\tilde{t}$, který způsobí nejnižší překladovou ztrátu $L_T(\tilde{s}, \tilde{t})$ (resp. překlady, pokud jich více splňuje danou podmínku), tedy:

$$\tilde{t}_G = \operatorname*{argmin}_{\tilde{t}} L_T(\tilde{s}, \tilde{t}) \tag{4.34}$$

a všechny ostatní překlady zamítneme (obecně lze opět vybrat N překladů s nejlepším skóre). Takto projdeme všechny větné páry a pro každou zdrojovou frázi uložíme všechny „dobré“ překlady spolu s informací, kolikrát byl daný překlad zvolen jako „dobrý“. Tím dostaneme finální frázovou tabulku (zdrojové fráze a jejich „dobré“ překlady), informace v ní uložené pak mohou být použity např. pro výpočet překladových pravděpodobností na základě četnosti výskytu [Koehn 03]. Celý postup je zachycen na Obrázku 4.2.

Zastavme se ještě u způsobu výběru „dobrých“ překladů založeném na principu minimální ztráty. Předpokládejme, že máme za úkol z nějaké množiny kandidátů vybrat jednoho kandidáta, který nejlépe odpovídá referenčnímu vzorku. Vztah mezi referenčním vzorkem a kandidáty je přitom popsán pomocí nějakého pravděpodobnostního modelu. Domníváme se, že obecně existují dvě možnosti jak to udělat. První možnost je založena na maximální aposteriorní pravděpodobnosti (a nazýváme jí angl. maximum a posteriori (MAP)), kdy vybereme kandidáta s maximální pravděpodobností danou modelem. Druhá možnost je pak založena na minimálním Bayesovském risku (a nazýváme jí angl. minimum Bayes risk (MBR)), kdy je vztah mezi referenčním vzorkem a kandidáty popsán pomocí nějaké ztrátové funkce L a kdy vybereme kandidáta, pro kterého tato ztrátová funkce nabývá minima. V případě výběru frází metoda založená na výběru N nejlepších překladů představuje aplikaci námi zmíněné MAP metody výběru. A metoda založená na principu minimální ztráty pak, jestliže je ztrátová funkce definována jako překladová ztráta L_T, představuje aplikaci námi zmíněné MBR metody výběru. V případě výběru frází a jejich překladů vycházíme totiž z představy, že nejlepší (nejpravděpodobnější) překlad zdrojové věty představuje určité množství pravděpodobnosti, které je složeno z dílčích pravděpodobností jednotlivých překladových párů (rozdělení vět na fráze ponechme nyní stranou). Nejlepší překlad tak můžeme dostat tak, že ke každé zdrojové frázi vybereme její nejpravděpodobnější překlad reprezentovaný nejvyšším skóre $c(\tilde{s}, \tilde{t})$ (MAP výběr) a jejich seskupením pak dostaneme nejlepší překlad zdrojové věty. Nebo tak, že jestliže nejlepší překlad představuje největší množství pravděpodobnosti, vybereme ke každé zdrojové frázi takový překlad, který představuje co nejmenší úbytek z tohoto ideálního množství pravděpodobnosti (MBR výběr). Pokud totiž k dané frázi vybereme jiný než optimální překlad (tj. překlad, který je součástí nejlepšího překladu), snížíme celkovou pravděpodobnost výsledného překladu. Z teoretického hlediska spočívá hlavní výhoda výběru založeného na principu minimální ztráty, oproti výběru založenému na výběru N nejlepších překladů, v zapojení kontextu, v kterém se daná fráze nachází ve zdrojové větě. Tento kontext se uplatňuje při výpočtu překladové ztráty L_T, kdy jsou prostřednictvím hodnot $c(\tilde{s}_i, \tilde{t})$, na rozdíl od výběru N nejlepších překladů, vzaty do úvahy i další fráze a jejich překlady, které se nachází ve zdrojové větě. Abychom tento předpoklad ověřili, porovnali jsme jednotlivé metody pro výběr „dobrých“ překladů pomocí následujícího experimentu.



Obrázek 4.3: Porovnání různých kritérií pro výběr „dobrých“ překladů.

Data z CSC korpusu jsme rozdělili na trénovací, vývojovou a testovací část. Poté jsme za použití stejných rysů a vah vytvořili z trénovacích dat pomocí každého z uvedených kritérií pro výběr frází frázovou tabulku. Vzniklou frázovou tabulku jsme pak vždy za stejných podmínek použili pro překlad vývojových dat. Přesnost překladu jsme změřili pomocí BLEU skóre. Výsledky jsou v grafu na Obrázku 4.3. Kde ML označuje metodu pro výběr frází založený na principu minimální ztráty, BestL výběr N nejlepších překladů pro každou zdrojovou frázi a BestG pak výběr N nejlepších překladových párů pro každou dvojici vět. Na ose X jsou hodnoty prahu N, které určují maximální počet vybraných frází při každém výběru, resp. tato hodnota určuje jaký počet nejlepších hodnot skóre c a jim odpovídajících překladových párů bude vybrán. Na levé ose Y jsou výsledky překladu a na pravé pak velikost výsledné frázové tabulky pro různé hodnoty prahu N. Z grafu je vidět, že metody ML a BestL dosahují nejlepšího výsledku při výběru vždy jen nejlepší hodnoty skóre c a jí odpovídajících překladových párů, s rostoucím množstvím vybraných frází kvalita překladu prudce klesá (s rostoucí velikostí obsahuje frázová tabulka více šumu, který znehodnocuje výsledky překladu). Dále je také vidět, že ML metoda založená na MBR výběru dosahuje lepších výsledků než obě metody BestL a BestG založené na MAP výběru (75,09 versus 72,83 a 72,31). Při výběru „dobrých“ překladů je tedy nejvhodnější použít metodu výběru založenou na principu minimální ztráty.

4.4.1 Rysy

Jak již bylo řečeno, log-lineární model pro výpočet skóre c umožňuje prostřednictvím různých rysů kombinaci informací z různých zdrojů. V případě této práce se však zatím při výběru frází omezíme jen na použití rysů založených na četnostech výskytu překladových párů a jednotlivých frází v trénovacím korpusu. Za tímto účelem jsme prošli celý trénovací korpus a spočetli jsme tyto četnosti: počet výskytů každé uvažované zdrojové fráze $N(\tilde{s})$ (odpovídá počtu zdrojových vět, v kterých se daná fráze objevila), počet výskytů každé cílové fráze $N(\tilde{t})$ (odpovídá počtu cílových vět, v kterých se daná fráze objevila), počet výskytů každého možného překladového páru $N(\tilde{s}, \tilde{t})$ (odpovídá počtu větných párů, v kterých se daný překladový pár objevil) a nakonec počet kolikrát byla daná cílová nebo zdrojová fráze považována za překlad $N_T(\tilde{t})$ a $N_T(\tilde{s})$ (odpovídá počtu všech frází, se kterými se daná fráze objevila ve všech větných párech jako jejich možný překlad). Na základě těchto četností uvažujeme následující pravděpodobnosti (rysy): překladovou pravděpodobnost ϕ_T, pravděpodobnost p_T, že je daná fráze překladem, tj. objeví se spolu s uvažovanou frází jako její překlad, obojí vždy pro oba směry překladu a nakonec překladovou pravděpodobnost založenou na vzájemné informaci p_{MI}.

Překladová pravděpodobnost ϕ_T je definována na základě relativních četností jako [Koehn 03]:

$$\phi_T(\tilde{s}|\tilde{t}) = \frac{N(\tilde{s}, \tilde{t})}{N(\tilde{t})} \quad \phi_T(\tilde{t}|\tilde{s}) = \frac{N(\tilde{s}, \tilde{t})}{N(\tilde{s})}. \tag{4.35}$$

Pravděpodobnost p_T, že je daná fráze překladem, tj. objeví se spolu s uvažovanou frází jako její překlad je definována jako:

$$p_T(\tilde{s}|\tilde{t}) = \frac{N(\tilde{s}, \tilde{t})}{N_T(\tilde{t})} \quad p_T(\tilde{t}|\tilde{s}) = \frac{N(\tilde{s}, \tilde{t})}{N_T(\tilde{s})}. \tag{4.36}$$

Pravděpodobnost založená na vzájemné informaci p_{MI} je definována jako [Lavecchia 08]:

$$p_{MI}(\tilde{s}, \tilde{t}) = \frac{MI(\tilde{s}, \tilde{t})}{\sum_{\tilde{t} \in T(\tilde{s})} MI(\tilde{s}, \tilde{t})}. \tag{4.37}$$

Vzhledem k tomu, že jsme pro každou zdrojovou i cílovou frázi spočetli dvě četnosti N a N_T můžeme na základě těchto četností spočítat i dvě vzájemné informace a jim odpovídající pravděpodobnosti založené na vzájemné informaci:

$$p_{MI}(\tilde{s}, \tilde{t}) = \frac{MI(\tilde{s}, \tilde{t})}{\sum_{\tilde{t} \in T(\tilde{s})} MI(\tilde{s}, \tilde{t})} \quad p_{MI_T}(\tilde{s}, \tilde{t}) = \frac{MI_T(\tilde{s}, \tilde{t})}{\sum_{\tilde{t} \in T(\tilde{s})} MI_T(\tilde{s}, \tilde{t})} \tag{4.38}$$

$$MI(\tilde{s}, \tilde{t}) = p(\tilde{s}, \tilde{t}) \cdot \log \frac{p(\tilde{s}, \tilde{t})}{p(\tilde{s}) \cdot p(\tilde{t})} \quad MI_T(\tilde{s}, \tilde{t}) = p_T(\tilde{s}, \tilde{t}) \cdot \log \frac{p_T(\tilde{s}, \tilde{t})}{p_T(\tilde{s}) \cdot p_T(\tilde{t})} \tag{4.39}$$

$$p(\tilde{s}, \tilde{t}) = \frac{N(\tilde{s}, \tilde{t})}{N_S} \quad p_T(\tilde{s}, \tilde{t}) = \frac{N(\tilde{s}, \tilde{t})}{N_T} \tag{4.40}$$

$$p(\tilde{s}) = \frac{N(\tilde{s})}{N_S},\ p(\tilde{t}) = \frac{N(\tilde{t})}{N_S} \quad p_T(\tilde{s}) = \frac{N_T(\tilde{s})}{N_T},\ p_T(\tilde{t}) = \frac{N_T(\tilde{t})}{N_T}, \tag{4.41}$$

kde N_S je počet všech párů vět v trénovacím korpusu a N_T je počet všech možných uvažovaných překladů, tj. jestliže má zdrojová věta pět a cílová devět frází, je do N_T započteno 45 možných překladů. Pravděpodobnosti ϕ_T a p_{MI} představují pravděpodobnost, že danou frázi $\tilde{s}$ přeložíme jako cílovou frázi $\tilde{t}$. Pravděpodobnosti p_T a p_{MI_T} pak představují pravděpodobnost, že se daná fráze $\tilde{t}$ objeví spolu s frází $\tilde{s}$ jako její překlad z množiny všech možných překladů $T(\tilde{s})$. Abychom porovnali jednotlivé rysy, použili jsme opět CSC korpus rozdělený jako v předešlém případě na

Rys	BLEU – 1	BLEU – 5	BLEU – 10	Počet frází
$\phi_T(\tilde{s}\|\tilde{t})$	39,06	39,67	40,27	1 231 830
$\phi_T(\tilde{t}\|\tilde{s})$	50,21	46,84	47,04	567 057
$p_T(\tilde{s}\|\tilde{t})$	70,09	73,12	72,93	25 715
$p_T(\tilde{t}\|\tilde{s})$	68,60	72,54	73,20	90 358
$p_{MI}(\tilde{s},\tilde{t})$	71,45	75,36	75,40	29 835
$p_{MI_T}(\tilde{s},\tilde{t})$	72,42	75,67	75,82	26 376

Tabulka 4.1: Přesnost překladu pro jednotlivé rysy.

trénovací, vývojovou a testovací část. Pro každý rys jsme pak vytvořili výše popsanou metodou frázovou tabulku vždy za použití jen tohoto rysu. Nakonec jsme vytvořenou frázovou tabulku použili pro překlad vývojových dat. Při tvorbě tabulky i překladu byly samozřejmě použity vždy stejné podmínky pro všechny rysy. Přesnost překladu jsme opět změřili pomocí BLEU kritéria. Výsledky jsou v Tabulce 4.1. BLEU – 1 označuje výsledky pro výběr překladů jen s nejvyšším skóre při překladu vývojových dat. BLEU – 5 a BLEU – 10 pak obdobně označuje výsledky pro výběr jen těch překladů, jejichž skóre patří mezi pět resp. deset nejvyšších. Jak je vidět z výsledků, jsou nejlepšími rysy pravděpodobnosti p_{MI_T}, p_{MI}. Jako nevhodné rysy se naopak jeví překladové pravděpodobnosti ϕ_T, které vedou na veliké tabulky s nízkou přesností překladu. Toto zjištění je poměrně překvapivé, neboť překladové pravděpodobnosti jsou při překladu jedním z hlavních zdrojů informace o zdrojovém a cílovém jazyce.

4.4.2 Vylepšení základní metody

V této části popíšeme několik vylepšení základní metody pro výběr frází založený na principu minimální ztráty (ML metoda), které by měli mít pozitivní vliv na přesnost překladu získaného pomocí takto vytvořené frázové tabulky a také na velikost této tabulky (ta je klíčová v případě rychlosti překladu).

Na základě analýzy chyb překladu vytvořeného pomocí frázové tabulky získané ML metodou jsme zjistili, že dochází k přílišné determinizaci u frází, které se vyskytli v korpusu jen jednou. To je způsobeno výběrem vždy jen jednoho nejlepšího překladu (resp. nejnižší hodnoty překladové ztráty L_T a jí odpovídajících překladů) zdrojové fráze pro každý větný pár. U zdrojových frází, které se vyskytly jen jednou, tak typicky dochází k výběru jen jednoho překladu, který je ve většině případů chybný. Abychom tomu zabránili, rozdělili jsme výběr „dobrých“ překladů podle četnosti výskytu dané zdrojové fráze. Pokud se fráze vyskytla jen jednou, jsou jako „dobré“ označeny všechny možné překlady bez ohledu na hodnotu překladové ztráty L_T. V ostatních případech jsou opět vybrány jen překlady s nejnižší hodnotou L_T.

Pokud se podíváme na výslednou frázovou tabulku, zjistíme, že obsahuje řadu nesmyslných překladových párů, kdy byla jako překlad vybrána špatná cílová fráze. Toto zašumění výsledné tabulky lze snížit pomocí techniky symetrizace frázové tabulky. Podobně jako v případě metody pro výběr frází založeném na slovním přiřazení, kdy je průnikem slovních přiřazení pro jednotlivé směry překladu vytvořeno symetrické slovní přiřazení, vytvoříme také novou frázovou tabulku jako výsledek průniku frázových tabulek odvozených ML metodou pro jednotlivé směry překladu. Do nové tabulky vybereme jen ty překladové páry, pro které platí, že tvoří vzájemný překlad, tj. že $\tilde{t}$ je překladem $\tilde{s}$ a zároveň $\tilde{s}$ je překladem $\tilde{t}$. Každý vybraný překlad je

do nové tabulky uložen spolu s původní informací, kolikrát byl vybrán jako „dobrý“ překlad.

Další možností, jak snížit šum ve vytvořené tabulce, je provést filtrování této tabulky. Vytvořenou tabulku, která obsahuje překladové páry spolu s hodnotami rysů používaných při překladu, použijeme k překladu trénovacích dat a zaznamenáme přitom, které překlady a kolikrát byly použity. Tuto informaci pak použijeme pro vytvoření nové frázové tabulky. Tato metoda může být použita na libovolnou frázovou tabulku. Podrobné výsledky pro jednotlivé úpravy výsledné tabulky a porovnání s dalšími frázovými tabulkami jsou v Kapitole 7.

Kapitola 5

Prohledávání

Cílem prohledávání je nalézt k dané zdrojové větě nejlepší odpovídající překlad. Tento postup se také nazývá generování nebo dekódování. Ze statistické teorie rozhodování vyplývá, že ze všech možných cílových vět bychom měli vybrat tu, která minimalizuje očekávanou ztrátu ([Duda 00]):

$$\hat{t}_1^{\hat{I}} = \underset{I,t_1^I}{\operatorname{argmin}} \left\{ \sum_{I',t_1'^{I'}} Pr(t_1'^{I'}|s_1^J) \cdot L(t_1^I, t_1'^{I'}) \right\}. \tag{5.1}$$

Tento výraz se nazývá Bayesovo rozhodovací pravidlo pro statistický automatický překlad ([Zens 08]). $L(t_1^I, t_1'^{I'})$ představuje ztrátovou funkci, která měří velikost chyby (ztrátu), jestliže zvolíme jako překlad t_1^I, když správný překlad je $t_1'^I$. $Pr(t_1'^{I'}|s_1^J)$ pak představuje aposteriorní rozdělení pravděpodobnosti přes všechny cílové věty $t_1'^{I'}$ pro danou zdrojovou větu s_1^J (v praxi je samozřejmě toto aposteriorní rozdělení neznámé, takže se nahrazuje hodnotami modelu vytvořeného z trénovacích dat). Jestliže ztrátovou funkci L zvolíme ve tvaru:

$$L_{0-1}(t_1^I, t_1'^{I'}) = \begin{cases} 0 & \text{jestliže } t_1^I = t_1'^{I'} \\ 1 & \text{jinak.} \end{cases} \tag{5.2}$$

Tato ztrátová funkce se pak nazývá 0–1 ztráta (všechny možnosti kromě správné obdrží ztrátu 1) a používá se v případě minimalizace větné nebo řetězcové chyby. A rozhodovací pravidlo lze zjednodušit na tvar:

$$\hat{t}_1^{\hat{I}} = \underset{I,t_1^I}{\operatorname{argmax}} \left\{ Pr(t_1^I|s_1^J) \right\}. \tag{5.3}$$

V případě log-lineárního modelu a maximální aproximace pak dostaneme:

$$\hat{t}_1^{\hat{I}} = \underset{I,t_1^I}{\operatorname{argmax}} \left\{ \max_{a_1^K} \sum_{m=1}^{M} \lambda_m h_m(t_1^I, a_1^K, s_1^J) \right\}. \tag{5.4}$$

Takovéto rozhodovací pravidlo se nazývá *maximální aposteriorní* (angl. maximal a posteriori (MAP)) *rozhodovací pravidlo*, neboť volíme kandidáta, který maximalizuje aposteriorní pravděpodobnost vytvořeného modelu. Díky jednoduché implementaci se toto pravidlo často používá pro dekódování a to i v případech, kdy se výsledná úspěšnost neměří pomocí 0–1 ztrátové funkce. Např. v případě automatického překladu se toto pravidlo často používá, i když všechny ztrátové funkce až na jednu (SER) používané pro vyhodnocení úspěšnosti výsledného překladu nejsou 0–1 ztrátové funkce. Jestliže použijeme jinou než 0–1 ztrátovou funkci, hovoříme pak o rozhodovacím pravidlu založeném na *minimálním Bayesovském risku* (angl. minimum Bayes

risk (MBR)). MBR dekódování pro různá kritéria používaná v oblasti automatického překladu (např. již zmíněná kritéria BLEU, WER, PER) je popsáno v práci [Kumar 04].

Prohledávání můžeme také interpretovat jako posloupnost rozhodnutí $(\tilde{t}_k, b_k, j_k)$ pro $k = 1, ..., K$. Kdy v každém kroku zvolíme zdrojovou frázi $\tilde{s}_k$ určenou její startovní a cílovou pozicí b_k, j_k a vybereme jí odpovídající překlad $\tilde{t}_k$. Abychom zajistili překlad celé zdrojové věty a splnili omezení kladená na segmentaci (žádné mezery ani překryvy), uchováváme si záznam o již přeložených („pokrytých“) zdrojových slovech. Tento záznam se nazývá *pokrytí* $C \subseteq \{1, ..., J\}$. Množinu všech možných segmentací a překladů můžeme reprezentovat pomocí grafu (stavový prostor), kde uzly představují jednotlivá pokrytí C a jim odpovídající překladové hypotézy a hrany pak rozhodnutí $(\tilde{t}_k, b_k, j_k)$. Počáteční uzel grafu je tedy označen *prázdným pokrytím* $C = \emptyset$ (žádné slovo ještě nebylo přeloženo) a cílový uzel je označen *plným pokrytím* $C = \{1, ..., J\}$ (všechna slova byla přeložena). Každá úplná cesta grafem představuje jeden možný překlad zdrojové věty, který vznikne spojením cílových frází $\tilde{t}$ podél cesty. Na základě překladového modelu, např. $\sum_{m=1}^{M} \lambda_m h_m(t_1^I, a_1^K, s_1^J)$, můžeme každému uzlu v grafu přiřadit jeho ohodnocení. Problém prohledávání pak můžeme definovat jako problém nalezení optimální cesty tímto grafem. Velikost grafu je exponenciální vzhledem k délce zdrojové věty. V pracích [Knight 99, Udupa 06] bylo dokázáno, že problém nalezení nejlepšího překladu spadá do třídy NP-obtížných (angl. NP-hard) problémů. Optimální řešení tedy nelze vždy nalézt v přijatelném čase. Abychom toho byli schopni, je třeba použít při prohledávání vhodné aproximace. Podle způsobu zpracování zdrojové věty můžeme prohledávání rozdělit na *monotónní* a *nemonotónní*. Poznamenejme, že cílová věta je vždy vytvářena postupně, tj. monotónně.

5.1 Monotónní prohledávání

V případě monotónního prohledávání jsou zdrojové fráze (slova) překládána jen v pořadí, v kterém jsou ve zdrojové větě. K přeuspořádání slov tedy může dojít jen uvnitř jednotlivých frází, které jsou pak přeloženy ve stejném pořadí, v jakém se nacházejí ve zdrojové větě. Díky tomuto omezení je problém prohledávání řešitelný pomocí dynamického programování. Výsledná složitost je lineární vzhledem k délce zdrojové věty. S ohledem na monotónní zpracování zdrojové věty platí:

$$b_k = j_{k-1} + 1, \quad k = 1, ..., K. \tag{5.5}$$

Jestliže definujeme $Q(j, \tilde{t})$ jako maximální skóre posloupnosti frází, která končí frází $\tilde{t}$ a pokrývá pozice 1 až j zdrojové věty, pak pro určení nejlepšího překladu, jemuž odpovídá skóre $\hat{Q}$, dostaneme následující rekurzivní algoritmus dynamického programování:

$$Q(0, \$) = 0 \tag{5.6}$$

$$Q(j, \tilde{t}) = \max_{\tilde{t}'} \left\{ Q(j_{prev}, \tilde{t}') + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t}' \oplus \tilde{t}, s_{j_{prev}+1}^{j}) \right\} \tag{5.7}$$

$$Q(J+1, \$) = \max_{\tilde{t}} \left\{ Q(J, \tilde{t}) + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t} \oplus \$, s_1^J) \right\}. \tag{5.8}$$

Kde \$ označuje hranice věty a j_{prev} označuje předchozí pokrytí, z kterého je vytvářeno nové pokrytí j, platí: $j - PhL_{max} \leq j_{prev} < j$, PhL_{max} je maximální uvažovaná délka fráze. Symbol $\oplus$ označuje spojení dvou řetězců pomocí mezery do jednoho řetězce. Celý postup algoritmu při dekódování vstupní zdrojové věty je popsán na Obrázku 5.1.

Vstupem je zdrojová věta, která je rozdělena na všechny možné fráze délky 1 až PhL_{max}, při zachování podmínky o spojitosti frází. Ve frázové tabulce jsou pak vyhledány překlady

Monotónní prohledávání

Vstup: Hash tabulka, která obsahuje všechna možná rozdělení zdrojové věty a jim odpovídající překlady:

$TA\{C(0,b),\{C_b(0,j),T(b+1,j)\}\}$

$(b,j): 0 \le b < J \cup b < j \le J$

$Trellis[C(0,0)][\$] = [Q(0,\$), C(0,0), \$]$

for všechna pokrytí $C(0,b) \in TA$ **do**

  for všechna pokrytí $C_b(0,j) \in TA[C(0,b)]$ **do**

    for všechny překlady $\tilde{t} \in T(b+1,j) = TA[C(0,b)][C_b(0,j)]$ **do**

      $Q_{max}(j,\tilde{t}) = -\infty$

      for všechny překlady $\tilde{t}\,' \in Trellis[C(0,b)]$ **do**

        $Q(j,\tilde{t}) = Trellis[C(0,b)][\tilde{t}\,'] + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t}\,' \oplus \tilde{t}, s_{b+1}^{j})$

        if $Q(j,\tilde{t}) > Q_{max}(j,\tilde{t})$

          $Q_{max} = Q(j,\tilde{t})$

          $\tilde{t}\,'_{max} = \tilde{t}\,'$

      $Trellis[C(0,j)][\tilde{t}] = [Q_{max}(j,\tilde{t}), C(0,b), \tilde{t}\,'_{max}]$

$\hat{Q}_{max}(J+1,\$) = -\infty$

for všechna pokrytí $\tilde{t} \in Trellis[C(0,J)]$ **do**

  $Q(J+1,\tilde{t}) = Trellis[C(0,J)][\tilde{t}] + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t} \oplus \$, s_1^J)$

  if $Q(J+1,\tilde{t}) > \hat{Q}_{max}(J+1,\tilde{t})$

    $\hat{Q}_{max}(J+1,\$) = Q(J+1,\$)$

    $\tilde{t}_{max} = \tilde{t}$

Počínaje $\tilde{t}_{max}$ projdi pomocí odkazů na předchozí pokrytí celý *Trellis* a nalezni nejlepší překlad zdrojové věty složený z nejlepších dílčích překladů $\tilde{t}\,'_{max}$ odpovídajících pokrytím náležícím k nejlepší nalezené cestě stavovým prostorem

Obrázek 5.1: Algoritmus monotónního prohledávání pro nalezení překladu.

všech těch frází, které se v ní nachází. Tím je vytvořena množina všech možných rozdělení zdrojové věty na fráze a jim odpovídajících překladů. Protože budeme zdrojovou větu překládat monotónně, seřadíme tuto množinu vzestupně podle začátků zdrojových frází. Dostaneme tak tabulku TA, kde jsou každé možné pozici b označují začátek zdrojové fráze přiřazeny všechny možné konce j, které představují různá pokrytí zdrojové věty začínající pozicí b a končící pozicí j, a jim odpovídající možné překlady $T(b+1,j)$ z frázové tabulky. Tuto tabulku nyní procházíme po řádcích a pro každou odpovídající si dvojici pokrytí $C(0,b)$ a $C(0,j)$ nalezneme pro každý překlad $\tilde{t} \in T(b+1,j)$ nejlepší předchozí překlad $\tilde{t}'_{max} \in C(0,b)$, který maximalizuje skóre $Q(j,\tilde{t})$, které vznikne rozšířením překladové hypotézy končící frází $\tilde{t}'_{max}$ o

novou frázi $\tilde{t}$. Do tabulky *Trellis* si pak uložíme do položky pro pokrytí $C(0,j)$ překlad $\tilde{t}$ se skórem $Q_{max}(j,\tilde{t})$, pokrytím $C(0,b)$ a nejlepším předchozím překladem $\tilde{t}'_{max}$. Skóre $Q_{max}(j,\tilde{t})$ slouží v dalším kroku pro výběr nejlepšího předchozího překladu $\tilde{t}'_{max}$, neboť v pozdějším kroku platí: $Q_{max}(j,\tilde{t}) = Q(j_{prev} = b, \tilde{t}') = \mathit{Trellis}[C(0,b)][\tilde{t}']$. Informace o pokrytí $C(0,b)$ a překladu $\tilde{t}'_{max}$ pak slouží, po projití všech položek v tabulce *TA* a přidání hranice věty \$, k vytvoření nejlepšího překladu vstupní věty, který vznikne spojením překladů $\tilde{t}'_{max}$ ležících na nalezené nejlepší cestě stavovým prostorem. Díky použití všech možných rozdělení na vstupu algoritmu a výběru vždy nejlepší hypotézy v každém kroku algoritmu je současně vybrán nejen nejlepší překlad, ale také Viterbiho (nejpravděpodobnější) rozdělení zdrojové a cílové věty na fráze.

Takto definovaný algoritmus představuje prohledávání stavového prostoru do šířky, neboť pro každé současné pokrytí $C(0,b)$ jsou vytvořena a do tabulky *Trellis* uložena pro pozdější použití všechna možná pokrytí $C(0,j)$ a jim odpovídající překlady, která mohou vzniknout z daného pokrytí $C(0,b)$. Algoritmus také definuje obecně n-gramovou závislost nově vytvářeného pokrytí a jemu odpovídajícího překladu na předchozích překladech, nebo-li je definován frázový n-gram, kdy současná fráze závisí na $n-1$ předchozích frázích (v algoritmu na Obrázku 5.1 je pro jednoduchost použita bigramová závislost, kdy nový překlad závisí jen na minulém překladu). Při výpočtu rysů $h(\cdot,\cdot)$ lze tedy využít informace až o $n-1$ předchozích zvolených překladech. Jestliže T_{max} představuje maximální množství možných překladů, které jsou uvažovány v každém kroku, pak výsledná složitost algoritmu používajícího frázový n-gram je $\mathcal{O}(J \cdot PhL_{max} \cdot {T_{max}}^n)$. Použitím efektivních datových struktur a zahrnutím předpokladu, že ke každé zdrojové frázy existuje méně než PhL_{max} možných pokrytí, můžeme provádět velmi efektivní prohledávání. Monotónní prohledávání je vhodné pro jazyky, které mají podobné slovní uspořádání, jde např. o jazykové páry španělština – angličtina, francouzština – angličtina, čeština – slovenština, v našem případě pak také čeština – znakovaná čeština.

5.2 Nemonotónní prohledávání

V případě nemonotónního prohledávání jsou zdrojové fráze překládány v různém pořadí, které může být odlišné od jejich pořadí ve zdrojové větě. Dochází tak k přeuspořádání slov nejen uvnitř jednotlivých frází, ale také mezi celými frázemi (při překladu lze např. přeskočit několik zdrojových frází a přeložit až následující frázi). Existují dvě možnosti, jak zahrnout přeuspořádání frází do hledání nejlepšího překladu. Za prvé můžeme zdrojové věty předzpracovat tak, že je pak lze přeložit pomocí monotónního prohledávání. To znamená, že je třeba vytvořit na základě trénovacích dat nějaký přeuspořádávací model, který uspořádá fráze ve zdrojové větě tak, aby jejich pořadí odpovídalo pořadí příslušných frází v cílové větě (tj. přiřazení mezi frázemi (slovy) je monotónní). Tento model je pak aplikován na každou zdrojovou větu, která je překládána. Druhou možností je zahrnout přeuspořádání přímo do prohledávání. Přidáním možnosti přeuspořádání frází se problém prohledávání stane NP-obtížným problémem. Jako takový pak odpovídá problému obchodního cestujícího, který může být vyřešen pomocí varianty Held-Karp algoritmu [Held 62], jak je ukázáno v práci [Tillmann 01]. Výpočetní složitost tohoto řešení je ovšem exponenciální vzhledem k délce vstupní věty. Provádění neomezeného prohledávání je tedy v praxi nepraktické. Jednou z možností jak snížit výpočetní složitost, je omezit dovolená přeuspořádání frází. Tato přeuspořádávací omezení budou popsána dále. Nyní popíšeme algoritmus pro nemonotónní prohledávání.

Stejně jako v případě monotónního prohledávání zavedeme skóre $Q(C,j,\tilde{t})$, které představuje maximální skóre posloupnosti frází, která končí frází $\tilde{t}$ a pokrývá pozice dané pokrytím C a končí na pozici j zdrojové věty. Pro určení nejlepšího překladu jemuž odpovídá skóre $\hat{Q}$

Nemonotónní prohledávání

Vstup: Hash tabulka, která obsahuje všechna možná rozdělení zdrojové věty a jim odpovídající překlady:

$TA\{C_c(b), \{C_{c+j-i}(j), T(i,j)\}\}$

$c \in \{0, \ldots, J\}, 0 \leq b < J \cap b \leq i < j \leq J$

$Trellis[C_0(0)][\$] = [Q(C_0(0), \$), C_0(0), \$]$

for všechna pokrytí $C_c(b) \in TA$ **do**

  for všechna pokrytí $C_{c+j-i}(j) \in TA[C_c(b)]$ **do**

    for všechny překlady $\tilde{t} \in T(i,j) = TA[C_c(b)][C_{c+j-i}(j)]$ **do**

      $Q_{max}(C_{c+j-i}(j), \tilde{t}) = -\infty$

      for všechny překlady $\tilde{t}\,' \in Trellis[C_c(b)]$ **do**

        $Q(C_{c+j-i}(j), \tilde{t}) = Trellis[C_c(b)][\tilde{t}\,'] + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t}\,' \oplus \tilde{t}, s_i^j)$

        if $Q(C_{c+j-i}(j), \tilde{t}) > Q_{max}(C_{c+j-i}(j), \tilde{t})$

          $Q_{max} = Q(C_{c+j-i}(j), \tilde{t})$

          $\tilde{t}\,'_{max} = \tilde{t}\,'$

      $Trellis[C_{c+j-i}(j)][\tilde{t}] = [Q_{max}(C_{c+j-i}(j), \tilde{t}), C_c(b), \tilde{t}\,'_{max}]$

$\hat{Q}_{max}(C_{J+1}(J), J+1, \$) = -\infty$

for všechna pokrytí $\tilde{t} \in Trellis[C_J(J)]$ **do**

  $Q(C_{J+1}(J), J+1, \tilde{t}) = Trellis[C_J(J)][\tilde{t}] + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t} \oplus \$, s_1^J)$

  if $Q(C_{J+1}(J), J+1, \tilde{t}) > \hat{Q}_{max}(C_{J+1}(J), J+1, \tilde{t})$

    $\hat{Q}_{max}(C_{J+1}(J), J+1, \$) = Q(C_{J+1}(J), J+1, \$)$

    $\tilde{t}_{max} = \tilde{t}$

Počínaje $\tilde{t}_{max}$ projdi pomocí odkazů na předchozí pokrytí celý *Trellis* a nalezni nejlepší překlad zdrojové věty složený z nejlepších dílčích překladů $\tilde{t}\,'_{max}$ odpovídajících pokrytím náležícím k nejlepší nalezené cestě stavovým prostorem

Obrázek 5.2: Algoritmus nemonotónního prohledávání pro nalezení překladu.

dostaneme opět rekurzivní algoritmus dynamického programování:

$$Q(\emptyset, \$, 0) = 0 \tag{5.9}$$

$$Q(C_c(j), \tilde{t}) = \max_{\tilde{t}'} \left\{ Q(C_{c_{prev}}(j_{prev}), \tilde{t}') + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t}' \oplus \tilde{t}, s^j_{j_{prev}+1}) \right\} \tag{5.10}$$

$$Q(J+1, \$) = \max_{\tilde{t}} \left\{ Q(C_J(J), \tilde{t}) + \sum_{m=1}^{M} \lambda_m h_m(\tilde{t} \oplus \$, s_1^J) \right\}. \tag{5.11}$$

Jak je vidět tento algoritmus se až na označení uzlů pomocí dvojice $(C_c(j), \tilde{t})$ (nevystačíme již

jen s j, neboť pokrytí C nemusí být spojité, tj. může obsahovat mezery) neliší od algoritmu pro monotónní prohledávání. Předchozí pokrytí z kterého je vytvářeno nové pokrytí C, je tak dáno dvojicí $(C_{c_{prev}}(j_{prev}), \tilde{t}')$. Jestliže označíme počet pokrytých slov pokrytí C jako jeho *kardinalitu* c, pak $C_{c_{prev}}$ náleží do množiny pokrytí, která mají nižší kardinalitu než je c, tedy platí: $c - PhL_{max} \leq c_{prev} < c$. Dále také musí platit, že mezi C_c a $C_{c_{prev}}$ je nulový překryv, tj. $C_{c_{prev}} \cap \{j_{prev}, j\} = \emptyset$. Algoritmus pro nemonotónní prohledávání na obrázku 5.2 je pak tak, až na rozdíly dané odlišným značením uzlů, stejný jako algoritmus pro monotónní prohledávání (srovnej Obrázek 5.1). Hlavní rozdíl je ve vstupní tabulce *TA*, která je nyní vzestupně uspořádána podle kardinality c jednotlivých pokrytí odpovídajících různým rozdělení zdrojové věty. Nejdříve jsou tedy pokrytí s kardinalitou 1 pak 2 až nakonec pokrytí s kardinalitou J. V tomto pořadí jsou také postupně zpracovávána a jsou vytvářeny jednotlivé překladové hypotézy. Toto uspořádání nám zajistí, že je prohledávaný graf procházen v topologickém uspořádání a při výpočtu nového pokrytí C jsou tak známa všechna předchozí pokrytí C_{prev}, z kterých je nové pokrytí vytvářeno. Uvažovaná rozdělení v tabulce *TA* mohou být dána nějakým *přeuspořádávacím omezením*, tj. při tvorbě tabulky *TA* nejsou uvažována všechna možná rozdělení, ale jen ta, která splňují danou podmínku (např. o maximálním počtu vynechaných zdrojových slov, více viz následující podkapitola).

Výpočetní složitost tohoto algoritmu je pak $\mathcal{O}(\sum_{c=1}^{J} PhL_{max} \cdot \binom{J}{c} \cdot J \cdot {T_{max}}^n)$. To můžeme zjednodušit na $\mathcal{O}(J \cdot {T_{max}}^n \cdot 2^J)$, jestliže vezmeme v úvahu, že $\sum_{c=1}^{J} \binom{J}{c} = 2^J$. Je zřejmé, že složitost tohoto algoritmu je exponenciální vzhledem k délce zdrojové věty. Abychom urychlili prohledávání, používá se v praxi prohledávání založené na *výřezové strategii prohledávání* (angl. beam search strategy) [Jelinek 98], kdy se uplatňuje prořezávání hypotéz na různých úrovních [Zens 08]. Pro každou úroveň se používají dvě varianty prořezávání: *prahové prořezávání* (angl. threshold nebo také beam pruning) a *histogramové prořezávání* (angl. histogram pruning). Prahové prořezávání znamená, že hypotézu dále uvažujeme jen, když je její skóre blízko nejlepšímu skóre srovnatelné hypotézy. Histogramové prořezávání pak znamená, že v každém kroku uvažujeme jen N nejlepších hypotéz, zbylé hypotézy jsou zamítnuty. Výhoda prahového prořezávání spočívá v autokorelaci vzhledem k neurčitosti. Jestliže je neurčitost vysoká, tj. mnoho hypotéz má podobné skóre, uvažujeme v rámci prohledávání mnoho hypotéz. Je-li naopak neurčitost malá, je zde jen jedna nebo několik hypotéz s vysokým skóre, uvažujeme dále jen malý počet hypotéz. To ovšem také znamená, že zde není žádný horní limit na počet uvažovaných hypotéz. To ovšem není z praktického hlediska vhodné, neboť množství uvažovaných hypotéz může být příliš velké. Jednoduchou cestou, jak omezit toto množství hypotéz, je použít spolu s prahovým také histogramové prořezávání. Zapojení strategie prořezávání do prohledávání způsobí, že nejlepší nalezené řešení již nemusí být optimální, ale jen suboptimální. Poznamenejme, že strategii prořezávání lze pro zrychlení zpracování použít i v případě monotónního prohledávání.

V práci [Zens 08] jsou definovány následující prořezávací strategie:

1. Prořezávání pozorování: Je omezen počet možných překladů pro každou zdrojovou frázi, provádí se ještě před samotným prohledáváním. Jestliže je τ_o práh pro prořezání pozorování a $q(j, j')$ označuje maximální skóre jakéhokoliv překladu $\tilde{t}$ libovolné zdrojové fráze $s_j, ..., s_{j'}$:

$$q(j, j') = \max_{\tilde{t}} \{ \sum_{m=1}^{M} \lambda_m h_m(\tilde{t}, s_{j'}^{j}) \}. \tag{5.12}$$

   Uvažujeme tedy jen ty překlady $\tilde{t}$ pro které platí:

$$\sum_{m=1}^{M} \lambda_m h_m(\tilde{t}, s_{j'}^{j}) + \tau_o \geq q(j, j'). \tag{5.13}$$

Dále aplikujeme také ještě histogramové prořezávání s parametrem N_o. Tedy jestliže je zde více než N_o překladů pro danou zdrojovou frázi, uvažujeme dále jen N_o nejlepších překladů.

2. Lexikální prořezávání na pokrytí: Pro každé pokrytí C uvažujeme jenom omezenou množinu lexikálních hypotéz. Jestliže označíme práh prořezávání τ_L a $Q(C)$ je maximální skóre jakékoliv hypotézy s pokrytím C:

$$Q(C) = \max_{\tilde{t},j}\{Q(C,\tilde{t},j) + R(C,j)\}, \tag{5.14}$$

kde $R(C,j)$ označuje odhad zbytkového skóre, které dostaneme, jestliže dokončíme danou hypotézu, tj. pokrytí C rozšíříme tak, aby byla pokryta celá věta, tedy vytvoříme z C pokrytí $\{1,...,J\}$. Dále uvažujeme jen ty hypotézy se skórem $Q(C,\tilde{t},j)$ pro které platí:

$$Q(C,\tilde{t},j) + R(C,j) + \tau_L \geq Q(C). \tag{5.15}$$

Opět také aplikujeme histogramové prořezávání s parametrem N_L.

3. Lexikální prořezávání na kardinalitu: V tomto případě porovnáváme všechny lexikální hypotézy s danou kardinalitou c. Tedy porovnáváme i hypotézy s různým pokrytím. Jestliže označíme práh prořezávání τ_c a $Q(c)$ je maximální skóre jakékoliv hypotézy s kardinalitou c:

$$Q(c) = \max_{C:|C|=c,\tilde{t},j}\{Q(C,\tilde{t},j) + R(C,j)\}. \tag{5.16}$$

Dále uvažujeme jen ty hypotézy se skórem $Q(C,\tilde{t},j)$ pro které platí:

$$Q(C,\tilde{t},j) + R(C,j) + \tau_c \geq Q(c). \tag{5.17}$$

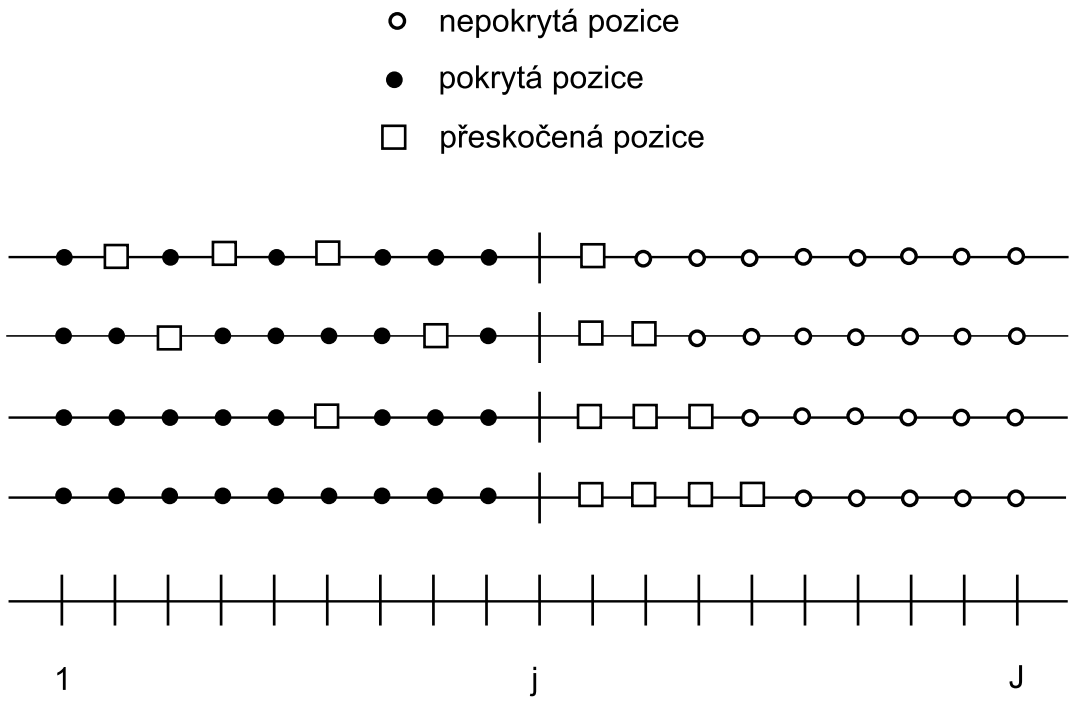
A opět také aplikujeme histogramové prořezávání s parametrem N_c.

4. Prořezávání pokrytí na kardinalitu: Pro danou kardinalitu c mezi sebou porovnáváme hypotézy s různými pokrytími C. Jak je definováno v rovnici 5.14, $Q(C)$ je maximální skóre jakékoliv hypotézy s pokrytím C. Jestliže označíme práh prořezávání τ_C, pak dále uvažujeme jen ty hypotézy pro které platí:

$$Q(C) + \tau_C \geq Q(c), \tag{5.18}$$

kde $Q(c)$ označuje maximální skóre jakékoliv hypotézy s danou kardinalitou c, tak jak je to definováno v rovnici 5.16. Nakonec aplikujeme také histogramové prořezávání s parametrem N_C. Jestliže vyloučíme pokrytí C, vyloučíme i všechny lexikální hypotézy s tímto pokrytím.

Během prořezávání porovnáváme hypotézy, které pokrývají různé části zdrojové věty. Je tedy důležité použít odhad zbytkového skóre $R(C,j)$ pro dokončení hypotézy, neboť jinak by se prohledávání soustředilo nejdříve jen na části zdrojové věty, které lze snadno přeložit. Heuristická funkce pro výpočet zbytkového skóre je popsána v práci [Och 02a]. Detailní popis algoritmu prohledávání s prořezáváním lze nalézt v práci [Zens 08].



Obrázek 5.3: Ilustrace IBM omezení [Tillmann 01].

5.3 Přeuspořádávací omezení

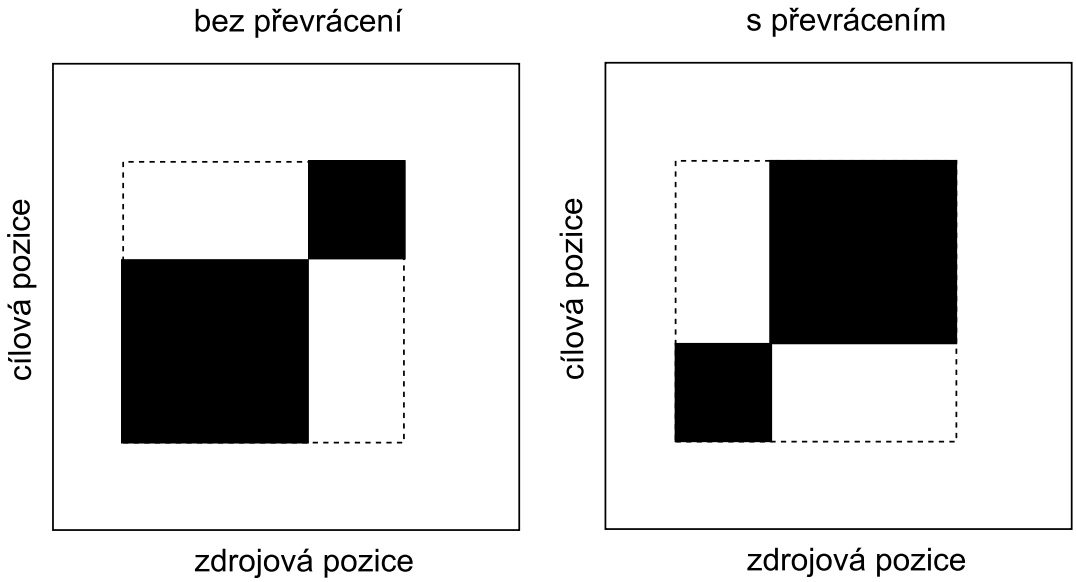
Cílem přeuspořádávacích omezení je, na úkor neomezeného přeuspořádávání, omezit počet hypotéz uvažovaných během prohledávání. Toto omezení je výhodné hned ze dvou důvodů. Za prvé použitím vhodných omezení lze problém prohledávání zjednodušit a tím potenciálně snížit množství chyb. A za druhé neomezené přeuspořádávání by bylo výhodné jen v případě, pokud bychom dokázali spolehlivě odhadnout přeuspořádávací pravděpodobnosti, což ale obvykle neplatí [Zens 08]. Nyní krátce popíšeme dvě často používaná omezení – IBM a ITG, jak jsou uvedena v práci [Zens 08].

5.3.1 IBM omezení

IBM omezení [Berger 96] je založeno na permutacích s omezeným posunem. Každou pozici ve zdrojové větě označíme jako pokrytou nebo nepokrytou. Na začátku jsou všechny pozice nepokryté. Nyní se tyto pozice prochází zleva doprava a je dovoleno pozici přeskočit a vrátit se k ní později. Vzhledem k IBM omezení může být další pozice jen jedna z k ještě nepokrytých pozic, tj. v každém čase může existovat jen maximálně $k-1$ přeskočených pozic. Fungování IBM omezení je zobrazeno na Obrázku 5.3.

Pro většinu cílových pozic je k dovolených zdrojových pozic. Pouze směrem ke konci věty je toto množství redukováno na počet zbývajících nepokrytých pozic. Jestliže n označuje délku zdrojové věty a r_n počet dovolených permutací pak v případě IBM omezení platí:

$$r_n = \begin{cases} k^{n-k} \cdot k! & n > k \\ n! & n \leq k. \end{cases}$$



Obrázek 5.4: Ilustrace ITG omezení.

Obvykle je k nastaveno na hodnotu 4. V tomto případě je asymptotická dolní a horní hranice 4^n, tj. $r_n \in \Theta(4^n)$.

5.3.2 ITG omezení

ITG (angl. inversion transduction grammar) omezení bylo představeno v práci [Wu 95], kde bylo navrženo pro použití k anotaci paralelního korpusu, závorkování, přiřazení a segmentaci. Využití základního ITG omezení pro automatický překlad pak bylo představeno v práci [Wu 96].

Na začátku máme zdrojovou větu jako posloupnost bloků, kde každá pozice ve větě má vlastní blok. Přeuspořádávací proces pak můžeme interpretovat jako postup, kdy zvolíme vždy dva následné bloky a spojíme je do jednoho bloku. A to buď ve stejném pořadí v jakém následují nebo je prohodíme. Jakmile jsou dva bloky spojeny, již se dále považují za jeden blok a mohou být dále spojovány jen jako celek. Alternativně lze ITG omezení popsat jako binární závorkování vstupní věty dvěma typy závorek: $[\bullet]$ pro monotónní spojení a $\langle\bullet\rangle$ pro převrácené spojení částí věty v závorkách. Např. uzávorkování $\langle[AB][\langle CD\rangle E]\rangle$ přepíšeme jako řetězec *DCEAB*. Fungování ITG omezení je zobrazeno na Obrázku 5.4.

Nyní prozkoumáme, kolik permutací dostaneme při použití ITG omezení. Permutace získané pomocí tohoto omezení můžeme reprezentovat jako binární strom, kde jsou vnitřní uzly buď černé (převrácené spojení) nebo bílé (monotónní spojení). To odpovídá parsovacímu stromu v jednoduché gramatice v [Wu 97]. Daná permutace pak může být pomocí tohoto omezení vytvořena řadou způsobů. Abychom dostali jedinečnou reprezentaci, každé permutace přidáme další podmínku pro vytvořený strom: jestliže je pravý syn uzlu vnitřní uzel, pak musí být obarven opačnou barvou. Díky této podmínce je každý strom unikátní a odpovídá parsovacímu stromu kanonické formy gramatiky v [Wu 97]. V práci [Shapiro 91] je ukázáno, že počet takovýchto binárních stromů s n uzly odpovídá $(n-1)$tému velkému *Schröderovu číslu* S_{n-1}:

$$(n+1)S_n = 3(2n-1)S_{n-1} - (n-2)S_{n-2},\ n \geq 2\ a\ S_0 = 1,\ S_1 = 2.$$

n	IBM				ITG
	k=2	3	4	5	
1	1	1	1	1	1
2	2	2	2	2	2
3	3	6	6	6	6
4	5	14	24	24	22
5	8	31	78	120	90
6	13	73	230	504	394
7	21	172	675	1 902	1 806
8	34	400	2 069	6 902	8 558
9	55	932	6 404	25 231	41 586
10	89	2 177	19 708	95 401	206 098
11	144	5 081	60 216	365 116	1 037 718
12	233	11 854	183 988	1 396 948	5 293 446
13	377	27 662	563 172	5 316 192	27 297 738
14	610	64 554	1 725 349	20 135 712	142 078 746
15	987	150 639	5 284 109	76 227 216	745 387 038

Tabulka 5.1: Počty permutací, které mohou být generovány pro IBM a ITG omezení.

Růst velikosti Schröderových čísel pak odpovídá: $(3+\sqrt{8})^n \approx 5,83^n$. V práci [Zens 08] je pak dále představeno rozšíření původního ITG omezení o použití spojitých a nespojitých frází a také jeho integrace do výřezového prohledávacího algoritmu. V Tabulce 5.1 jsou uvedeny počty permutací jako funkce délky zdrojové věty, které můžeme dostat použitím IBM a ITG omezení.

5.4 Rysy

V případě log-lineárního modelu lze použít a efektivně kombinovat různé zdroje informací o překladu dané zdrojové věty. V této části popíšeme rysy nejčastěji používané v oblasti automatického překladu. Ve většině případů se bude jednat o pravděpodobnostní model, jehož výstup (pravděpodobnost) je využíván při hledání překladu zdrojové věty.

5.4.1 Překladový model

Překladový model je hlavním rysem, který se při překladu používá. Jako takový popisuje vztah mezi zdrojovou a cílovou frází, tj. říká s jakou pravděpodobností je nějaká cílová fráze $\tilde{t}$ překladem dané zdrojové fráze $\tilde{s}$. Tyto frázové páry pak mohou být určeny různými způsoby např. ručně, jako je tomu v případě CSC korpusu nebo automaticky jednou z metod popsaných v Kapitole 4. Výpočet odhadu překladových pravděpodobností je založen na relativních

frekvencích výskytu jednotlivých frází [Koehn 03]:

$$\phi_T(\tilde{s}|\tilde{t}) = \frac{N(\tilde{s}|\tilde{t})}{N(\tilde{t})}, \tag{5.19}$$

kde $N(\tilde{s}|\tilde{t})$ označuje počet společných výskytů frází $\tilde{s}$ a $\tilde{t}$ a $N(\tilde{t})$ označuje počet výskytů samotné fráze $\tilde{t}$. Pro celou zdrojovou větu tak dostaneme výsledný rys:

$$h_T(t_1^I, a_1^K, s_1^J) = \sum_{k=1}^{K} \log \phi_T(\tilde{s}_k|\tilde{t}_k). \tag{5.20}$$

Log-lineární model nám umožňuje jednoduchou kombinaci různých modelů, jako výhodné se tak jeví použití i opačného překladového modelu:

$$h_{iT}(s_1^J, a_1^K, t_1^I) = \sum_{k=1}^{K} \log \phi_T(\tilde{t}_k|\tilde{s}_k), \tag{5.21}$$

jehož výpočet je obdobný jako v případě původního modelu, pouze zaměníme zdrojové a cílové fráze, přičemž jejich přiřazení zůstane stejné.

5.4.2 Jazykový model

Druhým velmi důležitým rysem je jazykový model. Úkolem jazykového modelu je zajistit, aby byl vytvořený překlad správný z hlediska cílového jazyka. Pomocí jazykového modelu se tedy modeluje správný slovosled cílového jazyka. Jako jedním z vhodných prostředků pro tento účel se ukázalo použití slovních n-gramů. Pravděpodobnost n-gramového jazykového modelu říká, s jakou pravděpodobností bude posloupnost $n-1$ slov pokračovat daným slovem. Odhad této pravděpodobnosti je opět založen na relativních frekvencích výskytu jednotlivých slov a jejich posloupností v korpusu obsahujícím texty v cílovém jazyce. Jazykový model k natrénování potřebuje jen jednojazyčná data, může tedy být natrénován, na rozdíl od překladového modelu, na významně větším množství dat, neboť jednojazyčná data lze získat mnohem jednodušeji než dvojjazyčná. Výsledný rys je tedy:

$$h_{LM}(t_1^I, a_1^K, s_1^J) = \sum_{i=0}^{I+1} \log p_{LM}(t_i|t_{i-n+1}^{i-1}). \tag{5.22}$$

Index i probíhá od 0 do $I+1$, protože na začátek a konec věty je vždy přidán symbol označující hranice vět. Čím vyšší n-gram použijeme tím lepší (z hlediska slovosledu) by měl být výsledný text. S vyššími n-gramy je ovšem spojen problém řídkosti trénovacích dat, kdy se hledané posloupnosti slov v datech nevyskytují vůbec nebo jen v malém počtu. Výsledkem jsou pak chybné odhady pravděpodobnosti. Navíc, ať už k natrénování použijeme sebevětší množinu trénovacích dat, vždy budou existovat posloupnosti slov, jejichž pravděpodobnost nebudeme znát. Naším cílem je ovšem přiřadit pravděpodobnost každé i v trénovacích datech neviděné posloupnosti slov. Toho lze docílit použitím vyhlazení získaných odhadů pravděpodobností. Princip vyhlazování je založen na použití pravděpodobnosti nižšího n-gramu v případě, že požadovaný n-gram nebyl viděn v trénovacích datech (jestliže v datech nebyl viděn ani 1-gram použije se pravděpodobnost 0-gramu, která je většinou dána jako apriorní pravděpodobnost výskytu libovolného slova a jako taková je určena jako 1/Velikost použitého slovníku). Nejčastěji se v praxi používají n-gramy druhého a třetího řadu, hovoříme pak o bigramech a trigramech (v případě 1-gramů hovoříme o unigramech, v případě 0-gramů o zerogramech).

V případě velkého množství dat lze použít i 4-gramy a 5-gramy. Poznamenejme, že problém jazykového modelování je široce diskutovanou problematikou, neboť nachází uplatnění v řadě problémů, jako je např. automatické rozpoznávání řeči, automatické rozpoznávání naskenovaného textu a řadě dalších úloh zabývajících se zpracováním přirozeného jazyka. Zde popsaný přístup k jazykovému modelování je pak nejčastěji používaným řešením. Podrobnější informace o jazykovém modelování lze nalézt v řadě publikací, za všechny jmenujme práci [Psutka 06].

5.4.3 Model slovní a frázové penalizace

V tomto případě se jedná o dva modely založené na jednoduchých heuristikách. Tyto modely ovlivňují průměrnou délku výsledné věty a použitých frází:

$$h_{WP}(t_1^I, a_1^K, s_1^J) = I \tag{5.23}$$

$$h_{PhP}(t_1^I, a_1^K, s_1^J) = K. \tag{5.24}$$

Model slovní penalizace jednoduše počítá délku cílové fráze. Představuje tak spolu se svojí váhou konstantní náklad na produkci jednoho cílového slova. Pomocí tohoto rysu tak můžeme jednoduše ovlivnit délku výsledného překladu. Jestliže totiž použijeme negativní váhu, jsou více penalizovány delší věty a překladový systém tak preferuje kratší překlady. V případě pozitivní váhy dostaneme opačný výsledek (preference delších překladů).

Podobně pak frázová penalizace představuje konstantní náklad na produkci jedné celé cílové fráze. Lze ji tedy využít k preferenci menšího počtu a tedy delších (záporná váha) nebo většího počtu a tedy kratších frází (kladná váha) ve výsledném překladu.

5.4.4 Model distanční penalizace

Tento model představuje základní přeuspořádávací model a díky své jednoduchosti nalezl široké využití v řadě překladových systémů. Je důležitý v případě nemonotónního překladu, neboť modeluje, v jakém pořadí mají být překládány zdrojové fráze. Je založen na přiřazení nákladů odpovídajících přechodu z cílové pozice současné zdrojové fráze na počáteční pozici další překládané zdrojové fráze:

$$h_{Dist}(t_1^I, a_1^K, s_1^J) = \sum_{k=1}^{K+1} q_{Dist}(b_k, j_{k-1}), \tag{5.25}$$

přičemž

$$q_{Dist}(j, j') = |j - j' + 1|. \tag{5.26}$$

Tento model tedy přiřadí nulové náklady v případě monotónního překladu, čím více frází je přeskupeno (přeskočeno) tím vyšší je distanční penalizace. Často se tento model používá také ve formě relativních přechodů, kdy je dvojice indexů b_k a j_{k-1} nahrazena absolutní hodnotou svého rozdílu, tj. modelují se tak jen přeskoky o daný počet slov bez ohledu na pozici ve zdrojové větě. Dále se také v praxi používá limit přeskoku D, který říká, jaký maximální počet slov může být přeskočen (tj. jaký je maximální rozdíl mezi koncem jedné a začátkem další překládané fráze).

Zde uvedené rysy představují jen základní rysy používané prakticky ve všech současných překladových systémech. Protože log-lineární model umožňuje jednoduchou a účelnou kombinaci různých rysů, je jedním z hlavních úkolů dneška v oblasti automatického překladu hledání dalších vhodných rysů. Cílem je nalézt rysy, které povedou k výběru lepších, tedy přirozenějších překladů. Tímto problémem se tak zabývá řada prací, uveďme např. již zmíněnou práci

[Zens 08], která popisuje další rysy zabývající se překladovými pravděpodobnostmi a také rysy související s přeuspořádáním cílových frází. Poznamenejme ještě, že hledání rysů vhodných pro překlad souvisí také s hledáním rysů pro výběr nejlepších frázových párů na základě log-lineárního modelu, který byl představen v Kapitole 4.

Kapitola 6

Trénování

Při trénování je naším cílem nastavit volné parametry systému pro automatický překlad na optimální hodnotu tak, aby přesnost překladu byla co nejvyšší. V případě log-lineárního modelu je tedy naším cílem nalézt optimální hodnoty vah λ_1^M v rovnici 1.4. V současné době je známo několik rozdílných způsobů, jak nalézt optimální hodnoty vah log-lineárního modelu pro automatický překlad. Tyto přístupy se liší především v použitém trénovacím kritériu, které je v průběhu trénování maximalizováno (minimalizováno) tak, aby byly nalezeny optimální hodnoty vah.

6.1 MMI trénování

Standardní kritérium pro trénování log-lineárního modelu je MMI kritérium, které může být odvozeno z principu maximální entropie, na němž jsou log-lineární modely založeny. Toto kritérium je ekvivalentní k populárnímu kritériu pro trénování parametrů založenému na maximalizaci pravděpodobnosti (angl. maximum likelihood (ML)) trénovacích dat. Cílem je nalézt hodnoty $\hat{\lambda}_1^M$, které maximalizují kritérium:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \log p_{\lambda_1^M}(\mathbf{t}_s|\mathbf{s}_s) \right\}, \tag{6.1}$$

kde $p_{\lambda_1^M}(\mathbf{t}_s|\mathbf{s}_s)$ je dáno rovnicí 1.4 a S je počet vět v množině dat určené k optimalizaci vah (obvykle se používá množina vývojových dat). Optimalizační problém v případě tohoto kritéria vede na jedinečné globální optimum a existují gradientní algoritmy, které garantují konvergenci k tomuto globálnímu optimu. Jedním z používaných algoritmů je GIS algoritmus [Darroch 72]. Tento přístup k nalezení optimálních hodnot vah byl použit v práci [Och 02b]. Aby však mohl být GIS algoritmus použit, bylo třeba vyřešit několik různorodých praktických problémů. Např. normalizace potřebná v rovnici 1.4 vyžaduje sumu přes velké množství všech možných vět, pro jejichž nalezení není znám efektivní algoritmus. Tato suma je nahrazena vzorkováním prostoru možných vět prostřednictvím velké množiny vysoce pravděpodobných vět, tj. pro každou větu je uvažováno N nejlepších kandidátů, kteří slouží jako aproximace prostoru všech možných vět. Na rozdíl např. od ASR máme pro každou větu ne jednu ale více referencí. To je třeba také zohlednit v trénovacím kritériu:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \frac{1}{R_s} \sum_{r=1}^{R_s} \log p_{\lambda_1^M}(\mathbf{t}_{s,r}|\mathbf{s}_s) \right\}, \tag{6.2}$$

kde R_s je počet referenčních překladů $\mathbf{t}_{s,1}, ..., \mathbf{t}_{s,R_s}$ pro větu $\mathbf{s}$. Dále také můžeme narazit na problém, že žádný z referenčních překladů není mezi N vybranými nejlepšími kandidáty. Za referenční překlad pak pouvažujeme ty kandidáty, kteří mají minimální počet chyb vzhledem ke kterémukoliv referenčnímu překladu.

6.2 MER trénování

Současnou standardní technikou pro nalezení hodnot vah je jejich optimalizace vzhledem ke kritériu použitému pro vyhodnocení přesnosti výsledného překladu [Och 03a]. Tento postup se nazývá trénování minimální chyby (MERT). Cílem je nalézt hodnoty $\hat{\lambda}_1^M$, které minimalizují kritérium (viz Kapitola 1.4):

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} E(r_s, \hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M)) \right\} \tag{6.3}$$

$$= \underset{\lambda_1^M}{\operatorname{argmin}} \left\{ \sum_{s=1}^{S} \sum_{k=1}^{K} E(\mathbf{r}_s, \mathbf{t}_{s,k}) \delta(\hat{\mathbf{t}}(s_s, \lambda_1^M), \mathbf{t}_{s,k}) \right\} \tag{6.4}$$

$$\hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M) = \underset{\mathbf{t} \in C_s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{t}, \mathbf{s}_s) \right\} \tag{6.5}$$

$$\delta(\hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M), \mathbf{t}_{s,k}) = \begin{cases} 1 & \text{jestliže } \hat{\mathbf{t}}(\mathbf{s}_s, \lambda_1^M) = \mathbf{t}_{s,k} \\ 0 & \text{jinak,} \end{cases}$$

kde E je chybové kritérium, $\mathbf{r}_s$ je referenční překlad zdrojové věty $\mathbf{s}_s$ a $C_s = \{\mathbf{t}_{s,1}, ..., \mathbf{t}_{s,K}\}$ je množina K různých překladů každé zdrojové věty $\mathbf{s}_s$. Jak je ukázáno v práci [Och 03a], lze tento postup použít pro různá chybová kritéria E (z již zmíněných např. BLEU, WER, PER). Nejlepších hodnot pro dané kritérium je vždy dosaženo při trénování s tímto kritériem (tj. nejvyšší hodnota BLEU skóre překladu testovacích dat je dosažena při MERT trénování s BLEU skóre jako chybovým kritériem E).

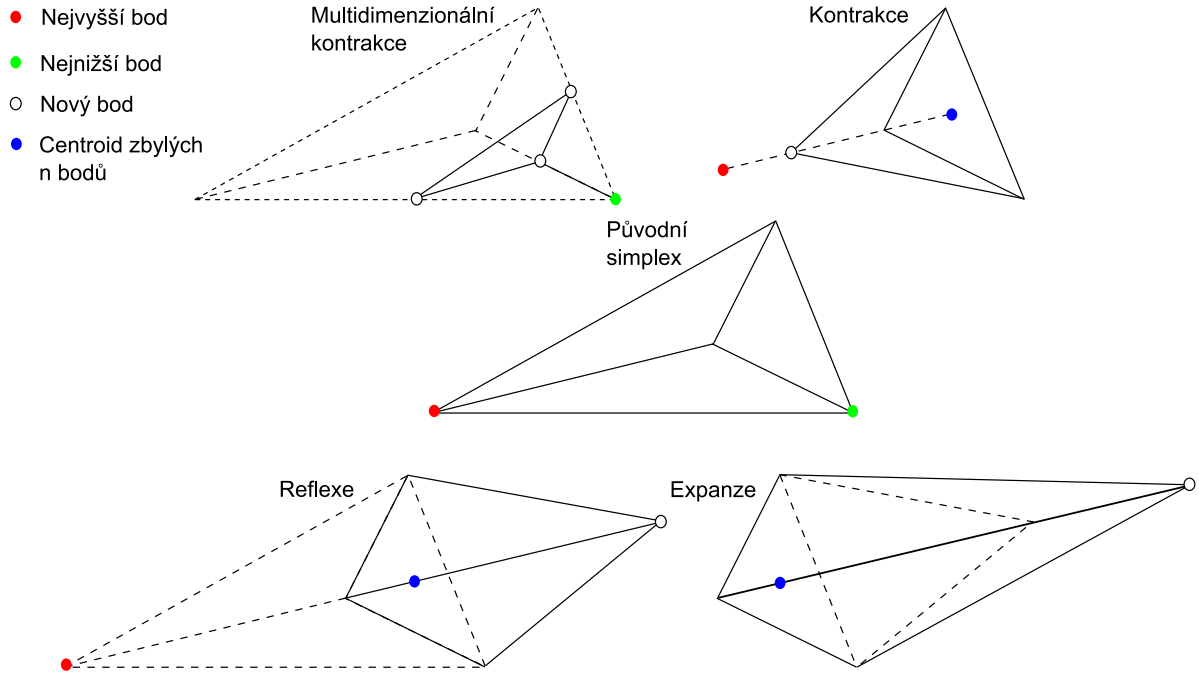
6.3 MEL trénování

Stejně jako v případě prohledávání i v případě trénování platí, že podle statistické teorie rozhodování by naším cílem mělo být maximalizovat střední zisk, tj. minimalizovat střední ztrátu (angl. minimal expected loss (MEL)):

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \sum_{\hat{\mathbf{t}}_s} Pr(\hat{\mathbf{t}}_s | \mathbf{s}_s) \cdot G(\hat{\mathbf{t}}_s, \mathbf{t}'_s) \right\}, \tag{6.6}$$

kde $\mathbf{t}'_s$ je referenční překlad zdrojové věty $\mathbf{s}_s$, $Pr(\hat{\mathbf{t}}_s|\mathbf{s}_s)$ představuje pravdivé rozdělení pravděpodobnosti, že $\hat{\mathbf{t}}_s$ je správným překladem dané zdrojové věty a G představuje funkci, která popisuje očekávaný zisk. Rozdělení Pr je však neznámé, takže se v praxi nahrazuje pravděpodobností log-lineárního modelu $p_{\lambda_1^M}(\mathbf{t}_s|\mathbf{s}_s)$ z rovnice 1.4. Dále je zde také suma přes všechny možné překlady $\hat{\mathbf{t}}_s$. Tato suma je v praxi aproximována použitím N nejlepších překladových hypotéz. Výsledné trénovací kritérium má tedy podobu:

$$\hat{\lambda}_1^M = \underset{\lambda_1^M}{\operatorname{argmax}} \left\{ \sum_{s=1}^{S} \sum_{\hat{\mathbf{t}}_s} p_{\lambda_1^M}(\hat{\mathbf{t}}_s | \mathbf{s}_s) \cdot G(\hat{\mathbf{t}}_s, \mathbf{t}'_s) \right\}. \tag{6.7}$$



Obrázek 6.1: Čtyři základní operace se simplexem v třídimenzionálním prostoru.

Z porovnání s kritériem pro MER trénování, je vidět, že je, na rozdíl od hypotézy s maximální aposteriorní pravděpodobností, využito celé pravděpodobnostní rozdělení [Zens 08].

Práce [Zens 08] pak obsahuje porovnání výše zmíněných trénovacích kritérií pro stejnou úlohu překladu. Jako nejlepší trénovací kritérium se jeví použití kritéria, které minimalizuje očekávanou ztrátu ať už v kombinaci s MAP nebo MBR dekódovacím pravidlem (obě poskytují téměř identické výsledky). Práce [Smith 06] pak také přináší srovnání tří předešlých kritérií a navíc vlastní metodu pro výpočet minimální očekávané ztráty a vlastní metodu pro optimalizaci tohoto kritéria založenou na deterministickém žíhání (angl. deterministic annealing). Jako nejlepší se opět, v případě automatického překladu, jeví kritérium založené na minimalizaci očekávané ztráty.

6.4 Nelder-Mead algoritmus

Optimalizaci všech výše uvedených kritérií lze provést např. pomocí Nelder-Mead algoritmu (nebo též algoritmus sestupného simplexu (angl. downhill simplex))[Nelder 65]. Jedná se o nelineární optimalizační algoritmus pro optimalizaci funkcí více proměnných. Princip hledání minima je založen na konceptu simplexu nebo n-simplexu pro funkci n proměnných, což je n-dimenzionální obdoba trojúhelníku (tj. pro funkci n proměnných použijeme n-simplex, který má $n+1$ vrcholů). V každém kroku algoritmu se hledá nový simplex, jehož nejvyšší vrchol má nižší hodnotu ohodnocení, než byla nejvyšší hodnota ohodnocení původního simplexu. Přitom se využívají čtyři základní operace se simplexem. Buď dojde k jeho expanzi, kontrakci nebo reflexi nejvyššího (v případě hledání minima) bodu prostřednictvím centroidu zbylých bodů nebo ke smrštění celého simplexu směrem k nejnižšímu bodu (viz Obrázek 6.1). Výhodou této metody je, že nepotřebuje znát derivaci optimalizované funkce (v průběhu výpočtu se zjišťují jen funkční hodnoty pro dané body), lze ji tedy využít pro řadu různých problémů. Nevýhodou pak může být pomalejší konvergence a možnost uváznutí v lokálním minimu (to

lze kompenzovat použitím různých počátečních podmínek při spuštění algoritmu).

Kapitola 7

Experimenty

V této kapitole popíšeme výsledky experimentů s obousměrným překladem mezi češtinou a znakovanou češtinou. K tomu budou využita data a informace shromážděné ve vytvořeném CSC korpusu. Bude vyzkoušena nová metoda pro výběr frází popsaná v části 4.4 a porovnána s ručně vytvořenými frázemi a frázemi získanými standardní metodou pro výběr frází. Porovnání se uskuteční prostřednictvím srovnání přesnosti překladu s danou frázovou tabulkou. Dále bude představen vlastní dekodér, který byl vytvořen podle poznatků v Kapitole 5. Výkonnost tohoto dekodéru bude srovnána s referenčním dekodérem, který představuje standard mezi současnými frázovými dekodéry. Bude otestována úspěšnost překladu základního systému a navrženy a otestovány úpravy tohoto systému pro obousměrný překlad mezi češtinou a znakovanou češtinou.

7.1 Evaluační kritéria

V průběhu výzkumu v oblasti automatického překladu byla vyvinuta řada kritérií pro porovnání různých překladů. Již z existence více ekvivalentních překladů příslušejících jedné zdrojové větě, je zřejmé, že porovnání dvou různých překladů jedné věty není triviální záležitost. Z tohoto hlediska se tak jako nejlepší jeví porovnání překladů prostřednictvím člověka (ideálně rodilého mluvčího daného cílového jazyka). Tento přístup však za prvé vede jen na subjektivní míru porovnání a za druhé je především časově a případně i finančně velmi náročný. Z těchto důvodů je tak jen velmi omezeně použitelný při vývoji automatického překladového systému, kdy požadujeme co nejrychlejší ověření přínosu provedené změny na přesnost překladu. Hlavní snahou tak bylo a stále je vyvinout kritérium, které by bylo možné použít pro automatické vyhodnocení přesnosti překladu a urychlit tak vývoj systémů pro automatický překlad. Ačkoliv se podařilo v tomto případě dosáhnout dílčích úspěchů (např. BLEU a NIST kritéria), neexistuje dosud jedno obecně přijímané a používané kritérium. Při vyhodnocení se tak upřednostňuje použití více kritérií. Z hlediska jejich povahy je lze rozdělit do dvou skupin na: *chybové míry* a *míry správnosti*.

7.1.1 Chybové míry

- SER (angl. sentence error rate) - Tato míra počítá poměr mezi vytvořenými větami, které jsou odlišné od referenčních (tj. jde o chybný překlad) a všemi vytvořenými větami:

$$SER = \frac{S_E}{S_G}, \tag{7.1}$$

kde S_E jsou všechny věty odlišné od referenčních a S_G jsou pak všechny přeložené věty. Tato míra se obvykle nepoužívá, neboť při překladu obecnějších a zvláště delších vět je díky vysoké míře nejednoznačnosti málokdy výsledný překlad identický s referenčním. V našem případě, kdy překládáme v omezené doméně, kde jsou časté krátké věty, je ovšem tato míra použitelná.

- WER (angl. word error rate) [Levenshtein 66] - Tato míra je převzata z oblasti automatického rozpoznávání řeči a počítá minimální počet operací substituce, vložení a smazaní, které musí být provedeny, aby vytvořená věta přesně odpovídala referenční větě. Tato míra se také nazývá Levenshteinova (editační) vzdálenost mezi dvěma řetězci (větami) a je dána předpisem:

$$WER = \frac{S + I + D}{N_R}, \tag{7.2}$$

kde S je počet nahrazených slov, I je počet vložených slov, D je počet smazaných slov a konečně N_R je počet všech slov v referenčním překladu.

- PER (angl. position-independent word error rate) [Tillmann 97b] - Jak název napovídá, tato míra porovnává dvě věty bez ohledu na pořadí jejich slov. Obchází tak nevýhodu WER míry, která vyžaduje stejné pořadí slov. Pořádek slov přijatelného překladu však může být odlišný od pořádku slov referenční věty a WER míra je pak zavádějící. Slova z referenční věty, která nejsou v generované větě se počítají jako substituce a v závislosti na délce vytvořené věty se chybějící nebo přebývající slova počítají jako vložení nebo smazaní:

$$PER = \begin{cases} N_R > N_H : \frac{S+I}{N_R} & S = N_H - R,\ I = N_R - N_H \\ N_R \leq N_H : \frac{S+D}{N_R} & S = N_R - R,\ D = N_H - N_R, \end{cases} \tag{7.3}$$

kde R jsou slova shodná pro obě věty (tedy správné překlady) a N_H je délka vytvořené věty.

- TER (angl. translation edit rate) [Snover 06] - Jedná se o poměrně novou míru, která je rozšířením WER míry. K již známým operacím přibyl ještě posun celých frází. Tato míra tak odráží množství práce, které musí provést překladatel, jestliže chce vytvořenou větu upravit do podoby správného překladu.

7.1.2 Míry přesnosti

- BLEU (angl. biLingual evaluation understudy) - Toto kritérium bylo navrženo v práci [Papineni 01] a je v současnosti nejpoužívanějším kritériem pro výpočet přesnosti překladu. Počítá modifikovanou n-gramovou přesnost vytvořeného překladu s ohledem na referenční překlad a s penalizací pro příliš krátké vytvořené věty. BLEU skóre pro vytvořený překlad t_1^I a referenční překlad $t_1'^{I'}$ se počítá jako:

$$BLEU(t_1^I, t_1'^{I'}) = BP(I, I') \cdot \prod_{n=1}^{4} Prec_n(t_1^I, t_1'^{I'})^{1/4} \tag{7.4}$$

s

$$BP(I, I') = \begin{cases} 1 & \text{jestliže } I \geq I' \\ e^{1-I/I'} & \text{jestliže } I < I' \end{cases} \tag{7.5}$$

$$Prec_n(t_1^I, t_1'^{I'}) = \frac{\sum_{w_1^n} \min\{C(w_1^n|t_1^I), C(w_1^n|t_1'^{I'})\}}{\sum_{w_1^n} C(w_1^n|t_1^I)}. \tag{7.6}$$

Kde $C(w_1^n|t_1^I)$ znamená počet, kolikrát se daný n-gram w_1^n vyskytl ve větě t_1^I. Jmenovatel n-gramové přesnosti pak počítá množství všech n-gramů ve vytvořené větě, tj. je roven: $I - n + 1$.

- NIST [Doddington 02] - Toto kritérium je podobné BLEU kritériu, neboť také počítá n-gramovou přesnost a používá penalizaci pro příliš krátké věty. Na rozdíl od geometrického průměru v případě BLEU kritéria však používá aritmetický průměr n-gramových počtů. Navíc jsou jednotlivá n-gramová skóre vážena svou informační hodnotou. NIST skóre pro vytvořený překlad t_1^I a referenční překlad $t_1'^{I'}$ se počítá jako:

$$NIST(t_1^I, t_1'^{I'}) = \sum_{n=1}^{N} BP(I, I') \cdot \sum_{\substack{\forall w_1^n : w_1^n \in t_1^I \\ \cap w_1^n \in t_1'^{I'}}} I(w_1^n) / \sum_{w_1^n \in t_1^I} (1) \tag{7.7}$$

s

$$BP(I, I') = e^{\beta \cdot \log_2 \min(L_{t_1^I} / \bar{L}_{t_1'^{I'}}, 1)} \tag{7.8}$$

$$I(w_1^n) = \log_2 \frac{N(w_1^{n-1})}{N(w_1^n)}. \tag{7.9}$$

Kde β je zvolena tak, aby penalizační faktor $BP = 0,5$, jestliže počet slov na výstupu je roven dvěma třetinám průměrného počtu slov v referenčním překladu. $L_{t_1^I}$ je pak počet slov ve vytvořeném překladu t_1^I a $\bar{L}_{t_1'^{I'}}$ je průměrný počet slov ve všech referenčních větách $t_1'^{I'}$. A nakonec $N(w_1^{n-1})$ je počet objevení n-gramu w_1^{n-1} a $N(w_1^n)$ obdobně počet objevení n-gramu w_1^n ve všech referenčních větách.

Jestliže je k dispozici více referenčních překladů ke každé vstupní větě, lze použít verze těchto kritérií pro výpočet s více referencemi. Jako na primární kritérium lze pohlížet na BLEU kritérium, které díky své možnosti automatického vyhodnocení znamenalo velký posun při vývoji systémů automatického překladu. V práci [Papineni 01] bylo ukázáno, že existuje vysoká korelace mezi posouzením pomocí tohoto kritéria a posouzením, které by provedl člověk. BLEU kritérium se také v současnosti používá v řadě MT evaluací jako oficiální kritérium. A to i přesto, že v práci [Callison-Burch 06] bylo ukázáno, že toto kritérium favorizuje frázové systémy na úkor nefrázových (důvodem je to, že BLEU kritérium favorizuje totiž ty systémy, které sdílejí očekávaný referenční slovník). Na druhou stranu bylo však také v té samé práci prokázáno, že jestliže porovnáváme různé varianty stejného systému nebo jen systémy založené na frázích, je toto kritérium adekvátní. Protože se v následujících experimentech chystáme porovnávat různé varianty jednoho systému a různé frázové systémy, je BLEU kritérium pro naše účely vhodným kritériem. Vytvoření BLEU kritéria vedlo k zájmu v oblasti zkoumání kritérií pro ohodnocení kvality překladu a byla vytvořena řada dalších kritérií, která se snažily překonat omezení daná konstrukcí BLEU kritéria a dosáhnout ještě lepší shody s posouzením, které by provedl člověk. Za všechny zmiňme dvě kritéria, která se v práci [Callison-Burch 07], která je výstupem tvůrčí dílny zabývající se porovnáním existujících překladových systémů a různých evaluačních kritérií při překladu řady jazyků do a z angličtiny, umístila z hlediska shody s posouzením provedeným lidmi před BLEU kritériem. Jde o kritérium, které je založeno na překrývání sémantických rolí (angl. semantic role overlap (SRO)) [Giménez 07] a METEOR kritérium [Banerjee 05]. SRO kritérium využívá syntaktické a sémantické rozbory vytvořeného překladu a referenčních překladů, na jejichž základě je pak určena lexikální shoda mezi sémantickými rolemi stejného typu ve vytvořeném a referenčním překladu. Díky tomu je překonána hlavní nevýhoda BLEU kritéria, kterou je lexikální orientace kritéria, tj. soustředění se jen

na porovnání slovní shody mezi překlady. Podobně METEOR kritérium využívá unigramovou přesnost a úplnost mezi vytvořeným a referenčním překladem, přitom se využívá flexibilní shody mezi slovy založené na lematizaci (hledání základního tvaru slova) a synonymech (používá se databáze WordNet[1]). Tím se také obchází přílišná závislost předchozích kritérií na shodě mezi konkrétními lexikálními tvary jednotlivých slov. Nevýhodou těchto dvou kritérií je pak to, že vyžadují použití pokročilých nástrojů pro zpracování přirozeného jazyka (jde např. o lematizátor, morfologický analyzátor, tagger, parser, databáze WordNet atd.), které však v případě znakované češtiny neexistují, takže tato kritéria nelze v našem případě použít.

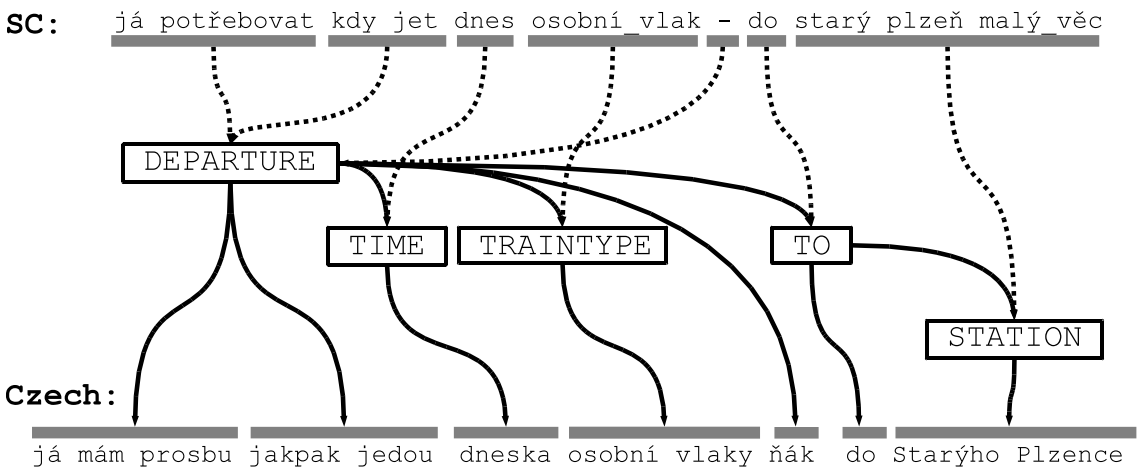
Hlavním problémem, kterému totiž čelíme při porovnání různých překladů, je potřeba zjistit, zda rozdílné slovní konstrukce dávají stejný smysl. Tento problém úzce souvisí s nejednoznačností přirozeného jazyka, kdy lze věci stejného mínění vyjádřit rozdílnými slovy (ať již jde např. o synonyma v podobě slov nebo části či dokonce celé věty, které mají stejný význam, ale používají různé syntaktické konstrukce i různá slova). Tento fenomén se v případě MT projevuje již přítomností obvykle více ekvivalentních referenčních překladů téže zdrojové věty. Při vytváření překladového systému pak dále také tím, že k jedné zdrojové frázi lze nalézt více odpovídajících si překladů, z kterých je pak třeba vybírat při překladu testovacích dat. Často se tak může stát, že při automatickém překladu dojde k vybrání sémantického synonyma (tj. majícího stejný význam), které však není obsaženo v žádném referenčním překladu, což v případě lexikálně orientovaného kritéria (všechna výše zmíněná kritéria s výjimkou SRO a METEOR kritéria) vede k nežádoucí penalizaci. Lexikální kritéria totiž většinou tíhnou k doslovné shodě s referenčním překladem (např. kritéria SER, PER, WER, TER, BLEU, NIST), což, jak ukazuje existence více ekvivalentních referenčních překladů, není nutná podmínka pro správný překlad. Jestliže tedy chceme adekvátně porovnávat různé překlady, je třeba zapojit i porovnání pomocí odpovídajících si významů (jako je tomu např. v případě SRO kritéria), což představuje nutnou podmínku správnosti překladu, a ne jen lexikálních tvarů. Jak již bylo řečeno, SRO kritérium však není pro překlad mezi češtinou a znakovanou češtinou použitelné. Díky použití HHTT korpusu jako základu CSC korpusu se však v případě znakované češtiny nabízí pro vyhodnocení smyslu překladu použití jiného, nového kritéria založeného na sémantické dimenzi značkovacího schématu dialogových aktů. Toto kritérium nyní popíšeme podrobněji v následující části.

7.1.3 SDO kritérium

SDO kritérium (angl. semantic dimension overlap (SDO)) je založeno na překryvu sémantické anotace přiřazené vytvořenému a referenčnímu překladu. Pomocí sémantického parseru je ke každému překladu (vytvořenému a referenčnímu) vytvořen odpovídající sémantický strom. Hodnota kritéria je pak daná velikostí shody mezi vytvořenými stromy. K parsování je využit parser, který je podrobně popsán v pracích [Jurčíček 07, Jurčíček 08]. Jedná se o HVS (angl. hidden vector state) parser, který byl rozšířen o možnost levého větvení, které je nezbytné v případě zpracování češtiny. Abychom ovšem mohli tento parser použít pro vyhodnocení kvality překladu do znakované češtiny, je třeba mít odpovídající trénovací data. Ta byla získána tak, že jsme ručně vytvořené parsovací stromy odpovídající jednotlivým promluvám v HHTT korpusu přiřadili také k jejich překladům do znakované češtiny. Při trénování parseru se předpokládá, že pořadí sémantických konceptů odpovídá pořadí slov ve větě. Protože přiřazení mezi češtinou a znakovanou češtinou je monotónní, platí, že lze jeden a ten samý sémantický strom přiřadit jak české promluvě, tak jejímu překladu do znakované češtiny (viz příklad na Obrázku 7.1).

V práci [Švec 07] je pro porovnání shody mezi sémantickými stromy navržen následující

[1] http://wordnet.princeton.edu/



Obrázek 7.1: Sémantický strom české promluvy a odpovídajícího překladu do ZČ.

postup založený na vzájemném zarovnání jednotlivých konceptů referenčního a hypotetického stromu. Toto zarovnání lze získat pomocí algoritmu pro výpočet vzdálenosti řazených stromů [Klein 98]. Při tomto výpočtu je určena celková minimální cena operací vedoucích k převedení dvojice stromů na stejný strom. Přitom jsou uvažovány tyto operace: vypuštění uzlu referenčního stromu (označme D - chyba vypuštění), vypuštění uzlu hypotetického stromu (označme I - chyba vložení), porovnání referenčního a hypotetického uzlu (označme S - chyba substituce, jestliže jsou uzly rozdílné, jinak jde o správné zarovnání). V případě vypuštění uzlu je tento uzel nahrazen svým podstromem. Cena pro všechny operace D, I, S je jedna. Ilustrace postupu při převodu stromů na stejný strom je na Obrázku 7.2.

Minimální cena provedených operací je pak získána pomocí výpočtu Levenshteinovy vzdálenosti mezi dvěma řetězci získanými z těchto stromů pomocí Eulerovské cesty stromem, přičemž do řetězce je každý uzel zahrnut právě dvakrát (jednou při sestupu k následníkům, podruhé při opuštění tohoto podstromu).

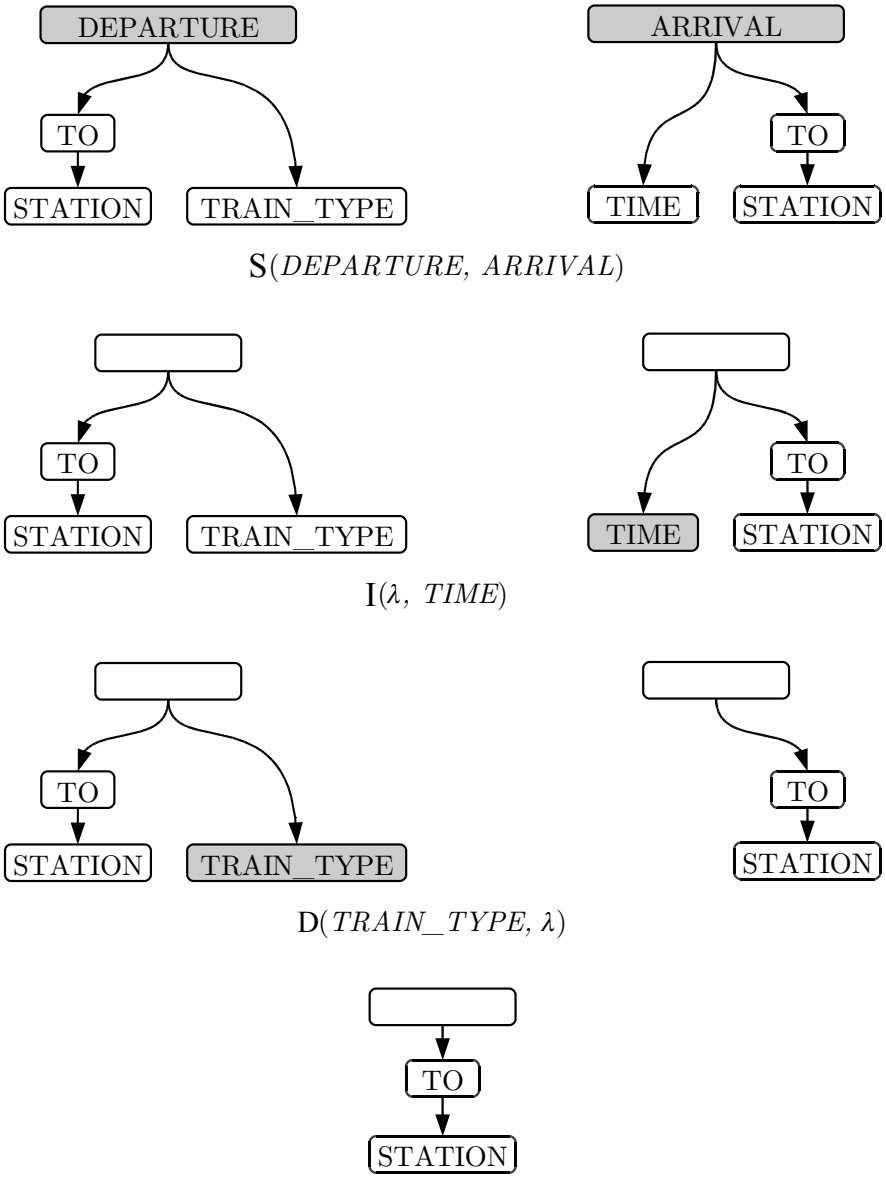
Na základě předchozích operací jsou v práci [Švec 07] definovány dvě následující míry vhodné pro vyhodnocení přesnosti sémantické analýzy. První mírou je konceptová přesnost CAcc (angl. concept accuracy), která je definována jako:

$$\begin{aligned} CAcc &= \frac{N - D - S - I}{N} \cdot [100\%] \\ &= \frac{H - I}{N} \cdot [100\%], \end{aligned} \tag{7.10}$$

kde H označuje počet správně zarovnaných konceptů v seznamu operací zarovnávajících referenční a hypotetický sémantický strom a N pak počet konceptů referenčního stromu. Druhou mírou je konceptová správnost CCorr (angl. concept correctness), která je definována jako:

$$CCorr = \frac{H}{N} \cdot [100\%]. \tag{7.11}$$

Z těchto dvou kritérií vybereme pro naše potřeby vyhodnocení shody mezi dvěma překlady CAcc míru, která dokáže adekvátně změřit míru shody mezi dvěma stromy.



Obrázek 7.2: Postupná úprava dvojice stromu na společný podstrom [Švec 07].

SDO kritérium nám v tomto případě tedy slouží k porovnání shody mezi dvěma překlady z hlediska jejich sémantické anotace určené pomocí sémantického parseru. Protože byla tato použitá sémantická anotace navržena pro potřeby řízení dialogu (viz část 3.2.3), můžeme na sémantický parser pohlížet jako na posuzovatele překladů z hlediska jejich správnosti v případě, že by byly použity jako vstup nebo výstup automatického dialogového systému. Při porovnání překladů pomocí SDO kritéria slouží sémantický strom referenčního překladu jako referenční a je porovnán se sémantickým stromem vytvořeného hypotetického překladu. Pokud jsou tyto dva stromy stejné, lze říci, že oba dané překlady jsou z hlediska jejich použití v automatickém dialogovém systému ekvivalentní, tj. případné odlišnosti v hypotetickém překladu nemají vliv z hlediska jeho sémantické anotace vytvořené použitým sémantickým parserem.

Pro vyhodnocení experimentů využijeme všechna zde uvedená kritéria s výjimkou kritéria TER. Pro výpočet SER a PER kritérií využijeme vlastní skript. Pro výpočet WER kritéria

pak program `HResults.exe` ze softwarového balíku HTK[2]. BLEU a NIST skóre je počítáno pomocí skriptu `mteval-v11b.pl`[3]. K výpočtu CAcc míry SDO kritéria využijeme skripty, které jsou součástí programového balíku pro trénování HVS parseru[4].

7.1.4 Statistická významnost a konfidenční intervaly

Test statistické významnosti využijeme k tomu, abychom mohli rozhodnout, zda je rozdíl mezi výsledky překladu dvou různých systémů významný, tj. pravděpodobnost, že tyto dva výsledky pocházejí ze stejné pravděpodobnostní distribuce, je nižší než zvolená hladina významnosti, na které ověřujeme hypotézu o stejné distribuční funkci. Na přesnost výsledného překladu můžeme totiž pohlížet jako na náhodnou proměnnou, která je závislá na použitých trénovacích a testovacích datech. Pro různá trénovací a testovací data můžeme při stejném nastavení dostat různé výsledky. Naším cílem je pak určit přesnost překladu pro neomezená testovací data, tj. zajímá nás, jak se daný systém při překladu chová bez ohledu na povahu testovacích dat. Protože však samozřejmě taková neomezená testovací data nemáme k dispozici, snažíme se určit alespoň odhad distribuční funkce, která popisuje pravděpodobnostní rozložení přesností překladu daného systému a tím získat odhad přesnosti překladu daného systému bez ohledu na povahu testovacích dat. K tomu, abychom však mohli určit odhad distribuční funkce, je třeba znát hodnoty přesnosti překladu pro různé testovací množiny. V případě omezeného množství dat k dispozici a neznalosti typu distribuce náhodné proměnné se nabízejí dvě možnosti, jak tyto hodnoty získat.

První možností je metoda *křížové validace* (angl. cross-validation), kdy jsou dostupná data opakovaně rozdělena na trénovací, vývojovou (je-li třeba) a testovací část. A to tak, že v každém kroku je testovací (případně vývojová) množina vybrána odlišným způsobem od předešlých množin (jestliže chceme např. provést křížovou validaci pro sto různých rozdělení, vybereme v prvním kroku do testovací množiny každou stou větu, ve druhém kroku pak každou stou první větu atd. až získáme sto různých testovacích množin). Vzhledem k tomu, že pro odhad potřebujeme řádově desítky hodnot (optimální je pak více než např. tisíc hodnot) a protože množství dat, které máme k dispozici je obvykle omezené (takže se snažíme o použití co nejvíce dat k trénování a navíc velikost testovací množiny je dána jako poměr velikosti celého korpusu k počtu zamýšlených rozdělení, takže výsledná velikost testovací množiny nemusí být dostatečná) a dále také, že trénování systému i samotný překlad testovací množiny mohou zabrat velké množství času, je většinou v praxi tato metoda nepoužitelná.

Druhou možností je použití metody *převzorkování* (angl. resampling), kdy jsou nové hodnoty vytvářeny na základě originálních změřených hodnot. K určení distribuční funkce vzorků se používá metoda *bootstrap*, podmínkou použitelnosti této metody je, že použité vzorky jsou reprezentanty souboru, ze kterého byly vybrány. Použití této metody pro určení statistické významnosti a konfidenčních intervalů v oblasti automatického překladu bylo navrženo v pracích [Koehn 04, Zhang 04]. Převzorkování v tomto případě funguje následovně. Nejprve je přeložena původní testovací množina, která obsahuje N vět, tím získáme množinu překladů T_0, další testovací množiny o velikosti N pak vytváříme tak, že náhodně vybíráme s vracením věty z původní množiny T_0. Opakováním tohoto výběru M-krát dostaneme sadu M uměle vytvořených testovacích množin (nazývaných také *bootstrapové vzorky*) $T_1, ..., T_M$ a jim odpovídajících hodnot přesnosti překladu, z kterých můžeme nyní spočítat parametry distribuční funkce, které nás zajímají, např. medián a konfidenční intervaly. Ve zmíněných pracích je i

[2] http://htk.eng.cam.ac.uk/
[3] http://www.nist.gov/speech/tools/
[4] http://code.google.com/p/extended-hidden-vector-state-parser/

navržen postup, jak metodu bootstrap použít k výpočtu statistické významnosti rozdílu mezi dvěma různými systémy. Mějme dva překladové systémy A a B. Vytvoříme opět sadu M testovacích množin pro každý systém a dostaneme tedy sady $T_0^A, ..., T_M^A$ a $T_0^B, ..., T_M^B$. Systém A dosáhl skóre S_0^A na množině T_0^A a systém B skóre S_0^B na množině T_0^B, jejich rozdíl je tedy $\delta_0 = S_0^A - S_0^B$. Jestliže nyní postupně spočteme rozdíly ve skóre pro zbylých M vzorků, tj. $\delta_i = S_i^A - S_i^B,\ i \in 1, ..., M$ dostaneme posloupnost rozdílů $\delta_0, ..., \delta_M$. Nyní vytvoříme např. 95 procentní konfidenční interval (tj. seřadíme skóre podle velikosti a najdeme 2,5 percentil a 97,5 percentil), pokud takto vytvořený interval neprotíná nulu, můžeme říct, že dané systémy jsou s 95 procentní pravděpodobností rozdílné, tedy rozdíl v přesnosti jejich překladu je statisticky významný na 5 procentní hladině významnosti [Zhang 04]. Nebo můžeme spočítat kolikrát je skóre S_i^A větší než skóre S_i^B. Je-li např. skóre systému A v 95 procentech větší než skóre systému B, pak lze opět říci, že rozdíl v přesnosti překladu těchto dvou systémů je statisticky významný na 5 procentní hladině významnosti [Koehn 04].

Později se však v pracích [Riezler 05, Collins 05] objevila kritika použití metody bootstrap pro zjišťování statistické významnosti. Hlavní nevýhoda spočívá v nižší přesnosti odhadu statistické významnosti, která je způsobena tím, že se metoda bootstrap soustředí především na odhad parametrů distribuce spíš než na porovnání rozdílu mezi dvěma systémy. Jako náhrada za bootstrap je v práci [Riezler 05] navrženo použití metody *aproximativní randomizace* (angl. approximate randomization, také známé jako náhodné nebo Monte Carlo permutační testy). Stejně jako v případě metody bootstrap jsou opět nové vzorky vytvářeny na základě originálních změřených hodnot. V případě automatického překladu postupujeme takto. Původní testovací množinu přeložíme oběma systémy a dostaneme množiny překladů T_0^A a T_0^B, nyní spočteme rozdíl ve skóre pro tyto dva systémy: $D_0 = |S_0^A - S_0^B|$. Dále vytváříme nové množiny překladů a to tak, že pro každou dvojici odpovídajících si překladů (překlady stejné zdrojové věty) z množin T_0^A a T_0^B náhodně s pravděpodobností 0,5 rozhodneme o zařazení každé věty do nové množiny zamíchaných překladů T_m^A nebo T_m^B. Po projití všech dvojic vět opět spočteme rozdíl ve skóre pro nově vytvořené množiny překladů T_m^A a T_m^B: $D_m = |S_m^A - S_m^B|$. Tento postup zopakujeme M-krát a počítáme kolikrát je nový rozdíl větší nebo roven původnímu rozdílu: if $D_m \geq D_0$ then $C = C + 1$. Pravděpodobnost nulové hypotézy (tj., že rozdíl mezi systémy není statisticky významný) je pak dána jako: $P = (C + 1)/(M + 1)$. Je-li tedy P nižší nebo rovno zvolené hladině významnosti, můžeme nulovou hypotézu na dané hladině významnosti zamítnout.

V práci [Riezler 05] je také dále poukázáno na důležitou okolnost, kterou je třeba v případě $k-$násobného párového porovnání statistické významnosti vzít do úvahy. V případě k párových porovnání totiž pravděpodobnost náhodného přiřazení statistické významnosti některému z porovnávaných rozdílů roste exponenciálně s k. Tato pravděpodobnost je dána vzorcem: $P_e \approx 1 - (1 - P_c)^k$, kde P_e je pravděpodobnost alespoň jednoho nesprávného zamítnutí nulové hypotézy při k porovnáních a P_c je pak hladina významnosti pro zamítnutí nebo přijetí nulové hypotézy. Chceme-li např. provést 5 porovnání s pravděpodobností zamítnutí nulové hypotézy 0,05, je pravděpodobnost alespoň jednoho špatného určení statistické významnosti rovna 0,23). Jednou z možností jak tuto okolnost vzít při testování do úvahy, je použít tzv. *Bonferroniho korekci*, kdy jednoduše požadovanou hladinu významnosti vydělíme počtem porovnání a tím dostaneme novou hladinu významnosti, kterou použijeme při samotném testování (v případě výše zmíněného příkladu dostaneme novou hladinu významnosti $P_c = 0,05/5 = 0,01$, které odpovídá chyba $P_e \approx 0,05$).

V následujících experimentech tedy pro určení parametrů distribuční funkce daných hodnot použijeme metodu bootstrap (konkrétně budou pro BLEU a NIST kritérium použity skripty,

které jsou volně k dispozici[5], pro SER, WER, PER a SDO pak vlastní skripty). Výsledkem jsou medián a konfidenční interval, který je zaznamenán v hranatých závorkách za výsledkem, popřípadě jako jeho dolní (levý konfidenční interval) a horní (pravý konfidenční interval) index. Pro zjištění statistické významnosti pak použijeme metodu aproximativní randomizace s adekvátně upravenou hladinou významnosti pomocí Bonferroniho korekce (použití vlastních skriptů). Výhoda obou těchto metod oproti metodě křížové validace je dostatečné (prakticky neomezené) množství hodnot pro statistický výpočet a malá časová náročnost, neboť stačí přeložit jenom jednu testovací sadu.

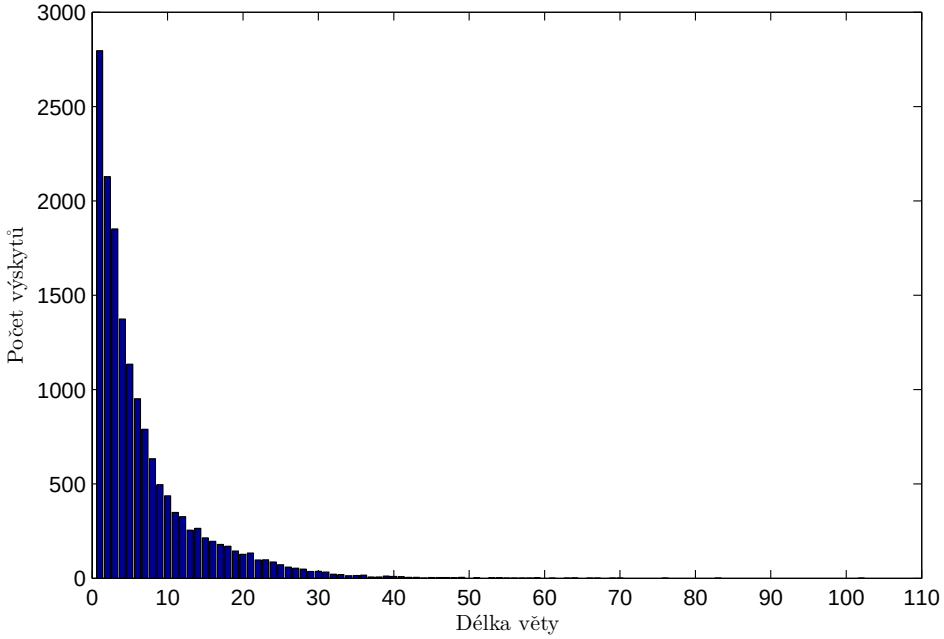
7.2 Dekodéry

Při následujících experimentech jsme využili tyto dva dekodéry:

- MOSES[6] [Koehn 07] – jedná se o volně přístupný (LGPL licence) frázový dekodér, který představuje současný „state of the art“ na poli frázových dekodérů. Vyhledávací algoritmus je založen na výřezové strategii prohledávání a umožňuje tak provádět monotónní i nemonotónní prohledávání. Pro modelování závislosti mezi zdrojovou a cílovou větou je použit log-lineární model, který umožňuje použití libovolných rysů. Pro nastavení vah tohoto modelu je použito MER trénování založené na algoritmu popsaném v práci [Och 03a]. Moses také dále nabízí řadu pokročilých technik pro automatický překlad, jako je MBR dekódování, faktorové překladové modely a dekódování konfuzních sítí. Spolu s dekodérem jsou dostupné i nástroje pro výběr frází, takže Moses představuje statistický systém pro automatický překlad, který dovoluje zcela automaticky vytvořit překladový systém pro libovolný jazykový pár. Vše co je k tomu potřeba, je jen paralelní korpus odpovídající velikosti. Celý systém je také navržen pro práci s velkým množstvím dat a to jak při trénování, tak i při samotném překladu. Klíčové části systému jsou tak implementovány v jazyce C++.

- SiMPaD – představuje vlastní implementaci monotónního frázového dekodéru. Základem dekodéru je algoritmus popsaný na Obrázku 5.1. Monotónnost je obsažena ve vytvářené tabulce *TA*, která umožňuje vytvořit všechna možná monotónní pokrytí vstupní věty, z nichž je při dekódování vybrán nejlepší překlad (při prohledávání je tak nalezeno zároveň nejlepší rozdělení vstupní věty na fráze a jemu odpovídající nejlepší překlad). Pokud bychom tedy chtěli provádět nemonotónní překlad, znamenalo by to přidat do systému možnost vytvářet tabulku *TA*, která by obsahovala i nemonotónní pokrytí (dané např. jedním z přeuspořádávacích omezení v části 5.3) dané vstupní věty, samotný dekódovací algoritmus by pak z praktického hlediska bylo vhodné rozšířit o prořezávání, např. tak jak je uvedeno v části 5.2. Při prohledávání je také využit frázový n-gram uvedený v části 5.1, konkrétně je implementován bigramový a trigramový model. Stejně jako Moses používá i Simpad log-lineární model, jehož váhy jsou opět nastaveny pomocí MER trénování, pro optimalizaci je však použit Nelder-Mead algoritmus zmíněný v části 6 (konkrétně jde o implementaci z knihy [Press 02]). Celý dekodér je implementován v jazyce Python, což umožňuje jeho snadnou údržbu, rozšiřitelnost, přenositelnost a možnost zapojení do různorodých aplikací.

[5] http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm
[6] http://www.statmt.org/moses/



Obrázek 7.3: Počty výskytů vět dané délky.

7.3 Úlohy a korpusy

7.3.1 Čeština – Znakovaná Čeština

V případě překladu mezi češtinou a znakovanou češtinou využijeme CSC korpus, jež je jedním z výsledků této práce. Jak již bylo řečeno, jedná se o korpus, který obsahuje textové záznamy telefonických dialogů mezi uživatelem a operátorem informačního centra vlakových jízdních řádů. Díky tomu je tento korpus poměrně úzce doménově vymezen. Protože CSC korpus vznikl z HHTT korpusu přidáním překladu českých vět do znakované češtiny, obsahuje, kromě paralelních textů, i řadu dalších informací (jde především o anotaci pomocí dialogových aktů), které jsou obsaženy v různých anotačních hladinách a mohou být využity při překladu. V Tabulce 7.1 jsou základní statistiky CSC korpusu. Na Obrázku 7.3 jsou pak znázorněny počty výskytů českých vět dané délky pomocí histogramového grafu. Jak je vidět z grafu, v korpusu převládají spíše kratší věty, především jde o věty kratší než 5 slov, což je způsobené hlavně povahou korpusu (jde o záznam telefonických dialogů). Díky této převaze krátkých vět je tak možné použít pro hodnocení překladů i SER kritérium.

7.4 Výběr frází založený na principu minimální ztráty

V této části prozkoumáme blíže metodu pro výběr frází založenou na principu minimální ztráty popsanou v části 4.4. Nejprve se zaměříme na použité rysy a způsob optimalizace jejich vah. A v dalším experimentu pak vyzkoušíme a porovnáme možná vylepšení této základní metody zmíněná na konci kapitoly 4.4. K experimentům byl použit CSC korpus, který jsme rozdělili na trénovací, vývojovou a testovací část, které jsou blíže popsány v Tabulce 7.2.

V prvním experimentu, jehož výsledky jsou v Tabulce 7.3, jsme nejprve vzali nejlepší rys

	Čeština	Znakovaná čeština
Větné páry	15 772	
Počet slov	107 953	107 663
Velikost slovníku	4 081	2 366
Počet singletonů	1 956	1 113

Tabulka 7.1: Základní statistiky CSC korpusu.

	Trénovací část		Vývojová část		Testovací část	
	Čeština	Znakovaná čeština	Čeština	Znakovaná čeština	Čeština	Znakovaná čeština
Větné páry	12 616		1 578		1 578	
Počet slov	86 690	86 389	10 700	10 722	10 563	10 552
Velikost slovníku	3 670	2 151	1 258	800	1 177	748
Počet singletonů	1 790	1 036	679	373	615	339
OOV(%)	–	–	240 (2,24)	122 (1,14)	208 (1,97)	105 (1,00)

Tabulka 7.2: Rozdělení CSC korpusu na trénovací, vývojovou a testovací část.

z Tabulky 4.1 a postupně jsme k němu přidávaly další rysy tak, jak je lze sestupně seřadit podle úspěšnosti překladu v Tabulce 4.1. Tj. nejprve jsme vzali samotný rys $p_{MI_T}(\tilde{s}, \tilde{t})$, v dalším kroku jsme k němu přidali rys $p_{MI}(\tilde{s}, \tilde{t})$ a postupně pak i další rysy v následujícím pořadí $p_T(\tilde{t}|\tilde{s})$, $p_T(\tilde{s}|\tilde{t})$, $\phi_T(\tilde{t}|\tilde{s})$, $\phi_T(\tilde{s}|\tilde{t})$. Tomu odpovídají jednotlivé řádky v Tabulce 7.3. Sloupce pak odpovídají použití MER trénování pro nalezení optimálních vah rysů při výběru frází a také optimálních vah překladového systému. V obou případech bylo optimalizováno BLEU skóre výsledného překladu a k optimalizaci byl použit Nelder-Mead algoritmus (viz. Kapitola 6). V překladovém systému byly použity tyto rysy: obousměrné překladové pravděpodobnosti ϕ_T, trigramový jazykový model p_{LM} a model slovní a frázové penalizace (celkem tedy 5 rysů). Při hledání překladu pak bylo použito monotónní prohledávání, které je dostatečné pro překládanou dvojici jazyků: čeština - znakovaná čeština. K nalezení překladů byl použit SiMPaD dekodér. Sloupec označený *X_X* znamená, že nebyla použita žádná optimalizace, sloupec označený *Dosim_X* pak značí použití jen optimalizaci vah rysů při výběru frází, sloupec označený *X_Dosim* pak obdobně použití jen optimalizace vah překladového systému a konečně sloupec označený *Dosim_Dosim* označuje použití optimalizace vah rysů při výběru frází a následné optimalizace vah překladového systému, v kterém je již použita optimální frázová tabulka získaná v předešlém kroku.

Z výsledků v Tabulce 7.3 je patrné, že použití optimalizace při výběru frází nebo při překladu popřípadě obou přináší vždy lepší výsledky než v případě, kdy není použita žádná optimalizace. Z hlediska možnosti použití optimalizace je pak nejlepší použít obě optimalizace společně, jak ukazují průměrné výsledky pro jednotlivé způsoby optimalizace. Nejlepšího výsledku pro vývojová data je také dosaženo při použití obou optimalizací, v případě testovacích dat je pak nejlepšího výsledku dosaženo při použití optimalizace jen vah překladového systému

Rys	*X_X*	*Dosim_X*	*X_Dosim*	*Dosim_Dosim*
	Vývojová data			
$p_{MI_T}(\tilde{s},\tilde{t})$	*75,69* [-1,41, 1,34]	76,20 [-1,29, 1,38]	*76,39* [-1,36, 1,35]	**76,91** [-1,36, 1,34]
$+p_{MI}(\tilde{s},\tilde{t})$	**75,81** [-1,40, 1,40]	76,26 [-1,42, 1,32]	**76,66** [-1,36, 1,32]	*76,88* [-1,31, 1,35]
$+p_T(\tilde{t},\tilde{s})$	*75,70* [-1,37, 1,36]	76,22 [-1,38, 1,35]	*76,30* [-1,33, 1,36]	*76,58* [-1,32, 1,31]
$+p_T(\tilde{s},\tilde{t})$	74,88 [-1,35, 1,41]	**76,32** [-1,33, 1,30]	*76,16* [-1,37, 1,32]	*76,79* [-1,32, 1,29]
$+\phi_T(\tilde{t},\tilde{s})$	*75,69* [-1,41, 1,42]	76,24 [-1,42, 1,33]	*76,44* [-1,39, 1,37]	*76,59* [-1,43, 1,34]
$+\phi_T(\tilde{s},\tilde{t})$	75,11 [-1,40, 1,45]	76,02 [-1,36, 1,42]	*76,46* [-1,34, 1,35]	*76,51* [-1,40, 1,36]
Průměr	75,48	76,21	76,40	**76,71**
	Testovací data			
$p_{MI_T}(\tilde{s},\tilde{t})$	*77,94* [-1,38, 1,33]	*78,31* [-1,38, 1,33]	*78,23* [-1,27, 1,27]	*78,80* [-1,34, 1,31]
$+p_{MI}(\tilde{s},\tilde{t})$	78,07 [-1,37, 1,37]	*78,41* [-1,43, 1,38]	78,43 [-1,31, 1,28]	**78,92** [-1,31, 1,31]
$+p_T(\tilde{t},\tilde{s})$	77,59 [-1,40, 1,41]	*78,39* [-1,46, 1,40]	78,55 [-1,38, 1,32]	*78,76* [-1,38, 1,29]
$+p_T(\tilde{s},\tilde{t})$	*77,77* [-1,37, 1,39]	*78,36* [-1,44, 1,38]	78,34 [-1,39, 1,31]	*78,58* [-1,35, 1,33]
$+\phi_T(\tilde{t},\tilde{s})$	**78,15** [-1,35, 1,38]	**78,50** [-1,42, 1,43]	**79,02** [-1,39, 1,29]	*78,89* [-1,40, 1,34]
$+\phi_T(\tilde{s},\tilde{t})$	*77,68* [-1,42, 1,40]	*78,42* [-1,40, 1,37]	*78,84* [-1,38, 1,33]	*78,87* [-1,38, 1,30]
Průměr	77,87	78,40	78,57	**78,80**

Tabulka 7.3: Výsledky překladu při postupném přidávání rysů a různých režimech trénování.

(použití obou optimalizací vede na druhý nejlepší výsledek). Z hlediska jednotlivých rysů je vidět, že při použití optimalizace nejsou až na výjimky rozdíly v přesnosti překladu statisticky významné (indikováno kurzívou, hladina významnosti $p = 0,015$). Zvláště při použití obou optimalizací, kdy nejsou rozdíly mezi výsledky pro vývojová i testovací data ani v jednom případě statisticky významné. Dále tedy budeme v experimentech s frázovou tabulkou získanou pomocí metody výběru frází založeného na principu minimální ztráty používat všech šest rysů a obě optimalizace k tvorbě této tabulky.

V Tabulce 7.4 je vývoj vah jednotlivých rysů pro výběr frází při jejich přidávání. Při optimalizaci těchto vah nebyla použita L_1 norma, takže váhy nemusí sčítat do jedné. V Tabulce 7.5 je pak vývoj samotných vah dekodéru při přidávání jednotlivých rysů pro výběr frází a při různých způsobech trénování. Z tabulky je patrné, že při použití obou optimalizací (*Dosim_Dosim*) dojde ke zvýšení váhy u překladového modelu $\phi_T(\tilde{s},\tilde{t})$ oproti váze při použití optimalizace jen vah dekodéru (*X_Dosim*). Toto zvýšení je na úkor vah pro rys $\phi_T(\tilde{t},\tilde{s})$, u kterého dojde ve všech případech ke snížení váhy. Zbylé váhy zůstávají víceméně stejné.

V dalším experimentu, jehož výsledky jsou v Tabulce 7.6, jsme se zaměřili na porovnání a možnou kombinaci metod pro zlepšení výběru frází založeného na principu minimální ztráty zmíněných na konci Kapitoly 4. K nalezení překladů byl ve všech metodách použit SiMPaD dekodér s pěti standardními rysy a monotónním dekódováním (viz předchozí experiment). V prvním sloupci je označení jednotlivých metod, ve druhém a třetím sloupci jsou pak hodnoty získané pro vývojová a testovací data a konečně v posledním je počet překladových dvojic ve

Dosim_X						
Rys	p_{MI_T}	p_{MI}	$p_T(\tilde{s},\tilde{t})$	$p_T(\tilde{t},\tilde{s})$	$\phi_T(\tilde{t},\tilde{s})$	$\phi_T(\tilde{s},\tilde{t})$
$p_{MI_T}(\tilde{s},\tilde{t})$	2,75	–	–	–	–	–
$+p_{MI}(\tilde{s},\tilde{t})$	2,73	0,23	–	–	–	–
$+p_T(\tilde{s},\tilde{t})$	2,24	1,80	-0,16	–	–	–
$+p_T(\tilde{t},\tilde{s})$	1,56	1,58	0,20	1,57	–	–
$+\phi_T(\tilde{t},\tilde{s})$	2,05	0,07	0,23	2,23	0,43	–
$+\phi_T(\tilde{s},\tilde{t})$	1,53	0,86	0,12	1,80	1,95	0,25

Tabulka 7.4: Vývoj vah rysů pro výběr frází při jejich postupném přidávání a optimalizaci na vývojových datech.

X_Dosim					
Rys	$\phi_T(\tilde{s},\tilde{t})$	$\phi_T(\tilde{t},\tilde{s})$	*PhP*	*LM*	*WP*
$p_{MI_T}(\tilde{s},\tilde{t})$	0,17	0,33	0,29	0,14	-0,07
$+p_{MI}(\tilde{s},\tilde{t})$	0,16	0,32	0,32	0,16	-0,03
$+p_T(\tilde{s},\tilde{t})$	0,09	0,47	0,18	0,12	-0,14
$+p_T(\tilde{t},\tilde{s})$	0,11	0,40	0,23	0,15	-0,11
$+\phi_T(\tilde{t},\tilde{s})$	0,07	0,49	0,29	0,15	-0,01
$+\phi_T(\tilde{s},\tilde{t})$	0,09	0,45	0,21	0,12	-0,13
Dosim_Dosim					
$p_{MI_T}(\tilde{s},\tilde{t})$	0,17	0,27	0,34	0,14	-0,08
$+p_{MI}(\tilde{s},\tilde{t})$	0,17	0,31	0,31	0,16	-0,06
$+p_T(\tilde{s},\tilde{t})$	0,22	0,36	0,27	0,14	-0,01
$+p_T(\tilde{t},\tilde{s})$	0,21	0,26	0,28	0,14	-0,11
$+\phi_T(\tilde{t},\tilde{s})$	0,20	0,28	0,28	0,15	-0,08
$+\phi_T(\tilde{s},\tilde{t})$	0,20	0,24	0,30	0,16	-0,10

Tabulka 7.5: Vývoj vah překladového systému při postupném přidávání rysů pro výběr frází a optimalizaci těchto vah na vývojových datech pro různé režimy trénování.

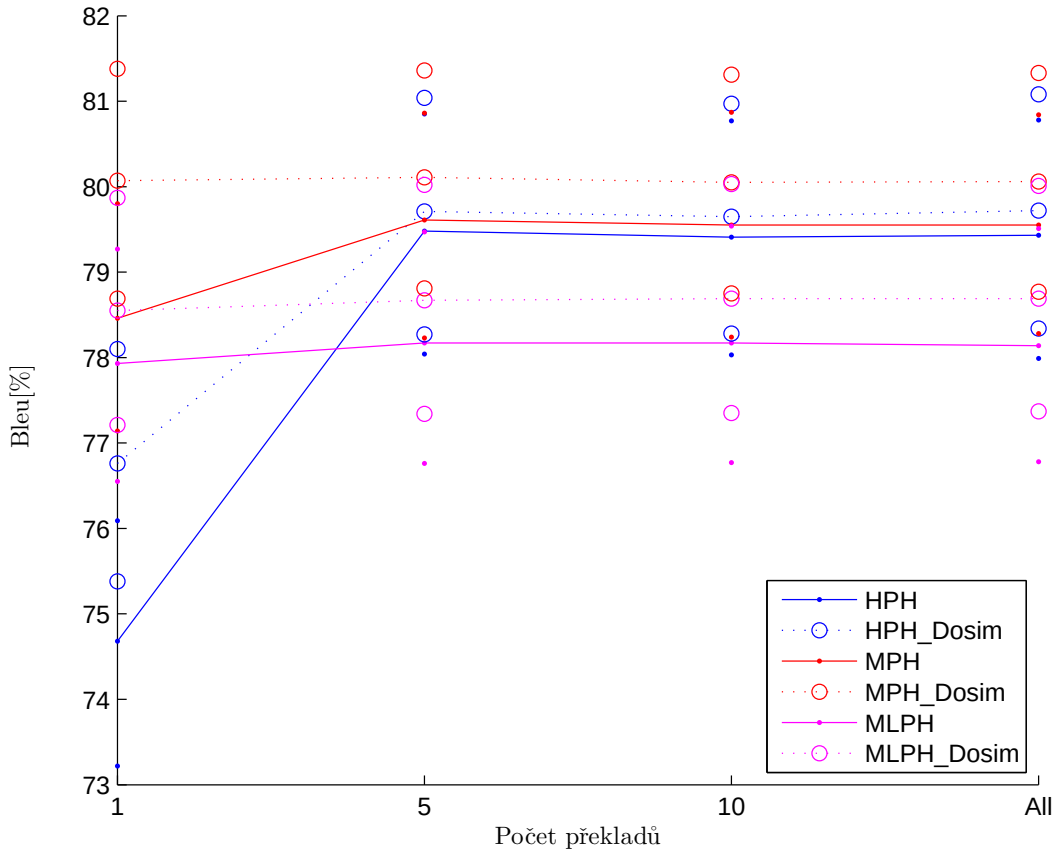
výsledné frázové tabulce. V prvním řádku (*Dosim_Dosim*) tabulky je základní hodnota, která byla získána při použití obou optimalizací a všech šesti rysů při výběru frází (s těmi pracují i další použité metody). Ve druhém řádku (*Once_All*) jsou výsledky při použití metody, kdy v případě, že se daná fráze vyskytuje jen jednou, jsou vybrány všechny možné překlady bez ohledu na velikost jejich překladové ztráty $L_T(\tilde{s},\tilde{t})$ (na rozdíl od standardního postupu, kdy jsou vybrány jen překlady s nejnižší překladovou ztrátou). Vzhledem k tomu, že fráze délky větší než jedna jsou uvažovány jen v případě, že se vyskytnou více než pětkrát, ovlivní tato metoda jenom výběr překladů samotných slov. Při trénování byla použita optimalizace vah rysů

	Vývojová data	Testovací data	Velikost tabulky
Dosim_Dosim	76,51 [-1,40, 1,36]	78,87 [-1,38, 1,30]	28 012
Once_All	76,32 [-1,41, 1,35]	78,55 [-1,40, 1,37]	102 193
Intersected_IDD	**77,35** [-1,41, 1,36]	**79,38** [-1,39, 1,36]	28 428
Intersected_DID	77,06 [-1,36, 1,34]	78,90 [-1,39, 1,36]	32 520
Filtered_XX	77,84 [-1,35, 1,35]	79,69 [-1,39, 1,30]	11 692
Filtered_DX	77,58 [-1,39, 1,35]	79,62 [-1,36, 1,35]	12 332
Filtered_XD	**78,17** [-1,32, 1,30]	**79,98** [-1,32, 1,38]	11 692
Filtered_DD	77,95 [-1,32, 1,33]	79,66 [-1,37, 1,34]	12 332
Intersected_ID_Filtered_D	**78,69** [-1,37, 1,33]	**80,22** [-1,34, 1,30]	11 585
Intersected_DI_Filtered_D	78,20 [-1,32, 1,32]	79,84 [-1,39, 1,26]	11 648
Filtered_Intersected_D	77,92 [-1,36, 1,34]	79,78 [-1,35, 1,34]	9 323

Tabulka 7.6: Porovnání výsledků různých metod pro zlepšení metody výběru frází založeného na principu minimální ztráty.

pro výběr frází i vah dekodéru. Z porovnání výsledků je zřejmé, že použití metody *Once_All* vede ke zhoršení výsledku na vývojových i testovacích datech. Při použití metody *Once_All* dojde také ke značnému nárůstu velikosti frázové tabulky (v tomto případě se tabulka zvětší pětinásobně). Další použité metody se pak především snaží o redukci velikosti této tabulky a další zlepšení přesnosti překladu. Protože tyto metody pracují na základě výběru vhodné podmnožiny překladových dvojic z již vytvořené frázové tabulky, je lepší použít jako jejich vstup tabulku vytvořenou právě pomocí metody *Once_All*, protože obsahuje více možností, z kterých lze následně vybírat.

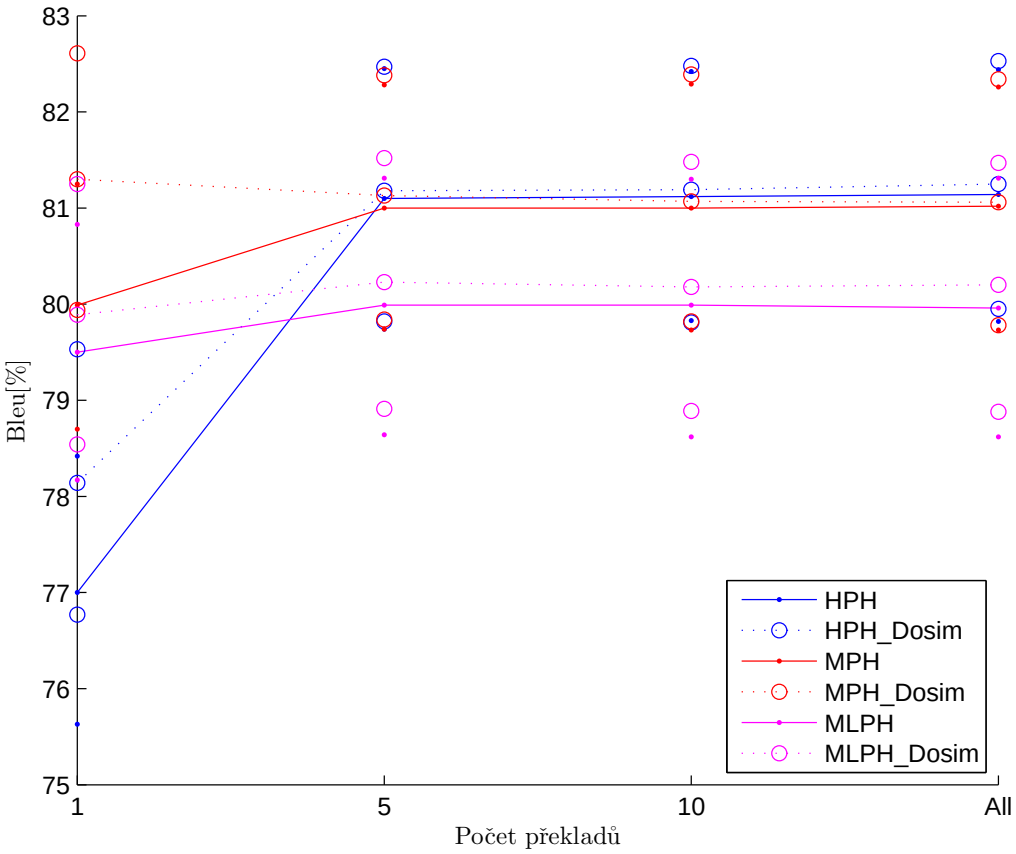
První metodou je metoda průniku (*Intersected*), kdy novou frázovou tabulku vytvoříme jako průnik frázových tabulek získaných pro každý směr překladu zvlášť. Do nové tabulky ukládáme ty překladové páry, které tvoří vzájemný překlad, tj. jestliže $\tilde{t}$ je překladem $\tilde{s}$ (frázová tabulka: zdrojový jazyk –> cílový jazyk), pak zároveň musí platit, že $\tilde{s}$ je překladem $\tilde{t}$ (frázová tabulka: cílový jazyk –> zdrojový jazyk) a do nové tabulky uložíme překladovou dvojici: ($\tilde{s}$, $\tilde{t}$) spolu s informací, kolikrát jsme tuto dvojici viděli. Takto projdeme všechny zdrojové fráze v původní tabulce, jestliže však daná zdrojová fráze nemá žádný vzájemný překlad, pak ji do nové tabulky uložíme s původními překlady. Na základě zjištěných hodnot o počtech odpovídajících si překladů pak spočteme nové překladové pravděpodobnosti ϕ_T. Máme v zásadě dvě možnosti, jak průnik tabulek použít při trénování. První možností je provést nejprve průnik jednotlivých tabulek, poté použít optimalizaci vah rysů pro výběr frází a nakonec provést optimalizaci vah dekodéru pro výslednou tabulku (v tabulce označeno jako *Intersected_IDD*). Při optimalizaci vah rysů pro výběr frází je vždy nejprve vytvořena na základě stávajících vah frázová tabulka pro každý směr zvlášť. Pak je proveden průnik těchto dvou tabulek a vzniklá tabulka je následně použita pro výpočet nových vah rysů. To opakujeme dokud není nalezena optimální tabulka, ta je pak použita pro výpočet optimálních vah dekodéru. Druhou možností je pak nejprve nalézt optimální frázovou tabulku pro každý směr (optimalizace vah rysů pro výběr frází), provést průnik těchto optimálních tabulek a nakonec opět optimalizovat váhy



Obrázek 7.4: Porovnání úspěšnosti překladu na vývojových datech pro různé frázové tabulky při různém počtu uvažovaných překladů.

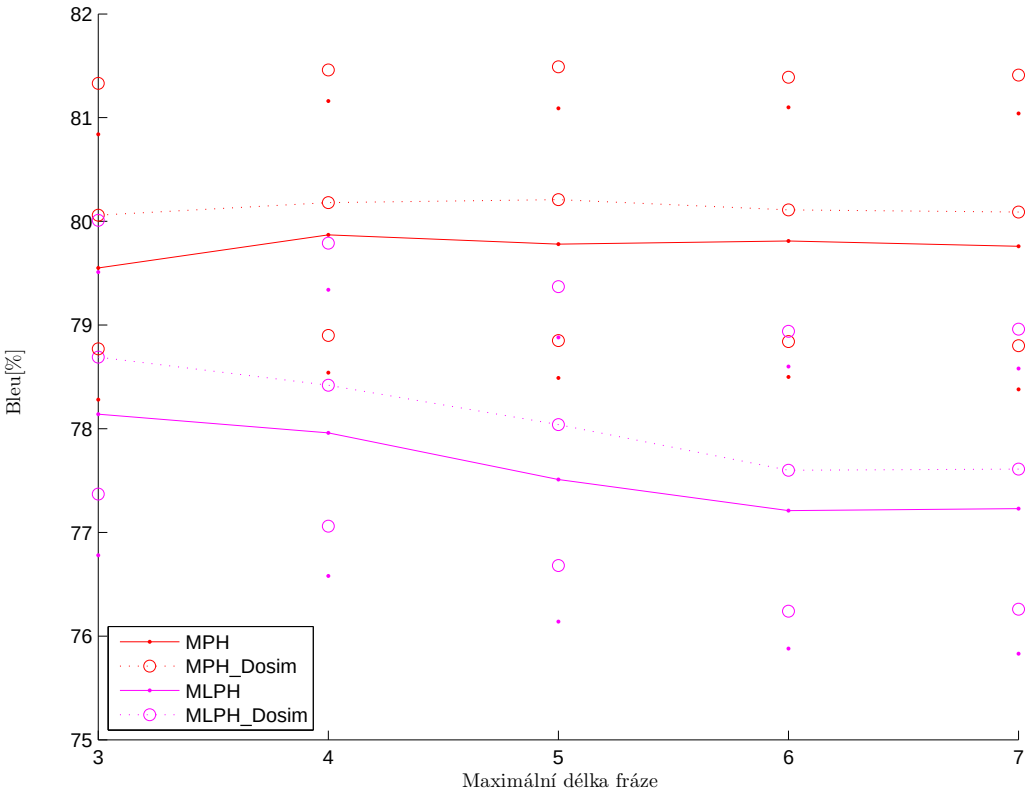
dekodéru pro novou tabulku (v tabulce označeno jako *Intersected_DID*). Jako vhodnější se z hlediska vývojových i testovacích dat jeví první postup, který vede k významnému zlepšení úspěšnosti překladu.

Druhou možností je použít metodu filtrování frázové tabulky (*Filtered*). V tomto případě použijeme původní frázovou tabulku k překladu nějakého zvoleného textu (lze použít např. trénovací data) a zaznamenáme si, které překlady a kolikrát byly použity. Na základě těchto hodnot pak vypočteme nové hodnoty překladových pravděpodobností ϕ_T a vytvoříme novou frázovou tabulku, která pro každou zdrojovou frázi obsahuje jenom ty překlady, které byly použity při překladu zvoleného textu. Zdrojové fráze, které nebyly použity při překladu, jsou do nové tabulky přidány s původními překlady a překladovými pravděpodobnostmi. V tabulce jsou výsledky pro čtyři různé možnosti postupu při filtrování tabulky. Nejprve jsou při filtrování i dekódování použity neoptimalizované váhy dekodéru (*Filtered_XX*), v dalším řádku je výsledek při použití optimalizovaných vah jen při překladu zvoleného textu (*Filtered_DX*), dále je výsledek při použití jen optimalizace vah dekodéru na novou tabulku (*Filtered_XD*) a konečně je výsledek při použití optimalizovaných vah při překladu zvoleného textu a následné optimalizace vah dekodéru na novou tabulku (*Filtered_DD*). Nejlepších výsledků dosáhneme při použití třetího postupu (*Filtered_XD*), kdy je nová tabulka vybrána pomocí základních vah dekodéru a následně jsou váhy dekodéru optimalizovány pro tuto novou tabulku.



Obrázek 7.5: Porovnání úspěšnosti překladu na testovacích datech pro různé frázové tabulky při různém počtu uvažovaných překladů.

Poslední tři řádky tabulky obsahují výsledky pro kombinaci dvou předchozích (*Intersected* a *Filtered*) metod a následnou optimalizaci vah dekodéru pro výslednou tabulku. Tyto dvě metody můžeme zkombinovat dvěma způsoby. V prvních dvou řádcích jsou výsledky postupu, kdy nejdříve provedeme průnik tabulek a poté filtraci výsledné tabulky (*Intersected_Filtered*). Protože průnik tabulek lze provést opět dvěma způsoby (*Intersected_ID* a *Intersected_DI*, viz výše), byly při této kombinaci vyzkoušeny obě varianty (*Intersected_ID_Filtered_D* a *Intersected_DI_Filtered_D*), tj. jako vstup pro filtraci byly použity obě možné tabulky, které lze získat použitím metody průniku. Druhou možnou kombinací je pak nejprve provést filtraci tabulky pro každý směr překladu zvlášť a poté aplikovat metodu průniku na takto získané tabulky (*Filtered_Intersected_D*). Ve všech případech použití metody filtrace byla použita nejlepší varianta (*Filtered_XD*) z předchozího porovnání. Z vyzkoušených možností možné kombinace metod vychází jako nejlepší použití filtrace na tabulku získanou provedením nejprve průniku tabulek pro jednotlivé směry překladu a následnou optimalizací vah pro výběr frází, tj. *Intersected_ID_Filtered_D*.



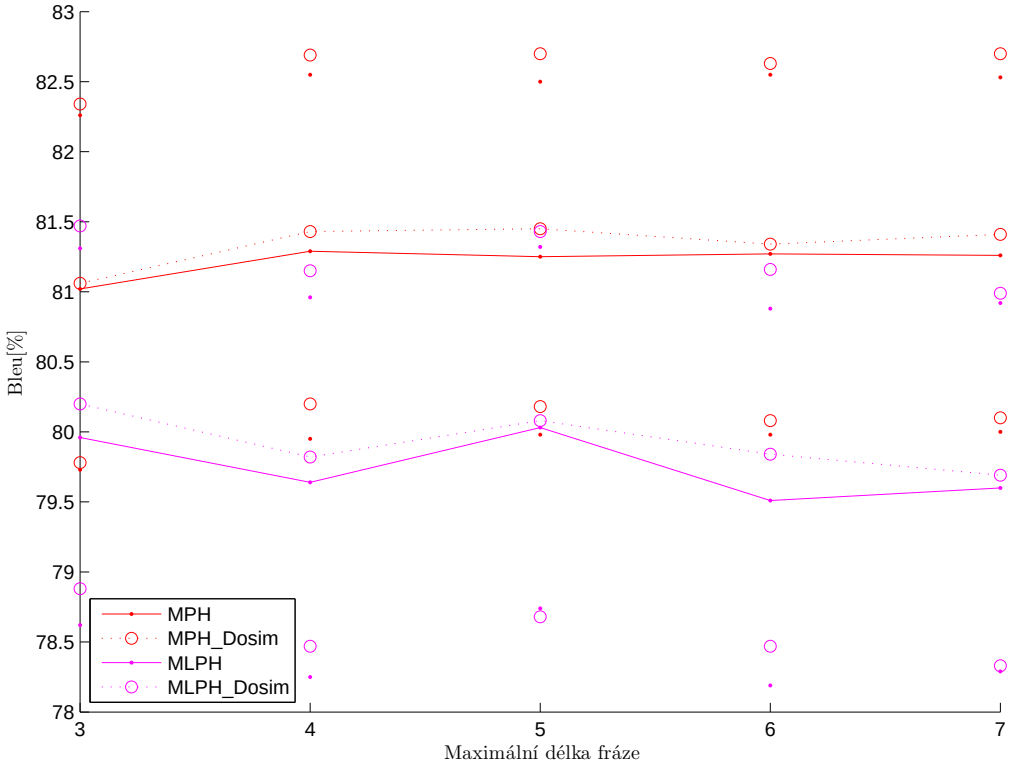
Obrázek 7.6: Porovnání úspěšnosti překladu na vývojových datech pro různé frázové tabulky při různé maximální délce frází.

7.5 Porovnání různých frázových tabulek

V rámci následujících experimentů jsme mezi sebou porovnali tři frázové tabulky, které byly získány různými postupy. První frázová tabulka (HPH) byla vytvořena na základě ručního přiřazení mezi zdrojovou frází a jejím překladem, které bylo vytvořeno při tvorbě CSC korpusu. Jedná se tedy o ručně vytvořenou frázovou tabulku. Další dvě tabulky pak byly vytvořeny pomocí různých automatických metod pro výběr frází.

První tabulka (MPH) byla vytvořena pomocí nástrojů, které jsou k dispozici u dekodéru MOSES. Při konstrukci frázové tabulky byla v tomto případě použita standardní metoda pro výběr frází, která je popsána v Kapitole 4.1.1. Z daných trénovacích dat jsou nejprve vytvořena slovní přiřazení pro oba směry překladu. Ta jsou následně spojena do jednoho symetrického přiřazení. Do frázové tabulky jsou pak vybrány ty spojité fráze do dané délky, které jsou konzistentní s vytvořeným symetrickým přiřazením. Tato metoda nabízí různé modifikace, takže při výběru bylo použito výchozí nastavení nástrojů.

Druhou tabulkou je pak frázová tabulka vytvořená pomocí výběru frází založeného na principu minimální ztráty (MLPH), tak jak byla popsána v předcházející části. Jako nejlepší způsob pro konstrukci tabulky vyšel postup označený *Intersected_ID_Filtered_D* v předchozí Tabulce 7.6, kdy je použita metoda optimalizace vah pro výběr frází na frázové tabulce, která je vytvořena průnikem tabulek pro jednotlivé směry překladu. Na takto získanou optimální

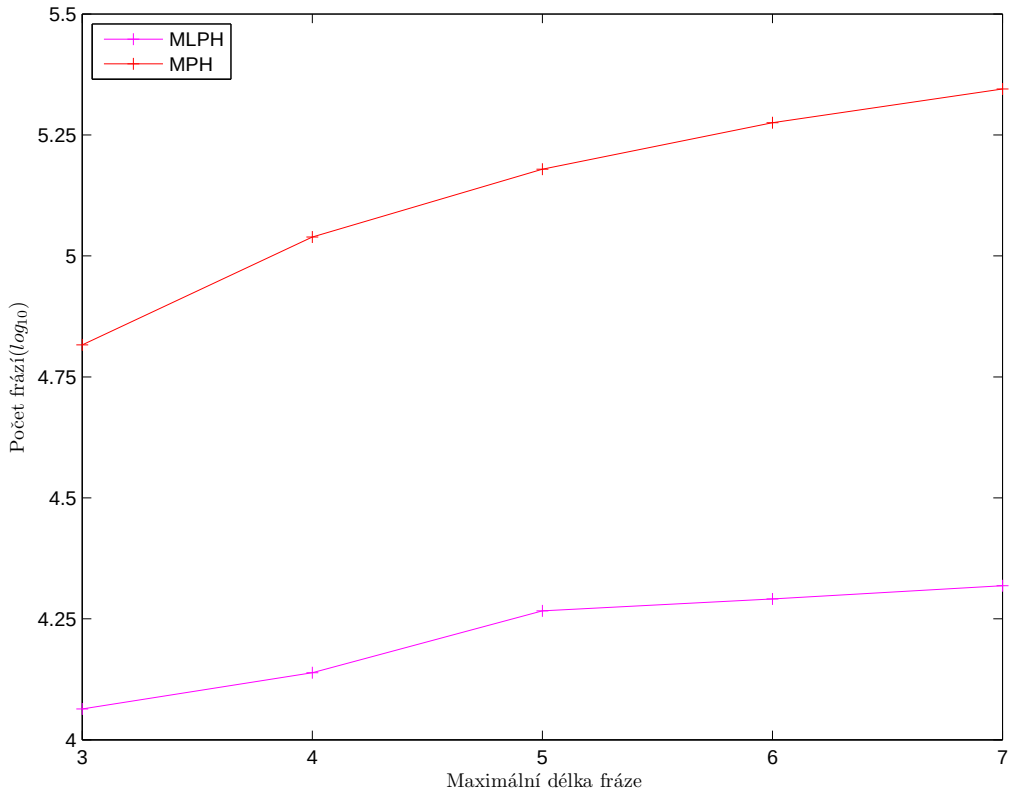


Obrázek 7.7: Porovnání úspěšnosti překladu na testovacích datech pro různé frázové tabulky při různé maximální délce frází.

tabulku je pak aplikována metoda filtrace, kdy jsou při překladu zvoleného textu zjištěny používané překlady a nepoužité překlady jsou z tabulky odstraněny.

Na Obrázcích 7.4 a 7.5 je porovnání výsledků BLEU kritéria pro všechny tři tabulky při použití různého maximálního počtu překladů pro každou zdrojovou frázi. Při překladu byl použit SiMPaD dekodér se standardními pěti rysy (viz výše). Na Obrázku 7.4 je průběh hodnoty BLEU kritéria na vývojových datech. Plnou čarou je vyznačen průběh pro neoptimalizované váhy, tečkovanou pak pro optimalizované váhy. Pomocí bodů odpovídající barvy a označení jsou označeny konfidenční intervaly pro každou hodnotu. Byly vždy vyzkoušeny čtyři různé hodnoty maximálního počtu překladů uvažovaných pro každou zdrojovou frázi při překladu. Nejprve byla použita jen jedna nejlepší fráze, poté pět, deset a nakonec všechny možné překlady. Na Obrázku 7.5 jsou pak obdobně výsledky pro testovací data. Z průběhů je vidět, že od použití pěti nejlepších překladů jsou mezi výsledky minimální rozdíly.

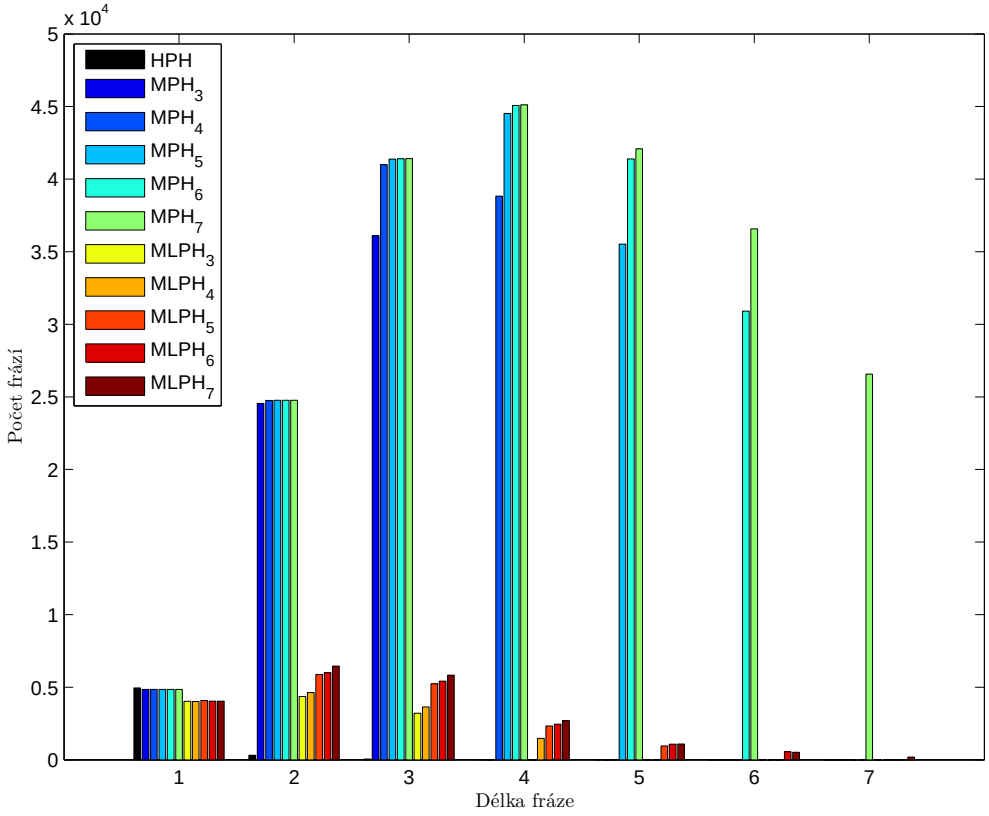
Při tvorbě obou tabulek získaných automaticky, lze definovat maximální délku vybíraných frází. Na Obrázcích 7.6 a 7.7 jsou tak výsledky BLEU kritéria na vývojových a testovacích datech při použití různé maximální délky vybraných frází. V grafech je použito stejné značení jako u předchozích porovnání z hlediska maximálního počtu uvažovaných frází. Jako počáteční hodnota byla zvolena maximální délka fráze tři. V případě MPH tabulky je z výsledků zřejmé, že od maximální délky čtyři jsou rozdíly mezi jednotlivými výsledky minimální (resp. dochází spíše k mírnému zhoršení). V případě MLPH je pak zřejmý pokles přesnosti překladu pro fráze delší než tři. Na Obrázku 7.8 je pak průběh nárůstu velikosti frázové tabulky při zvětšování



Obrázek 7.8: Vývoj velikosti frázové tabulky při různé maximální délce uvažovaných frází pro obě automaticky získané frázové tabulky.

maximální délky frází. V obou případech můžeme pozorovat lineární nárůst velikosti tabulky, který je v případě MPH (z 65 000 na 220 000) tabulky strmější než v případě MLPH tabulky (z 11 000 na 21 000), což vede až k řádovému rozdílu ve velikosti tabulky pro maximální délku frází sedm.

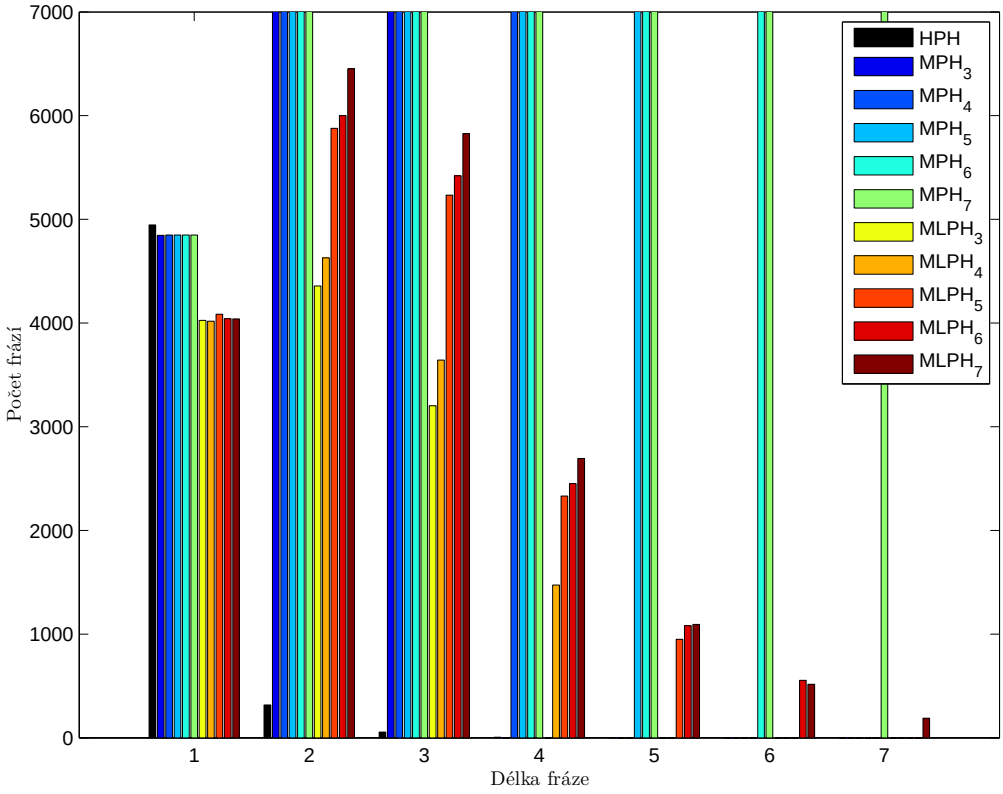
Dané tři tabulky lze dále porovnat z hlediska celkové velikosti a také zastoupení frází dané délky. Toto porovnání je na Obrázcích 7.9 a 7.10, které obsahují histogramy zastoupení frází dané délky pro všechny tři typy tabulek. V případě automaticky získaných tabulek je zobrazeno zastoupení frází pro tabulky s maximální délkou frází od tří do sedmi (maximální délka frází je v popisu označena pomocí dolního indexu u příslušného označení tabulky). Z porovnání je vidět, že ručně vytvořená HPH tabulka obsahuje v drtivé míře fráze délky jedna, tedy samotná slova, a dále pak v mnohem menší míře fráze délky dva a tři. Delší fráze se pak již téměř nevyskytují (resp. jsou zastoupeny již maximálně v jednotkových počtech). V případě automaticky získaných tabulek je tomu pak spíše naopak, kdy je zastoupení delších frází vyšší než zastoupení frází délky jedna. Především to platí pro tabulku MPH, u které zastoupení frází delších než jedna několikanásobně překonává zastoupení frází délky jedna ve všech případech, kdy byly fráze dané délky vybírány při tvorbě tabulky. U tabulky MLPH to pak platí především pro fráze délky dva a tři při maximální délce frází větší nebo rovné pěti. Pokud jde o porovnání zastoupení frází dané délky vzhledem k celkovému počtu frází v tabulce, pak HPH tabulka tíhne k výběru jen jednotlivých slov, doplněných o několik málo často frekventovaných delších



Obrázek 7.9: Porovnání zastoupení frází dané délky pro různé frázové tabulky.

frází. Naopak MPH tabulka se soustředí na výběr delších frází, především pak frází délky tři, čtyři a pět, jejichž počty významně překonávají počty frází kratších. V případě MLPH tabulky jsou pak počty frází do délky tři celkem vyrovnané, přičemž u delších frází pak prudce klesají na rozdíl od MPH tabulky. Tento rozdíl je daný především povahou metod použitých při výběru těchto dvou tabulek. Při výběru MPH tabulky se vychází ze slovního přiřazení na jehož základě se vybírají všechny fráze konzistentní s tímto přiřazením, takže výsledkem je tabulka, která obsahuje velké množství relevantních, ale překrývajících se frází. Výběr frází do MLPH tabulky je pak ovlivněn tím, že fráze délky větší než jedna jsou uvažovány jen v případě, že se vyskytnou minimálně pětkrát a dále také tím, že vybíráme jen fráze společné pro oba směry překladu (*Intersected*), které ještě dále filtrujeme na fráze skutečně používané při překladu (*Filtered*). Z hlediska porovnání celkové velikosti různých typů tabulky je nejmenší HPH tabulka, která obsahuje 5 325 frází (z toho je 4 946 délky jedna). Druhá je nejmenší MLPH tabulka s maximální délkou fráze tři, která obsahuje 11 585 frází. A poslední pak nejmenší MPH tabulka (opět maximální délka fráze tři), která obsahuje 65 494 frází.

Dále je porovnání přesnosti překladu všech tří tabulek z hlediska různých kritérií. Obě automaticky vytvořené tabulky byly použity ve verzi s maximální délkou fráze rovnou tři, při překladu pak byly vždy uvažovány všechny dostupné překlady. K překladu byly použity oba dekodéry MOSES (M) i SiMPaD (S) s neoptimalizovanými i optimalizovanými vahami (přípona M a D v tabulce, ve všech případech byla použita optimalizace z hlediska BLEU kritéria). V případě HPH a MLPH tabulek bylo použito pět standardních rysů: obousměrné překladové



Obrázek 7.10: Porovnání zastoupení frází dané délky pro různé frázové tabulky - bližší pohled.

pravděpodobnosti, jazykový model, frázová a slovní penalizace. V případě MPH tabulky jsou ještě navíc použity obousměrné lexikální překladové pravděpodobnosti [Koehn 03]:

$$p_w(\tilde{s}|\tilde{t}, a) = \prod_{i=1}^{n} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall(i,j)\in a} w(s_i|t_j) \tag{7.12}$$

$$a = \{(i,j)\}, i = 1, ..., n, j = 0, 1, ..., m$$

$$w(s|t) = \frac{N(s,t)}{\sum_{s'} N(s',t)},$$

kde $N(s,t)$ udává počet, kolikrát jsme viděli slovo s společně s překladem t a $\sum_{s'} N(s',t)$ pak kolikrát jsme viděli samotný překlad t. Cílem lexikální překladové pravděpodobnosti je ověřit kvalitu frázového páru prostřednictvím toho, jak dobrými překlady jsou mezi sebou jednotlivá slova. K tomu je použita lexikální překladová pravděpodobnost $w(s|t)$.

V Tabulkách 7.7 až 7.9 jsou postupně výsledky přesnosti překladu pro všechna uvažovaná kritéria. V případě chybových kritérií (SER, WER, PER) platí, že nižší je lepší, v případě kritérií přesnosti (BLEU, NIST, SDO) pak vyšší je lepší. V Tabulce 7.8 jsou výsledky SDO kritéria pro porovnání mezi sémantickým stromem referenčních překladů a sémantickým stromem automaticky vytvořených překladů. V Tabulce 7.9 jsou pak výsledky SDO kritéria pro

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
	BLEU[%]					
M	$79,33^{+1,38}_{-1,40}$	$79,54^{+1,29}_{-1,40}$	$78,16^{+1,39}_{-1,43}$	$81,06^{+1,31}_{-1,35}$	$80,87^{+1,31}_{-1,31}$	$79,84^{+1,39}_{-1,34}$
M_M	$79,91^{+1,26}_{-1,34}$	$80,27^{+1,28}_{-1,33}$	$78,82^{+1,35}_{-1,36}$	$81,29^{+1,27}_{-1,29}$	$81,05^{+1,29}_{-1,30}$	$80,21^{+1,32}_{-1,36}$
S	$79,40^{+1,40}_{-1,32}$	$79,59^{+1,27}_{-1,29}$	$78,16^{+1,35}_{-1,40}$	$81,14^{+1,33}_{-1,33}$	$81,01^{+1,23}_{-1,27}$	$79,97^{+1,33}_{-1,34}$
S_D	$79,72^{+1,34}_{-1,37}$	$80,07^{+1,28}_{-1,30}$	$78,69^{+1,27}_{-1,39}$	$81,22^{+1,31}_{-1,31}$	$81,08^{+1,27}_{-1,32}$	$80,20^{+1,28}_{-1,33}$
	NIST					
M	$11,55^{+0,14}_{-0,14}$	$11,44^{+0,14}_{-0,15}$	$11,33^{+0,15}_{-0,15}$	$11,67^{+0,13}_{-0,13}$	$11,57^{+0,14}_{-0,14}$	$11,49^{+0,14}_{-0,14}$
M_M	$11,54^{+0,14}_{-0,14}$	$11,44^{+0,15}_{-0,14}$	$11,29^{+0,15}_{-0,16}$	$11,65^{+0,13}_{-0,14}$	$11,57^{+0,13}_{-0,14}$	$11,47^{+0,14}_{-0,14}$
S	$11,67^{+0,13}_{-0,13}$	$11,57^{+0,14}_{-0,14}$	$11,49^{+0,14}_{-0,14}$	$11,63^{+0,14}_{-0,14}$	$11,58^{+0,14}_{-0,13}$	$11,44^{+0,15}_{-0,15}$
S_D	$11,63^{+0,14}_{-0,14}$	$11,58^{+0,14}_{-0,13}$	$11,44^{+0,15}_{-0,15}$	$11,65^{+0,13}_{-0,13}$	$11,58^{+0,14}_{-0,14}$	$11,44^{+0,14}_{-0,14}$
	SER[%]					
M	$41,95^{+3,42}_{-3,36}$	$41,83^{+3,42}_{-3,30}$	$43,16^{+3,42}_{-3,49}$	$38,53^{+3,49}_{-3,42}$	$39,16^{+3,42}_{-3,42}$	$40,56^{+3,55}_{-3,42}$
M_M	$41,70^{+3,42}_{-3,49}$	$40,37^{+3,42}_{-3,36}$	$42,65^{+3,55}_{-3,49}$	$38,15^{+3,49}_{-3,30}$	$38,21^{+3,42}_{-3,42}$	$40,56^{+3,49}_{-3,36}$
S	$41,76^{+3,55}_{-3,42}$	$42,02^{+3,36}_{-3,55}$	$44,61^{+3,55}_{-3,49}$	$38,59^{+3,42}_{-3,30}$	$39,35^{+3,42}_{-3,36}$	$42,59^{+3,61}_{-3,49}$
S_D	$41,70^{+3,36}_{-3,49}$	$41,13^{+3,49}_{-3,30}$	$44,11^{+3,68}_{-3,30}$	$38,53^{+3,42}_{-3,30}$	$38,59^{+3,49}_{-3,36}$	$42,90^{+3,55}_{-3,36}$
	WER[%]					
M	$13,83^{+1,27}_{-1,21}$	$14,16^{+1,32}_{-1,22}$	$15,56^{+1,43}_{-1,41}$	$13,12^{+1,31}_{-1,25}$	$13,49^{+1,31}_{-1,31}$	$14,67^{+1,41}_{-1,34}$
M_M	$13,72^{+1,29}_{-1,23}$	$13,63^{+1,26}_{-1,23}$	$15,20^{+1,43}_{-1,35}$	$13,14^{+1,33}_{-1,29}$	$13,43^{+1,36}_{-1,31}$	$14,48^{+1,41}_{-1,35}$
S	$13,81^{+1,38}_{-1,25}$	$14,17^{+1,30}_{-1,27}$	$15,86^{+1,39}_{-1,34}$	$13,13^{+1,33}_{-1,30}$	$13,49^{+1,26}_{-1,25}$	$15,02^{+1,34}_{-1,39}$
S_D	$13,67^{+1,33}_{-1,25}$	$13,85^{+1,31}_{-1,25}$	$15,47^{+1,42}_{-1,34}$	$13,06^{+1,32}_{-1,25}$	$13,43^{+1,31}_{-1,25}$	$14,88^{+1,42}_{-1,33}$
	PER[%]					
M	$12,34^{+1,18}_{-1,13}$	$12,85^{+1,19}_{-1,14}$	$13,75^{+1,30}_{-1,23}$	$11,55^{+1,15}_{-1,17}$	$12,00^{+1,18}_{-1,17}$	$12,82^{+1,25}_{-1,19}$
M_M	$12,38^{+1,18}_{-1,12}$	$12,17^{+1,22}_{-1,12}$	$13,55^{+1,31}_{-1,24}$	$11,64^{+1,22}_{-1,17}$	$11,85^{+1,21}_{-1,16}$	$12,95^{+1,21}_{-1,18}$
S	$12,39^{+1,18}_{-1,13}$	$12,83^{+1,19}_{-1,12}$	$14,07^{+1,32}_{-1,21}$	$11,63^{+1,19}_{-1,16}$	$11,97^{+1,22}_{-1,15}$	$13,16^{+1,25}_{-1,18}$
S_D	$12,33^{+1,22}_{-1,14}$	$12,41^{+1,19}_{-1,18}$	$13,81^{+1,32}_{-1,24}$	$11,72^{+1,20}_{-1,13}$	$11,93^{+1,20}_{-1,13}$	$13,24^{+1,26}_{-1,16}$

Tabulka 7.7: Porovnání výsledků pro různé frázové tabulky a oba dekodéry.

porovnání mezi referenčním sémantickým stromem získaným z CSC korpusu a sémantickým stromem automaticky vytvořených překladů.

Z porovnání výsledků pro jednotlivé tabulky vyplývá, že HPH a MPH tabulky dosahují srovnatelných výsledků (výsledky HPH tabulky jsou ve většině případů lepší, ale rozdíl je

	SDO[%]					
	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
M	$90,69^{+2,25}_{-2,63}$	$90,15^{+2,28}_{-2,75}$	$89,72^{+2,28}_{-2,53}$	$92,25^{+1,96}_{-2,37}$	$92,12^{+1,97}_{-2,33}$	$90,07^{+2,26}_{-2,54}$
M_M	$90,12^{+2,30}_{-2,80}$	$90,81^{+2,17}_{-2,49}$	$89,75^{+2,31}_{-2,51}$	$92,08^{+1,95}_{-2,37}$	$92,12^{+2,03}_{-2,30}$	$90,84^{+2,07}_{-2,49}$
S	$90,67^{+2,22}_{-2,59}$	$90,13^{+2,33}_{-2,66}$	$89,57^{+2,27}_{-2,63}$	$92,24^{+1,94}_{-2,38}$	$92,12^{+1,97}_{-2,35}$	$90,53^{+2,23}_{-2,53}$
S_D	$90,33^{+2,30}_{-2,56}$	$90,69^{+2,12}_{-2,54}$	$89,90^{+2,30}_{-2,59}$	$92,25^{+1,96}_{-2,30}$	$92,11^{+2,01}_{-2,39}$	$90,82^{+2,13}_{-2,51}$

Tabulka 7.8: Porovnání výsledků SDO kritéria pro různé frázové tabulky a oba dekodéry.

	SDO[%]					
	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
M	$55,94^{+3,17}_{-3,17}$	$55,59^{+3,05}_{-3,22}$	$55,44^{+2,96}_{-3,14}$	$55,80^{+3,02}_{-3,11}$	$55,76^{+3,21}_{-3,16}$	$55,39^{+3,08}_{-3,09}$
M_M	$55,77^{+3,06}_{-3,26}$	$56,11^{+3,01}_{-3,13}$	$55,47^{+2,97}_{-3,21}$	$55,77^{+3,00}_{-3,27}$	$55,86^{+3,19}_{-3,08}$	$55,31^{+3,04}_{-3,08}$
S	$55,90^{+2,97}_{-3,23}$	$55,74^{+3,04}_{-3,21}$	$55,26^{+2,87}_{-3,16}$	$55,84^{+3,11}_{-3,16}$	$55,77^{+3,08}_{-3,20}$	$55,32^{+3,07}_{-3,19}$
S_D	$55,95^{+3,05}_{-3,06}$	$56,04^{+2,94}_{-3,13}$	$55,34^{+3,03}_{-3,15}$	$55,52^{+2,96}_{-3,25}$	$56,11^{+3,07}_{-3,12}$	$55,37^{+3,11}_{-3,11}$

Tabulka 7.9: Porovnání výsledků SDO kritéria pro různé frázové tabulky a oba dekodéry (referenční sémantická data).

zanedbatelný). MLPH tabulka pak zaostává v závislosti na kritériu o jedno až více než dvě procenta, což není z praktického hlediska nijak velká ztráta. Výhodou HPH a MLPH tabulek oproti MLPH tabulce je pak jejich mnohem menší velikost. Z grafů na Obrázcích 7.6 a 7.7 je vidět, že při použití maximální délky frází čtyři můžeme v případě MPH tabulky dostat ještě o něco lepší výsledky než v případě použité délky tři (např. BLEU na testovacích datech z 81,08 na 81,43) toto zlepšení je však vykoupeno takřka zdvojnásobením velikosti výsledné tabulky (ze 65 000 na 110 000). Jinou možností je použití metody filtrace také na HPH a MPH tabulku. Jak je vidět z výsledků v Tabulce 7.10 (byl použit SiMPaD dekodér) vede tento postup ke zlepšení přesnosti překladu pro obě tabulky a v případě HPH tabulky je ještě navíc dosaženo redukce velikosti tabulky takřka o 20 procent (v případě MPH tabulky jde jen o minimální 4 procentní redukci).

7.6 Srovnání dekodérů

Srovnání výsledků obou dekodérů pro všechny tři použité frázové tabulky lze nalézt v předešlé části v Tabulkách 7.7 až 7.9. Ze srovnání vyplývá, že výsledky obou dekodérů jsou plně srovnatelné. Největší rozdíl je mezi přesností překladu pro MLPH tabulku, kdy např. v případě SER kritéria poskytuje MOSES dekodér významně lepší výsledky. V případě BLEU kritéria, z hlediska kterého byly váhy dekodéru optimalizovány, jsou však rozdíly mezi výsledky obou dekodérů minimální.

	Vývojová data	Testovací data	Velikost tabulky
HPH	79,71 [-1,44, 1,33]	81,18 [-1,36, 1,29]	5 325
HPH_Filtered	**79,96** [-1,38, 1,28]	**81,51** [-1,32, 1,31]	4 373 (82%)
MPH	80,11 [-1,30, 1,25]	81,13 [-1,29, 1,25]	65 494
MPH_Filtered	**80,28** [-1,35, 1,27]	**81,50** [-1,29, 1,27]	62 833 (96%)

Tabulka 7.10: Výsledky použití metody filtrace frázové tabulky pro různé frázové tabulky (BLEU kritérium).

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
SiMPaD_Bi_2	79,33	79,35	77,70	81,12	80,60	79,36
SiMPaD_Bi_3	79,41	79,58	78,13	81,15	81,01	79,99
SiMPaD_Bi_4	79,41	79,59	78,15	81,14	81,02	80,00
SiMPaD_Tri_2	79,28	79,40	77,73	81,12	80,58	79,29
SiMPaD_Tri_3	79,34	79,52	78,20	81,03	80,88	79,86
SiMPaD_Tri_4	79,33	79,55	78,18	81,07	80,88	79,85

Tabulka 7.11: Porovnání různých řádů frázového n-gramu u SiMPaD dekodéru a různých řádů n-gramu jazykového modelu (BLEU kritérium).

V případě SiMPaD dekodéru, který využívá při dekódování frázový n-gram (viz Kapitola 5.1), lze mezi sebou porovnat použití různých řádů tohoto n-gramu při překladu. Konkrétně jsme mezi sebou porovnali použití frázového bigramu (v tabulce označeno příponou Bi) a trigramu (označeno příponou Tri) a zároveň jsme také porovnali kombinaci s různými řády jazykového modelu (příslušný řád modelu je vyznačen odpovídajícím dolním indexem). Výsledky jsou v Tabulce 7.11. Z těchto výsledků lze odvodit, že pro danou úlohu překladu je vhodné použít bigramový frázový model pro SiMPaD dekodér spolu s trigramovým jazykovým modelem. Bigramový frázový model totiž poskytuje stejné nebo dokonce mírně lepší výsledky než trigramový model a zároveň nabízí větší rychlost prohledávání danou tím, že se při prohledávání prochází méně hypotéz. V případě trigramového jazykového modelu je pak volba dána tím, že trigramový model poskytuje lepší výsledky než bigramový a další zvyšování řádu n-gramu již nepřináší významné zlepšení přesnosti překladu.

V případě MOSESU máme možnost porovnat monotónní (v tabulce označeno příponou Mon) a nemonotónní (označeno příponou Nonmon) způsob překladu. Jak již bylo řečeno při překladu z češtiny do znakované češtiny bychom měli vystačit s monotónním překladem. Z výsledků v Tabulce 7.12 je zřejmé, že monotónní i nemonotónní překlad poskytují stejné výsledky, takže předpoklad o dostatečnosti monotónního způsobu překladu byl správný. Pro monotónní překlad pak hovoří i mnohem větší rychlost překladu.

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
Moses_Mon	79,34	79,53	78,18	81,07	80,89	79,82
Moses_Nonmon	77,58	77,86	76,61	79,61	79,48	78,43
Moses_Mon_M	79,91	80,28	78,82	81,32	81,05	80,21
Moses_Nonmon_M	79,95	80,36	78,87	81,10	81,06	80,18

Tabulka 7.12: Porovnání monotónního a nemonotónního překladu při použití dekodéru MOSES (BLEU kritérium).

7.7 Vylepšení základního systému pro překlad

V předchozích experimentech byl představen základní systém pro překlad mezi češtinou a znakovanou češtinou. Díky dalším informacím obsaženým v CSC korpusu a povaze překládané dvojice jazyků lze přesnost překladu ještě dále vylepšit.

První možností je využití třídního jazykového modelu založeného na pojmenovaných entitách, které jsou obsaženy v CSC korpusu. Podobně jako v případě ASR tak i v oblasti automatického překladu je řada chyb způsobena slovy chybějícími ve slovníku (angl. out-of-vocabulary words (OOV)). V případě překladu můžeme totiž přeložit jen ta slova (fráze), která máme v překladovém slovníku, tj. známe jejich překlad. Z analýzy vytvořených překladů vyplynulo, že řada OOV slov je způsobena chybějícími jmény stanic a osob. Protože řešená úloha překladu se omezuje na doménu dialogů z vlakového informačního centra, lze problém OOV slov řešit podobně jako v práci [Hoidekr 06]. Kde byl třídní jazykový model využit pro titulkování hokejových zápasů probíhající v reálném čase. Do standardního jazykového modelu byly přidány třídy pro jména hráčů, národností a států. Podobně jsme i my přidali do našeho jazykového modelu třídy pro všechna známa jména stanic: STATION a osob: PERSON a dále také pro oblast: AREA a typ vlaku: TRAIN. Protože sémantická anotace CSC korpusu obsahuje označení těchto pojmenovaných entit, můžeme je v datech jednoduše nahradit odpovídající třídou a vytvořit z nich příslušné překladové slovníky. Použití třídního jazykového modelu vede ke zlepšení v řádu jednoho až dvou a půl procenta přesnosti překladu v závislosti na daném kritériu (viz porovnání výsledků v Tabulkách 7.7 a 7.8 versus výsledky v Tabulce 7.13). To je způsobeno především redukcí OOV slov na nulu a také poklesem perplexity jazykového modelu o 20,06 procenta (resp. o 17,86 procenta v případě vývojových dat). Jedinou výjimkou je NIST kritérium, v jehož případě došlo k poklesu jeho hodnoty. Rozdíly v přesnosti překladu mezi jednotlivými tabulkami zůstaly zachovány na stejné úrovni (pokud porovnáme hodnotu BLEU kritéria s ohledem, na nějž byla provedena optimalizace).

Dále je možné použít pozpracování získaných překladů, které upraví získaný překlad do konečné podoby. Za prvé můžeme z výsledného překladu vynechat slova, která se nepřekládají (resp. byla přeložena použitím znaku „_“ pro prázdný překlad). V trénovacích datech je ovšem dobré tato slova ponechat, neboť pak dostaneme lepší výsledky překladu díky detailnějším a tím jednoznačnějším překladovým a jazykovým modelům. Za druhé můžeme neznámá slova nahradit znakem pro hláskování (spelling), neboť slova, pro něž ve znakované řeči neexistuje znak, se hláskují pomocí prstové abecedy. Použitím pozpracování na výsledky překladu získaného s využitím třídního jazykového modelu dostaneme jednak konečnou podobu překladu (odstranění prázdných překladů), kterou lze prezentovat jako výstup překladového systému, a dále má použití pozpracování také pozitivní vliv na přesnost vytvořeného překladu (to platí

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
# frází	4 203	56 549	11 281	4 203	56 549	11 281
	BLEU[%]					
S	$80,26^{+1,36}_{-1,37}$	$80,20^{+1,33}_{-1,34}$	$79,03^{+1,31}_{-1,40}$	$82,08^{+1,33}_{-1,26}$	$82,10^{+1,18}_{-1,25}$	$81,42^{+1,29}_{-1,30}$
S_D	$80,62^{+1,36}_{-1,32}$	$80,75^{+1,36}_{-1,39}$	$79,58^{+1,36}_{-1,33}$	$82,40^{+1,23}_{-1,30}$	$82,33^{+1,24}_{-1,25}$	$81,21^{+1,31}_{-1,31}$
	NIST					
S	$11,24^{+0,14}_{-0,14}$	$11,15^{+0,15}_{-0,15}$	$11,03^{+0,14}_{-0,15}$	$11,34^{+0,13}_{-0,13}$	$11,28^{+0,13}_{-0,14}$	$11,23^{+0,14}_{-0,15}$
S_D	$11,23^{+0,15}_{-0,14}$	$11,23^{+0,15}_{-0,14}$	$11,08^{+0,15}_{-0,14}$	$11,33^{+0,13}_{-0,14}$	$11,34^{+0,13}_{-0,14}$	$11,20^{+0,14}_{-0,14}$
	SER[%]					
S	$39,73^{+3,36}_{-3,36}$	$40,30^{+3,55}_{-3,42}$	$42,71^{+3,42}_{-3,36}$	$36,25^{+3,36}_{-3,36}$	$37,14^{+3,42}_{-3,30}$	$39,42^{+3,42}_{-3,49}$
S_D	$40,05^{+3,49}_{-3,42}$	$39,29^{+3,42}_{-3,30}$	$42,08^{+3,49}_{-3,36}$	$36,31^{+3,23}_{-3,42}$	$36,44^{+3,30}_{-3,36}$	$40,43^{+3,55}_{-3,49}$
	WER[%]					
S	$12,90^{+1,28}_{-1,24}$	$13,20^{+1,27}_{-1,23}$	$14,68^{+1,42}_{-1,34}$	$12,05^{+1,31}_{-1,24}$	$12,25^{+1,21}_{-1,17}$	$13,32^{+1,30}_{-1,25}$
S_D	$12,80^{+1,25}_{-1,24}$	$12,83^{+1,27}_{-1,21}$	$14,34^{+1,40}_{-1,28}$	$11,94^{+1,30}_{-1,21}$	$12,06^{+1,26}_{-1,19}$	$13,39^{+1,31}_{-1,31}$
	PER[%]					
S	$11,47^{+1,15}_{-1,16}$	$11,86^{+1,17}_{-1,15}$	$12,87^{+1,26}_{-1,19}$	$10,56^{+1,17}_{-1,13}$	$10,84^{+1,13}_{-1,07}$	$11,61^{+1,19}_{-1,13}$
S_D	$11,46^{+1,16}_{-1,13}$	$11,37^{+1,21}_{-1,09}$	$12,68^{+1,23}_{-1,17}$	$10,60^{+1,18}_{-1,12}$	$10,57^{+1,19}_{-1,13}$	$11,80^{+1,24}_{-1,16}$
	SDO[%]					
S	$90,54^{+2,35}_{-2,64}$	$91,30^{+2,16}_{-2,50}$	$90,45^{+2,21}_{-2,57}$	$91,99^{+2,10}_{-2,44}$	$92,04^{+2,04}_{-2,46}$	$91,78^{+2,04}_{-2,29}$
S_D	$90,80^{+2,20}_{-2,54}$	$91,30^{+2,19}_{-2,55}$	$90,48^{+2,18}_{-2,65}$	$92,32^{+1,95}_{-2,43}$	$92,86^{+1,88}_{-2,29}$	$91,41^{+2,09}_{-2,39}$

Tabulka 7.13: Výsledky použití třídního jazykového modelu při překladu z češtiny do znakované češtiny.

především pro optimalizované BLEU kritérium, viz výsledky v Tabulce 7.14). To je opět, jako v případě použití třídního jazykového modelu, způsobeno poklesem perplexity jazykového modelu o dalších 17,49 procenta (resp. o 18,33 procenta v případě vývojových dat).

7.8 Překlad ze znakované češtiny do češtiny

Stejné metody a postupy, které jsme použili v případě překladu z češtiny do znakované češtiny můžeme použít i pro opačný směr překladu, tedy ze znakované češtiny do češtiny. Pokrytí obou směrů překladu je výhodné z hlediska použití překladového systému pro zajištění obousměrného dialogu mezi slyšícím a neslyšícím. Aby mohla být stávající data z CSC korpusu použita pro natrénování systému pro opačný směr překladu, bylo třeba je ještě dodatečně upravit. Tato úprava spočívá ve vypuštění prázdných překladů (znak „_“) a jim odpovídajícím českým slovům. Zatímco totiž v předešlém směru překladu byla vstupem česká věta, která

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
	BLEU[%]					
S	$81,42^{+1,45}_{-1,49}$	$81,03^{+1,45}_{-1,45}$	$80,30^{+1,33}_{-1,41}$	$83,20^{+1,32}_{-1,39}$	$82,95^{+1,32}_{-1,34}$	$82,61^{+1,36}_{-1,38}$
S_D	$81,46^{+1,42}_{-1,48}$	$81,62^{+1,46}_{-1,46}$	$80,73^{+1,36}_{-1,47}$	$83,21^{+1,35}_{-1,38}$	$83,38^{+1,29}_{-1,37}$	$82,42^{+1,33}_{-1,39}$
	NIST					
S	$11,15^{+0,14}_{-0,15}$	$11,17^{+0,15}_{-0,15}$	$11,02^{+0,14}_{-0,14}$	$11,22^{+0,14}_{-0,14}$	$11,25^{+0,13}_{-0,14}$	$11,14^{+0,14}_{-0,14}$
S_D	$11,18^{+0,15}_{-0,15}$	$11,20^{+0,15}_{-0,14}$	$11,05^{+0,14}_{-0,15}$	$11,24^{+0,14}_{-0,14}$	$11,27^{+0,14}_{-0,14}$	$11,13^{+0,14}_{-0,14}$
	SER[%]					
S	$39,04^{+3,36}_{-3,42}$	$39,48^{+3,55}_{-3,36}$	$41,32^{+3,42}_{-3,42}$	$35,68^{+3,42}_{-3,42}$	$36,06^{+3,42}_{-3,42}$	$38,47^{+3,49}_{-3,36}$
S_D	$39,48^{+3,42}_{-3,49}$	$38,34^{+3,30}_{-3,30}$	$41,00^{+3,55}_{-3,30}$	$35,49^{+3,49}_{-3,36}$	$35,30^{+3,36}_{-3,23}$	$39,54^{+3,49}_{-3,36}$
	WER[%]					
S	$13,07^{+1,30}_{-1,24}$	$13,19^{+1,34}_{-1,28}$	$14,49^{+1,40}_{-1,36}$	$12,17^{+1,35}_{-1,25}$	$12,11^{+1,30}_{-1,22}$	$13,26^{+1,38}_{-1,33}$
S_D	$12,95^{+1,32}_{-1,23}$	$12,92^{+1,35}_{-1,24}$	$14,22^{+1,42}_{-1,28}$	$12,11^{+1,27}_{-1,28}$	$11,92^{+1,29}_{-1,32}$	$13,28^{+1,38}_{-1,33}$
	PER[%]					
S	$10,91^{+1,14}_{-1,13}$	$11,17^{+1,14}_{-1,11}$	$12,12^{+1,22}_{-1,15}$	$10,21^{+1,20}_{-1,08}$	$10,34^{+1,17}_{-1,08}$	$11,20^{+1,22}_{-1,18}$
S_D	$10,87^{+1,16}_{-1,10}$	$10,74^{+1,16}_{-1,09}$	$11,98^{+1,26}_{-1,21}$	$10,19^{+1,17}_{-1,05}$	$9,99^{+1,18}_{-1,05}$	$11,39^{+1,17}_{-1,16}$
	SDO[%]					
S	$90,78^{+2,20}_{-2,58}$	$90,68^{+2,19}_{-2,57}$	$89,55^{+2,34}_{-2,67}$	$92,01^{+2,11}_{-2,45}$	$92,88^{+1,94}_{-2,48}$	$91,44^{+2,24}_{-2,59}$
S_D	$90,65^{+2,27}_{-2,59}$	$91,27^{+2,13}_{-2,55}$	$89,38^{+2,36}_{-2,66}$	$92,23^{+2,08}_{-2,49}$	$92,78^{+1,97}_{-2,42}$	$90,90^{+2,19}_{-2,70}$

Tabulka 7.14: Výsledky překladu z češtiny do znakované češtiny po použití pozpracování a třídního jazykového modelu.

obsahovala všechna zapsaná slova, nyní je vstupem věta ve znakované češtině, která kdyby byla produkována samotným neslyšícím, by žádné prázdné překlady neobsahovala. A jí tedy také tím pádem odpovídá česká věta, která vznikne vynecháním českých slov, jež se do znakované češtiny nepřekládají. V případě obousměrného překladu mezi češtinou a znakovanou češtinou tak neplatí předpoklad o vzájemné převoditelnosti mezi odpovídajícími si překlady, který se v případě mluvených jazyků považuje za všeobecně platný. Tj. jestliže máme dvojjazyčný korpus odpovídajících si překladů, předpokládáme v případě mluvených jazyků, že překladový systém pro každý směr překladu lze získat prostým prohozením zdrojového a cílového jazyka.

Podle očekávání jsou výsledky pro překlad ze znakované češtiny do češtiny horší než v opačném případě směru překladu. Rozdíl činí asi 20 procentních bodů v případě BLEU kritéria, s ohledem na nějž byla opět přesnost překladu optimalizována (viz výsledky v Tabulce 7.15). Tento rozdíl je způsoben především tím, že v případě překladu ze znakované češtiny do češtiny je přítomna větší nejednoznačnost, neboť překládáme z menšího slovníku na větší, než v opačném případě. Z hlediska jednotlivých tabulek je situace obdobná jako u předešlého směru překladu. HPH a MPH tabulky opět dosahují srovnatelných výsledků, zatímco MLPH tabulka

	Vývojová data			Testovací data		
	HPH	MPH	MLPH	HPH	MPH	MLPH
# frází	3 621	45 711	9 217	3 621	45 711	9 217
	BLEU[%]					
S	$60,32^{+1,95}_{-1,96}$	$60,33^{+1,98}_{-1,97}$	$59,02^{+2,04}_{-2,07}$	$63,38^{+1,98}_{-2,06}$	$63,40^{+2,03}_{-2,01}$	$61,28^{+1,97}_{-2,02}$
S_D	$61,25^{+2,00}_{-2,01}$	$61,89^{+2,00}_{-2,00}$	$59,54^{+1,96}_{-1,96}$	$63,80^{+2,06}_{-2,01}$	$63,98^{+1,97}_{-1,98}$	$61,69^{+2,01}_{-2,07}$
	NIST					
S	$9,33^{+0,19}_{-0,18}$	$9,35^{+0,19}_{-0,18}$	$9,26^{+0,19}_{-0,18}$	$9,58^{+0,19}_{-0,20}$	$9,61^{+0,19}_{-0,19}$	$9,43^{+0,19}_{-0,20}$
S_D	$9,36^{+0,19}_{-0,19}$	$9,49^{+0,19}_{-0,18}$	$9,22^{+0,19}_{-0,19}$	$9,57^{+0,20}_{-0,19}$	$9,67^{+0,19}_{-0,19}$	$9,42^{+0,20}_{-0,20}$
	SER[%]					
S	$51,17^{+3,46}_{-3,52}$	$51,04^{+3,65}_{-3,46}$	$51,76^{+3,59}_{-3,52}$	$51,09^{+3,59}_{-3,47}$	$51,28^{+3,53}_{-3,53}$	$53,08^{+3,40}_{-3,40}$
S_D	$51,11^{+3,52}_{-3,65}$	$49,15^{+3,46}_{-3,52}$	$51,63^{+3,59}_{-3,52}$	$50,80^{+3,43}_{-3,43}$	$50,19^{+3,53}_{-3,59}$	$53,02^{+3,40}_{-3,53}$
	WER[%]					
S	$21,34^{+1,65}_{-1,59}$	$21,42^{+1,74}_{-1,62}$	$22,70^{+1,82}_{-1,78}$	$20,14^{+1,64}_{-1,57}$	$20,08^{+1,61}_{-1,58}$	$21,74^{+1,66}_{-1,62}$
S_D	$21,05^{+1,70}_{-1,63}$	$20,52^{+1,76}_{-1,58}$	$22,75^{+1,81}_{-1,82}$	$20,11^{+1,63}_{-1,56}$	$19,70^{+1,59}_{-1,54}$	$21,55^{+1,74}_{-1,64}$
	PER[%]					
S	$20,34^{+1,59}_{-1,57}$	$20,48^{+1,61}_{-1,60}$	$21,18^{+1,65}_{-1,57}$	$18,79^{+1,58}_{-1,43}$	$19,09^{+1,52}_{-1,52}$	$20,19^{+1,58}_{-1,59}$
S_D	$20,07^{+1,60}_{-1,57}$	$19,47^{+1,57}_{-1,49}$	$21,33^{+1,81}_{-1,72}$	$18,83^{+1,65}_{-1,50}$	$18,66^{+1,53}_{-1,50}$	$20,15^{+1,63}_{-1,59}$
	SDO[%]					
S	$84,66^{+2,38}_{-2,56}$	$85,26^{+2,23}_{-2,52}$	$83,89^{+2,49}_{-2,58}$	$84,13^{+2,57}_{-2,81}$	$84,46^{+2,44}_{-2,75}$	$83,69^{+2,50}_{-2,85}$
S_D	$84,79^{+2,44}_{-2,58}$	$85,49^{+2,42}_{-2,54}$	$83,72^{+2,48}_{-2,79}$	$84,64^{+2,49}_{-2,82}$	$84,69^{+2,37}_{-2,64}$	$83,89^{+2,50}_{-2,86}$

Tabulka 7.15: Výsledky překladu ze znakované češtiny do češtiny.

za nimi zaostává, přičemž její ztráta oproti předešlému směru překladu ještě narostla, např. v případě BLEU kritéria je ztráta vyšší přibližně o dalších 1,3 procenta (Tabulka 7.14 versus Tabulka 7.15). Výhoda menší velikosti HPH a MLPH tabulek oproti MPH tabulce zůstala zachována.

7.9 Použití slovního přiřazení pro výběr frází založený na principu minimální ztráty

Dosud jsme v metodě pro výběr frází založený na principu minimální ztráty využívali jen rysy založené na různých četnostech výskytu překladových dvojic a jednotlivých frází. Takto vytvořená frázová tabulka však zaostává v přesnosti překladu za tabulkou získanou standardní metodou výběru frází založenou na určeném slovním přiřazení. Jednou z možností jak zlepšit přesnost překladu tabulky založené na principu minimální ztráty, je použití dalších rysů. Jak je ukázáno v práci [Deng 08], vhodnými rysy pro výběr frází jsou rysy založené na aposteriorní

	Vývojová data			Testovací data		
	-	CLM	CLM_PP	-	CLM	CLM_PP
	Simpad					
HPH	$79,40^{+1,40}_{-1,32}$	$\mathbf{80,26}^{+1,36}_{-1,37}$	$\mathbf{81,42}^{+1,45}_{-1,49}$	$\mathbf{81,14}^{+1,33}_{-1,33}$	$82,08^{+1,33}_{-1,26}$	$\mathbf{83,20}^{+1,32}_{-1,39}$
MPH	$\mathbf{79,59}^{+1,27}_{-1,29}$	$80,20^{+1,33}_{-1,34}$	$81,03^{+1,45}_{-1,45}$	$81,05^{+1,29}_{-1,30}$	$\mathbf{82,10}^{+1,18}_{-1,25}$	$82,95^{+1,32}_{-1,34}$
MLPH	$78,16^{+1,35}_{-1,40}$	$79,03^{+1,31}_{-1,40}$	$\mathit{80,30}^{+1,33}_{-1,41}$	$79,97^{+1,33}_{-1,34}$	$\mathit{81,42}^{+1,29}_{-1,30}$	$\mathit{82,61}^{+1,36}_{-1,38}$
MLPH$_{WA}$	$\mathit{79,00}^{+1,36}_{-1,34}$	$\mathit{79,86}^{+1,34}_{-1,43}$	$\mathit{80,51}^{+1,37}_{-1,45}$	$\mathit{80,78}^{+1,30}_{-1,38}$	$81,85^{+1,32}_{-1,31}$	$\mathit{82,81}^{+1,32}_{-1,35}$
	Simpad_Dosim					
HPH	$\mathit{79,72}^{+1,34}_{-1,37}$	$\mathit{80,62}^{+1,36}_{-1,32}$	$\mathit{81,46}^{+1,42}_{-1,48}$	$\mathbf{81,22}^{+1,31}_{-1,31}$	$\mathbf{82,40}^{+1,23}_{-1,30}$	$\mathit{83,21}^{+1,35}_{-1,38}$
MPH	$\mathbf{80,07}^{+1,28}_{-1,30}$	$\mathbf{80,75}^{+1,36}_{-1,39}$	$\mathbf{81,62}^{+1,46}_{-1,46}$	$81,08^{+1,27}_{-1,32}$	$\mathit{82,33}^{+1,24}_{-1,25}$	$\mathbf{83,38}^{+1,29}_{-1,37}$
MLPH	$78,69^{+1,27}_{-1,39}$	$79,58^{+1,36}_{-1,33}$	$80,73^{+1,36}_{-1,47}$	$80,20^{+1,28}_{-1,33}$	$81,21^{+1,31}_{-1,31}$	$82,42^{+1,33}_{-1,39}$
MLPH$_{WA}$	$\mathit{79,46}^{+1,33}_{-1,34}$	$\mathit{80,21}^{+1,29}_{-1,29}$	$\mathit{80,94}^{+1,33}_{-1,39}$	$80,83^{+1,32}_{-1,33}$	$\mathit{81,91}^{+1,30}_{-1,28}$	$\mathit{82,97}^{+1,28}_{-1,39}$

Tabulka 7.16: Porovnání výsledků pro různé frázové tabulky a oba dekodéry.

pravděpodobnosti získané ze slovního přiřazení. V následujícím experimentu tedy nastíníme vliv použití takovýchto rysů založených na využití pravděpodobnostního modelu slovního přiřazení na výběr frází založený na principu minimální ztráty.

Ke stávajícím šesti rysům pro výběr frází: p_{MI_T}, p_{MI}, $p_T(\tilde{t}|\tilde{s})$, $p_T(\tilde{s}|\tilde{t})$, $\phi_T(\tilde{t}|\tilde{s})$, $\phi_T(\tilde{s}|\tilde{t})$ jsme přidali nový rys p_{WA} založený na modelu slovního přiřazení:

$$p_{WA}(\tilde{s}|\tilde{t}) = \frac{n * \sum_{i=1}^{m} \sum_{j=1}^{n} w(s_i|t_j)}{m}, \tag{7.13}$$

kde $w(s_i|t_j)$ je lexikální překladová pravděpodobnost dvojice slov s_i, t_j (tj. pravděpodobnost, že s_i se přeloží jako t_j) získaná na základě vytvořeného slovního přiřazení a $\tilde{s} = s_1, ..., s_m$, $\tilde{t} = t_1, ..., t_n$. Tento nový rys se tak pomocí lexikálních pravděpodobností snaží určit odpovídající si překladové páry. Je tak obdobou lexikální překladové pravděpodobnosti p_w používané v případě MPH frázové tabulky (pro výpočet je použita stejná slovní překladová pravděpodobnost w, takže pro fráze délky jedna jsou hodnoty obou rysů p_w a p_{WA} shodné). Na rozdíl od výpočtu pravděpodobnosti p_w však při výpočtu p_{WA} není využito slovního přiřazení a mezi jednotlivými slovy frází v překladovém páru. Tento rys je pak tedy z tohoto pohledu jednodušší než rys p_w a jeho cílem je tak spíše jen nastínit možnosti použití rysů založených na slovním přiřazení pro výběr frází, než aby představoval plné využití všech dostupných informací ze slovního přiřazení (rysy zabývající se širším využitím informací ze slovního přiřazení lze nalézt v již zmíněné práci [Deng 08]). Výsledky frázové tabulky získané s využitím rysu p_{WA} jsou v Tabulce 7.16 vždy v řádku označeném popiskem MLPH$_{WA}$. Tato tabulka dále obsahuje hodnoty BLEU kritéria pro všechny tři zkoumané frázové tabulky (HPH, MPH a MLPH) a pro různé úpravy vstupních a výstupních dat překladového systému (CLM, CLM_PP). Rozdíly mezi hodnotami byly porovnány z hlediska statistické významnosti na hladině významnosti $p = 0,02$. Nejvyšší hodnota je vždy vyznačena tučně, statisticky nevýznamný rozdíl pak kurzívou. Z výsledků je vidět, že použití rysu p_{WA} vedlo ve všech případech ke zlepšení přesnosti překladu oproti původní MLPH tabulce. Navíc došlo v případě MLPH$_{WA}$ tabulky, s výjimkou jednoho případu, k znevýznamění u některých dříve statisticky významných rozdílů v přesnosti

překladu MLPH tabulky oproti nejlepší tabulce v dané skupině. Velikost MLPH_{WA} tabulky (11 658 a 11 143 frází) oproti MLPH tabulce (11 585 a 11 281 frází) pak zůstala víceméně nezměněna. Je tedy zřejmé, že použitím informací ze slovního přiřazení lze ještě dále zlepšit přesnost překladu vytvořeného frázovou tabulkou získanou metodou pro výběr frází založený na principu minimální ztráty.

Kapitola 8

Závěr

Hlavní cíl, tedy vytvoření automatického systému pro obousměrný překlad mezi češtinou a znakovanou řečí, byl splněn. Za tímto účelem byl vytvořen obecný statistický frázový překladový systém a paralelní korpus obsahující český text a jeho překlad do znakované češtiny. Tento korpus pak posloužil pro vytvoření systému pro automatický obousměrný překlad mezi češtinou a znakovanou češtinou. Vraťme se nyní k dílčím cílům definovaným v Kapitole 2 a zaměřme se na to, jak byly splněny za účelem dosažení hlavního cíle.

Jako první dílčí cíl bylo definováno vytvoření paralelního korpusu čeština – znakovaná čeština, který pak bude moci být použit při konstrukci systému pro automatický překlad. Vzhledem k současné situaci v případě českého znakového jazyka a obtížnosti zachycení jakéhokoliv znakového jazyka pomocí textu byla v tomto případě, kdy se jedná o vytvoření prvního existujícího paralelního korpusu čeština – znakovaná řeč, dána přednost znakované češtině (více o znakované řeči a znakované češtině viz Kapitola 1). Jako základ paralelního korpusu byl zvolen již existující český Human–Human Train Timetable dialogový korpus, do kterého byl přidán překlad dialogů do znakované češtiny. Tím vznikl Czech – Signed Czech paralelní korpus obsahující 15 772 dvojic vět, které na straně češtiny tvoří 4 081 jedinečných slov a na straně znakované češtiny pak 2 366 jedinečných znaků českého znakového jazyka. Volba HHTT korpusu, který obsahuje přepisy telefonních dotazů do informačního centra vlakových jízdních řádů, je výhodná především proto, že představuje dobře definované a ohraničené téma a také obsahuje řadu anotačních vrstev přinášejících informace, které mohly být využity k dalšímu zpřesnění automatického překladu. Za účelem přidání překladu do ZČ byla vytvořena dosud neexistující psaná forma ZČ. Vzhledem k tomu, že ZČ sdílí gramatická pravidla češtiny a používá jednotlivé znaky ČZJ, byla psaná forma ZČ navržena jako záznam znaků ČZJ v pořadí odpovídajícím českým slovům v překládané větě. Každý znak ČZJ byl pro tento účel reprezentován jedinečným řetězcem. Abychom zajistily konzistentnost překladů (tj. jednotné a jedinečné pojmenování znaků) používali všichni překladatelé při překladu stejný slovník. Jako základ tohoto slovníku jsme použili textovou verzi největšího dostupného slovníku ČZJ (viz [Langer 04], tento slovník obsahuje 3 063 znaků). Slovník jsme dále upravili tak, aby obsahoval pro každý znak jedinečný název, přidali jsme upřesňující popis u znaků, které to vyžadovaly, a také jsme přidali dva speciální znaky: znak „_“, který znamená prázdný překlad a znak „spelling“, který se používá pro slova, jež se hláskují pomocí prstové abecedy. Dále jsme také do tohoto slovníku přidali další znaky, které byly třeba pro překlad textů v korpusu. Tyto znaky byly buď převzaty z dalších slovníků nebo byly zjištěny přímo dotazem ve Spolku neslyšících Plzeň, celková velikost použitého slovníku byla tedy 3 185 znaků. Tím byla zajištěna skutečná existence všech znaků (tj. známe jejich prostorovou podobu), které byly použity při překladu tak, aby výsledný překladový systém byl použitelný pro automatickou syntézu znako-

vané řeči (při ní je třeba znát prostorovou podobu všech syntetizovaných znaků). Pro potřeby překladu a zajištění jeho konzistence byl také dále upraven stávající anotační nástroj použitý při vytváření anotace HHTT korpusu. Při překladu vět z HHTT korpusu bylo také anotátory vyznačeno slovní přiřazení mezi českou větou a jejím překladem do ZČ. Tato informace pak byla použita k vytvoření frázové tabulky pro automatický překlad.

Druhým dílčím úkolem bylo navržení vlastní metody výběru frází z paralelního korpusu. Navržená metoda je založena na využití log-lineárního modelu pro ohodnocení jednotlivých překladových párů a následném výběru nejlepších překladových párů z hlediska daného kritéria. Použití log-lineárního modelu umožňuje, díky kombinaci libovolných rysů, využití různorodých informací a díky optimalizaci vah tohoto modelu také optimální výběr frází a jejich překladů z hlediska kritéria použitého pro měření přesnosti výsledného překladu. Pro výběr nejlepších překladových párů bylo definováno nové kritérium založené na principu minimální ztráty. Tj. pro danou zdrojovou frázi ve větě, je jako nejlepší překlad vybrána ta cílová fráze (resp. ty cílové fráze) z odpovídajícího překladu, která použita jako překlad zbylých zdrojových frází způsobí nejmenší překladovou ztrátu, definovanou jako suma ohodnocení překladových párů, které tvoří vždy zvolená cílová fráze a popořadě zbylé zdrojové fráze. Toto kritérium bylo za stejných podmínek porovnáno se dvěma alternativními kritérii. Prvním kritériem bylo kritérium odpovídající MAP přístupu k výběru frází, kdy je pro každou zdrojovou frázi vybrán jen překladový pár s nejvyšším ohodnocením (resp. N překladových párů s nejvyšším ohodnocením). Druhým porovnávaným kritériem pak byla modifikace předešlého kritéria použitá v práci [Deng 08], kdy je vybráno N překladových párů s nejvyšším ohodnocením jen pro každý větný pár. Z tohoto porovnání vyšlo nově navržené kritérium jako nejvhodnější, neboť dosáhlo nejlepšího BLEU skóre 75,09, zbylá dvě kritéria pak jen skóre 72,83 resp. 72,31. Nově navržená metoda pak byla použita pro výběr frází při překladu z češtiny do znakované češtiny. Při tom byly vyzkoušeny různé kombinace rysů pro výběr frází navržené v Kapitole 4.4.1 a také možné kombinace použití optimalizace vah rysů pro výběr frází a optimalizace vah použitého dekodéru. Z těchto experimentů podle očekávání vyplynulo, že nejlepší je použít při výběru frází všechny navržené rysy a také obě optimalizace, takto bylo dosaženo BLEU skóre 76,51 pro vývojová, resp. 78,87 pro testovací data. Na základě výsledků těchto experimentů byla také navržena a vyzkoušena možná zlepšení této nové metody pro výběr frází založený na principu minimální ztráty. Jako první bylo na základě analýzy chyb výsledného překladu navrženo rozdělit výběr nejlepšího překladového páru podle frekvence výskytu dané zdrojové fráze. Bylo totiž zjištěno, že výběr jen překladového páru s nejnižší ztrátou vede v případě frází, které se vyskytly v korpusu jen jednou, v řadě případů k volbě chybného překladu. Výběr nejlepšího páru byl tedy rozdělen na dvě možnosti. V případě, že se zdrojová fráze vyskytla jen jednou, jsou do frázové tabulky vybrány všechny možné překladové páry bez ohledu na velikost překladové ztráty, jinak je vybrán jen překladový pár (páry) s nejnižší překladovou ztrátou. Další zlepšení se týkají použití průniku frázových tabulek pro jednotlivé směry překladu při vytváření výsledné frázové tabulky a dále metody filtrování výsledné tabulky, kdy je tato tabulka použita k překladu nějakého zvoleného textu (lze použít např. trénovací data) a je zaznamenáno, které překlady a kolikrát byly použity a na základě těchto informací je následně vytvořená nová frázová tabulka. Obě tato zlepšení umožňují různé způsoby použití optimalizace vah i vzájemné kombinace. Z vyzkoušených možností vyšla nejlépe kombinace obou metod označená v Tabulce 7.6 jako *Intersected_ID_Filtered_D*, kdy je metoda filtrace a optimalizace vah dekodéru použita na výslednou tabulku získanou prováděním nejprve průniku tabulek pro jednotlivé směry překladu a následnou optimalizací vah pro výběr frází, dokud není nalezena výsledná optimální tabulka (tabulka vybraná na základě optimálních vah pro výběr frází). Použitím tohoto postupu bylo dosaženo BLEU skóre 78,69 pro vývojová a 80,22 pro testovací data (absolutně tak došlo ke zlepšení o 2,18, resp. o 1,35 bodu BLEU skóre oproti

základní verzi metody pro výběr frází založený na principu minimální ztráty).

Třetím dílčím cílem bylo vytvoření vlastního dekodéru použitelného pro úlohu automatického překladu. Vlastní dekodér SiMPaD byl vytvořen implementací algoritmu pro monotónní prohledávání popsaného v Kapitole 5.1. Tento algoritmus využívá techniku dynamického programování pro nalezení nejlepšího překladu a také nově definuje obecně n-gramovou frázovou závislost právě vytvářeného překladu na předchozích překladech (frázích). Vstupem algoritmu je tabulka možných pokrytích a rozdělení vstupní věty na fráze a jim odpovídajících překladů, na niž je za účelem nalezení nejlepšího překladu (tj. nejlepší cesty stavovým prostorem, který je tvořen možnými překlady zdrojové věty) aplikováno prohledávání do šířky. V případě použití nemonotónního překladu pak stačí jen změnit vstupní tabulku tak, aby obsahovala i nemonotónní pokrytí zdrojové věty. Z praktického hlediska je pak ještě dále nutné použít prořezávání možných překladových hypotéz tak, aby výsledný překlad byl nalezen v přijatelném čase. V experimentech s dekodérem SiMPaD byl vyzkoušen bigramový a trigramový frázový n-gram v kombinaci s bigramovým až 4-gramovým jazykovým modelem. Jako nejvhodnější z hlediska rychlosti a přesnosti překladu vyšla v tomto případě kombinace bigramového frázového n-gramu s trigramovým jazykovým modelem, tato kombinace také byla následně použita v dalších experimentech. Ze srovnání se standardním frázovým dekodérem MOSES vyplynulo, že oba dekodéry poskytují plně srovnatelné výsledky, viz následující dílčí úkol.

V rámci čtvrtého dílčího úkolu jsme vzájemně porovnali výsledky překladu z češtiny do znakované češtiny s ručně vytvořenými frázemi (HPH) s úspěšností překladu s frázemi získanými automaticky a to standardní metodou založenou na automaticky vytvořeném slovním přiřazení popsanou v Kapitole 4.1.1 (MPH) a nově navrženou metodou pro výběr frází založený na principu minimální ztráty popsanou v této práci v Kapitole 4.4 (MLPH). K experimentům byl použit zmíněný CSC korpus, který byl rozdělen na trénovací (12 616 vět), vývojovou (1 578 vět) a testovací (1 578 vět) část. Pro srovnání byly při překladu použity oba dekodéry, váhy pro výběr frází a váhy použitých dekodérů byly optimalizovány z hlediska BLEU kritéria. Pro porovnání referenčního a vytvořeného překladu byla také definována nová míra SDO založená na překryvu jejich sémantické anotace vytvořené sémantickým parserem, který byl natrénován na datech CSC korpusu (data potřebná pro natrénování parseru pro ZČ byla získána vhodnou úpravou jedné z anotačních vrstev HHTT korpusu). Jako sémantický parser byl použit upravený HVS parser popsaný v prácích [Jurčíček 07, Jurčíček 08]. Jako nejlepší z tohoto srovnání vyšla HPH tabulka, která pro testovací data dosáhla 81,29 bodu BLEU skóre (resp. 81,22 v případě SiMPaD dekodéru), druhá pak MPH tabulka, která dosáhla 81,05 (resp. 81,08) a poslední MLPH tabulka s 80,21 (resp. 80,20) bodu BLEU skóre. Z rozdílů mezi tabulkami vyplývá, že automatické metody pro výběr frází dosahují výsledků, které jsou v případě naší úlohy plně srovnatelné s výsledky ručně vytvořených frází (rozdíly mezi jednotlivými tabulkami jsou z praktického hlediska zanedbatelné). Nevýhoda MPH tabulky oproti zbylým dvěma pak spočívá v její velikosti, neboť obsahuje 65 494 frází (fráze maximální délky tři), zatímco MLPH tabulka jen 11 585 a HPH tabulka dokonce jen 5 325 frází.

Poslední dva dílčí úkoly se týkají otestování základního systému pro překlad z češtiny do znakované češtiny a návrh a otestování jeho úprav při obousměrném překladu těchto jazyků. K experimentům byl opět použit CSC korpus se stejným rozdělením na části jako v předešlých experimentech a SiMPaD dekodér, všechny váhy byly opět optimalizovány z hlediska BLEU skóre. Základní systém i všechny úpravy byly vyzkoušeny pro všechny tři frázové tabulky. Hodnoty základního systému pro překlad jsou známy z předchozího experimentu a jsou to: 81,22 pro HPH, 81,08 pro MPH a 80,20 bodu BLEU skóre pro MLPH tabulku. Pro zlepšení přesnosti překladu byla navržena dvě možná zlepšení. První možností je využití třídního jazykového modelu založeného na pojmenovaných entitách, které jsou obsaženy v CSC kor-

pusu. Do standardního jazykového modelu byly přidány třídy pro všechna známa jména stanic: STATION a osob: PERSON a dále také pro oblast: AREA a typ vlaku: TRAIN. Protože sémantická anotace CSC korpusu obsahuje označení těchto pojmenovaných entit, můžeme je v datech jednoduše nahradit příslušnou třídou a vytvořit z nich odpovídající překladové slovníky. Použitím třídního jazykového modelu jsme dostali tyto výsledky: 82,40 pro HPH, 82,33 pro MPH a 81,21 bodu BLEU skóre pro MLPH tabulku. Ve všech případech tak došlo ke zlepšení překladu o více než jedno procento. To je způsobeno především redukcí OOV slov na nulu a také poklesem perplexity jazykového modelu pro testovací data o 20,06 procenta. Rozdíly v přesnosti překladu mezi jednotlivými tabulkami zůstaly zachovány na stejné úrovni. Dále je možné použít pozpracování získaných překladů, které upraví získaný překlad do konečné podoby. Za prvé můžeme z výsledného překladu vynechat slova, která se nepřekládají (resp. jsou přeložena použitím znaku „_“ pro prázdný překlad). V trénovacích datech je ovšem dobré tato slova ponechat, neboť pak dostaneme lepší výsledky překladu díky detailnějším a tím jednoznačnějším překladovým a jazykovým modelům. Za druhé můžeme neznámá slova nahradit znakem pro hláskování (spelling), neboť slova, pro něž ve znakované řeči neexistuje znak, se hláskují pomocí prstové abecedy. Použitím pozpracování na výsledky překladu získaného s využitím třídního jazykového modelu dostaneme jednak konečnou podobu překladu (odstranění prázdných překladů), kterou lze prezentovat jako výstup překladového systému, a dále má použití pozpracování také pozitivní vliv na přesnost vytvořeného překladu, pro který jsme dostali výsledky: 83,21 pro HPH (4 203 frází), 83,38 pro MPH (56 549 frází) a 82,42 bodu BLEU skóre pro MLPH (11 281 frází) tabulku. Ve všech případech tak došlo opět ke zlepšení překladu o další více než jedno procento (kromě HPH tabulky, kde je zlepšení nižší). To je opět jako v případě použití třídního jazykového modelu způsobeno poklesem perplexity jazykového modelu o dalších 17,49 procenta. Celkově se nám tedy povedlo absolutně zlepšit výsledky překladu o 1,99 pro HPH, o 2,30 pro MPH a o 2,22 bodu BLEU skóre pro MLPH tabulku oproti výsledkům základního překladového systému.

Stejný postup, tj. aplikace třídního jazykového modelu, byl použit i v případě opačného směru překladu ze znakované češtiny do češtiny. Aby mohla být stávající data z CSC korpusu použita pro natrénování systému pro daný směr překladu, bylo třeba je ještě dodatečně upravit. Tato úprava spočívala ve vypuštění prázdných překladů (znak „_“) a jim odpovídajícím českým slovům. Zatímco totiž v předešlém směru překladu byla vstupem česká věta, která obsahovala všechna zapsaná slova, nyní je vstupem věta ve znakované češtině, která by, kdyby byla produkována samotným neslyšícím, žádné prázdné překlady neobsahovala. A jí tedy také tím pádem odpovídá česká věta, která vznikne vynecháním českých slov, jež se do znakované češtiny nepřekládají. Podle očekávání jsou výsledky pro překlad ze znakované češtiny do češtiny horší než v opačném případě směru překladu (viz výše). Rozdíl činí asi 20 procentních bodů v případě BLEU kritéria, s ohledem na nějž byla opět přesnost překladu optimalizována (konkrétně bylo dosaženo těchto přesností překladu: 63,80 pro HPH, 63,98 pro MPH a 61,69 bodu BLEU skóre pro MLPH tabulku). Tento rozdíl je způsoben především tím, že v případě překladu ze znakované češtiny do češtiny je přítomna větší nejednoznačnost, neboť překládáme z menšího slovníku na větší, než v opačném případě. Z hlediska jednotlivých tabulek je situace obdobná jako u předešlého směru překladu. HPH a MPH tabulky opět dosahují srovnatelných výsledků, zatímco MLPH tabulka za nimi zaostává, přičemž její ztráta oproti předešlému směru překladu ještě narostla, v případě BLEU kritéria je ztráta vyšší přibližně o dalších 1,3 procenta. Výhoda menší velikosti HPH (3 621 frází) a MLPH (9 217 frází) tabulek oproti MPH (45 711 frází) tabulce zůstala zachována.

8.1 Další práce

Hlavním úkolem z hlediska překladu mezi češtinou a znakovanou řečí je nyní vytvořit paralelní korpus českého znakového jazyka tak, aby mohl být stávajícími metodami vyvinutými v této práci vytvořen automatický překladový systém mezi češtinou a českým znakovým jazykem. Jako klíčová se pak v tomto případě jeví psaná forma českého znakového jazyka, která by měla být navržena tak, aby dokázala zachytit všechny fenomény znakového jazyka naznačené v Kapitole 1.2. A dále byla použitelná pro systém automatické syntézy z českého textu do znakované řeči, jehož součástí by vytvořený systém automatického překladu měl být. Ideální by pak také bylo, kdyby tato snaha o vytvoření psané formy ČZJ vzešla přímo z komunity Neslyšících a byla jimi tak plně podporována a nebo přinejmenším s komunitou Neslyšících a mluvčích ČZJ co možná nejdůkladněji konzultována.

V případě SiMPaD dekodéru a jeho použití pro překlad ČZJ i dalších jazyků je prioritní implementace nemonotónního překladu. Znamená to tedy především přidání možnosti generovat tabulku možných pokrytí vstupní věty tak, aby obsahovala i nemonotónní pokrytí zdrojové věty. Samotné prohledávání je pak třeba doplnit o prořezávání možných hypotéz tak, aby mohl být výsledný překlad nalezen v přijatelném čase. V případě použitého frázového n-gramu se pak nabízí možnost návrhu a využití nových rysů profitujících ze zachycené frázové závislosti.

Z porovnání různých frázových tabulek vyplývá, že současné automatické metody pro výběr frází poskytují z hlediska přesnosti překladu výsledky plně srovnatelné s ručně vytvořenou frázovou tabulkou. Zároveň však z toho samého porovnání také vyplývá, že zaostávají ve velikosti vytvořené tabulky, která je několikanásobně větší než srovnatelně přesná, ručně vytvořená tabulka. Řešením tohoto problému, tedy výběru malé tabulky a zároveň co nejlepší z hlediska přesnosti překladu, by mohla být nová metoda pro výběr frází založený na principu minimální ztráty představená v této práci. Tato metoda však zatím v přesnosti překladu mírně zaostává za dosud nejpoužívanější automatickou metodou. Vzhledem k tomu však, že tato nová metoda využívá k výběru frází log-lineární model, lze její přesnost překladu jednoduše vylepšit přidáním dalších rysů, které zaručí výběr správných překladových párů. Takovýmito rysy mohou být např. rysy založené na slovním přiřazení, které stojí za přesností překladu stávajících metod pro výběr frází. Vliv rysů založených na slovním přiřazení na přesnost překladu vybrané tabulky byl pak vyzkoušen v posledním experimentu v předcházející kapitole, kdy byl pro ilustraci použit jednoduchý rys založený na lexikální překladové pravděpodobnosti. Z uvedených výsledků je zřejmé, že použití rysu založeného na slovním přiřazení má pozitivní vliv na přesnost překladu při zachování kompaktní velikosti tabulky. Další možnosti jak zlepšit překladovou přesnost nové metody dále spočívají v použití nových ztrátových funkcí pro výběr „dobrých“ překladů, rozdělení výběru „dobrých“ překladů na více skupin podle počtu objevení se dané zdrojové fráze (resp. lze uvažovat o rozdělení výběru např. podle spolehlivosti odhadu překladové pravděpodobnosti daného překladového páru, tj. jestliže je odhad spolehlivý, můžeme vybrat méně „dobrých“ překladů, v opačném případě je výhodnější ponechat více možných překladů a výběr nechat až na samotném dekodéru). Přesnost lze také dále zlepšit použitím účinnějších metod pro hledání optimálních vah log-lineárního modelu (nabízí se např. deterministické žíhání a další). Metodu filtrace lze pak dále vylepšit např. použitím N nejlepších hypotéz, na jejichž základě budou vybrány překlady používané při překladu zvoleného textu a dále také porovnáním vytvořených překladů s referenčním překladem zvoleného textu.

Příloha A

Ukázka překladu

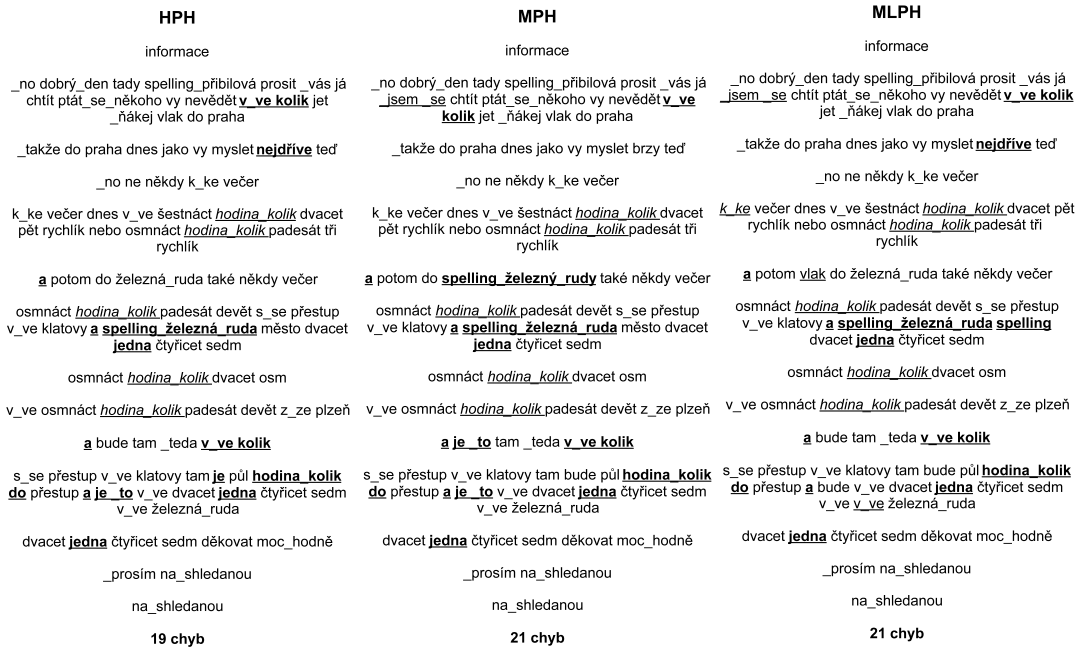
Na následujících obrázcích je ukázka překladu jednoho celého dialogu z CSC korpusu. Jsou ukázány překlady pro různé úpravy vstupních i výstupních dat a pro všechny tři testované frázové tabulky. Nechybí ani ukázka opačného směru překladu ze znakované češtiny do češtiny. Na Obrázku A.1 jsou vstupní data, tj. český text, který se překládá. Vlevo je zdrojový text v základní podobě a vpravo pak při použití třídního jazykového modelu. Na následujícím Obrázku A.2 jsou pak odpovídající referenční překlady daného vstupního textu. Vlevo je opět referenční překlad odpovídající zdrojovému textu v základní podobě, uprostřed pak referenční překlad odpovídající použití třídního jazykového modelu a konečně vpravo je referenční překlad odpovídající použití pozpracovaní a třídního jazykového modelu. Na Obrázcích A.3 až A.5 jsou překlady daného zdrojového textu vytvořené použitím všech tří porovnávaných frázových tabulek (vlevo je vždy výsledek pro ručně vytvořenou frázovou tabulku (HPH), uprostřed pro frázovou tabulku vytvořenou automaticky standardním postupem (MPH) a vpravo pro tabulku vytvořenou automaticky metodou popsanou v této práci, založenou na principu minimální ztráty (MLPH)). Nejprve jsou výsledky překladu pro základní podobu zdrojového textu, pak při použití třídního jazykového modelu a konečně při použití pozpracování a třídního jazykového modelu. Na dalších dvou Obrázcích A.6 a A.7 jsou ukázky sémantické anotace přiřazené použitým a vytvořeným cílovým textům. Na Obrázku A.6 je nahoře referenční sémantická anotace odpovídající cílovému textu a pod ní jsou sémantické anotace vytvořené HVS parserem pro všechny tři možné podoby cílového textu (základní, použití třídního jazykového modelu a použití pozpracování a třídního jazykového modelu). Na Obrázku A.7 jsou pak sémantické anotace vytvořené stejným HVS parserem pro překlady získané použitím všech tří frázových tabulek. V horní části je anotace odpovídající překladu zdrojového textu v základní podobě, pod ní pak anotace odpovídající použití třídního jazykového modelu (anotace pro pozpracování a použití třídního jazykového modelu je stejná). Tyto sémantické anotace jsou použity pro porovnání při vyhodnocení SDO kritéria. Na posledních dvou Obrázcích A.8 a A.9 je pak ukázka opačného směru překladu daného dialogu ze znakované češtiny do češtiny, tj. zdrojový a cílový text jsou adekvátně prohozeny a upraveny, viz Kapitola 7.8.

Zdrojový text	Zdrojový text
informace	informace
no dobrý den tady přibilová prosím vás já jsem se chtěla zeptat nevíte v kolik jede ňákej vlak do prahy	no dobrý den tady person prosím vás já jsem se chtěla zeptat nevíte v kolik jede ňákej vlak do station
takže do prahy dneska jako myslíte nejbližší teďko	takže do station dneska jako myslíte nejbližší teďko
no ne někdy k večeru	no ne někdy k večeru
k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík	k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík
a potom do železný_rudy taky někdy večer	a potom do station taky někdy večer
osmnáct padesát devět s přestupem v klatovech a železná_ruda město dvacet jedna čtyřicet sedm	osmnáct padesát devět s přestupem v station a station město dvacet jedna čtyřicet sedm
osmnáct dvacet osm	osmnáct dvacet osm
v osmnáct padesát devět z plzně	v osmnáct padesát devět z station
a je to tam teda v kolik	a je to tam teda v kolik
s přestupem v klatovech tam je půl hodiny na přestup a je to ve dvacet jedna čtyřicet sedm v železné_rudě	s přestupem v station tam je půl hodiny na přestup a je to ve dvacet jedna čtyřicet sedm v station
dvacet jedna čtyrycet sedm děkuju moc	dvacet jedna čtyrycet sedm děkuju moc
prosím nashledanou	prosím nashledanou
nashledanou	nashledanou

Obrázek A.1: Ukázka zdrojového textu, vlevo v základní podobě, vpravo při použití třídního jazykového modelu.

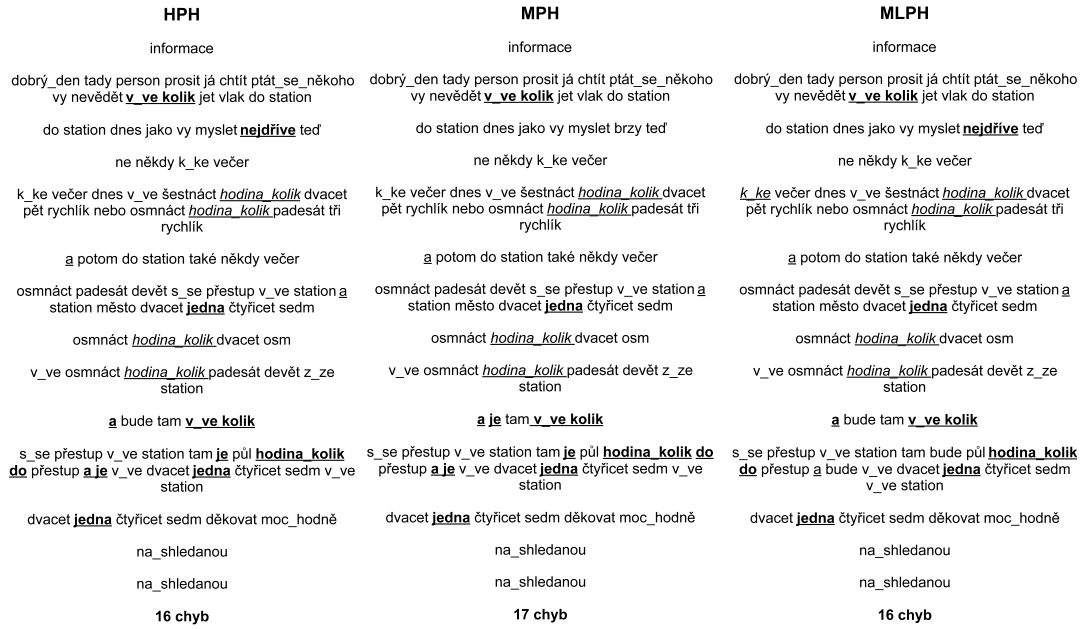
Cílový text	Cílový text	Cílový text
informace	informace	informace
_no dobrý_den tady spelling_přibilová prosit _vás já chtít ptát_se_někoho vy nevědět kdy jet _ňákej vlak do praha	_no dobrý_den tady person prosit _vás já chtít ptát_se_někoho vy nevědět kdy jet _ňákej vlak do station	dobrý_den tady person prosit já chtít ptát_se_někoho vy nevědět kdy jet vlak do station
_takže do praha dnes jako vy myslet brzy teď	_takže do station dnes jako vy myslet brzy teď	do station dnes jako vy myslet brzy teď
_no ne někdy k_ke večer	_no ne někdy k_ke večer	ne někdy k_ke večer
k_ke večer dnes v_ve šestnáct hodina_kolik dvacet pět rychlík nebo osmnáct hodina_kolik padesát tři rychlík	k_ke večer dnes v_ve šestnáct hodina_kolik dvacet pět rychlík nebo osmnáct hodina_kolik padesát tři rychlík	k_ke večer dnes v_ve šestnáct hodina_kolik dvacet pět rychlík nebo osmnáct hodina_kolik padesát tři rychlík
_a potom do železná_ruda také někdy večer	_a potom do station také někdy večer	potom do station také někdy večer
osmnáct hodina_kolik padesát devět s_se přestup v_ve klatovy _a železná_ruda město dvacet jedna_hodina čtyřicet sedm	osmnáct hodina_kolik padesát devět s_se přestup v_ve station _a station město dvacet jedna_hodina čtyřicet sedm	osmnáct hodina_kolik padesát devět s_se přestup v_ve station station město dvacet jedna_hodina čtyřicet sedm
osmnáct hodina_kolik dvacet osm	osmnáct hodina_kolik dvacet osm	osmnáct hodina_kolik dvacet osm
v_ve osmnáct hodina_kolik padesát devět z_ze plzeň	v_ve osmnáct hodina_kolik padesát devět z_ze station	v_ve osmnáct hodina_kolik padesát devět z_ze station
_a bude tam _teda kdy	_a bude tam _teda kdy	bude tam kdy
s_se přestup v_ve klatovy tam bude půl hodina_jak_dlouho pro přestup _a bude v_ve dvacet jedna_hodina čtyřicet sedm v_ve železná_ruda	s_se přestup v_ve station tam bude půl hodina_jak_dlouho pro přestup _a bude v_ve dvacet jedna_hodina čtyřicet sedm v_ve station	s_se přestup v_ve station tam bude půl hodina_jak_dlouho pro přestup bude v_ve dvacet jedna_hodina čtyřicet sedm v_ve station
dvacet jedna_hodina čtyřicet sedm děkovat moc_hodně	dvacet jedna_hodina čtyřicet sedm děkovat moc_hodně	dvacet jedna_hodina čtyřicet sedm děkovat moc_hodně
_prosím na_shledanou	_prosím na_shledanou	na_shledanou
na_shledanou	na_shledanou	na_shledanou

Obrázek A.2: Ukázka referenčních překladů předchozího zdrojového textu, vlevo překlad odpovídající základní podobě, uprostřed překlad odpovídající použití třídního jazykového modelu a vpravo překlad odpovídající použití pozpracování a třídního jazykového modelu.

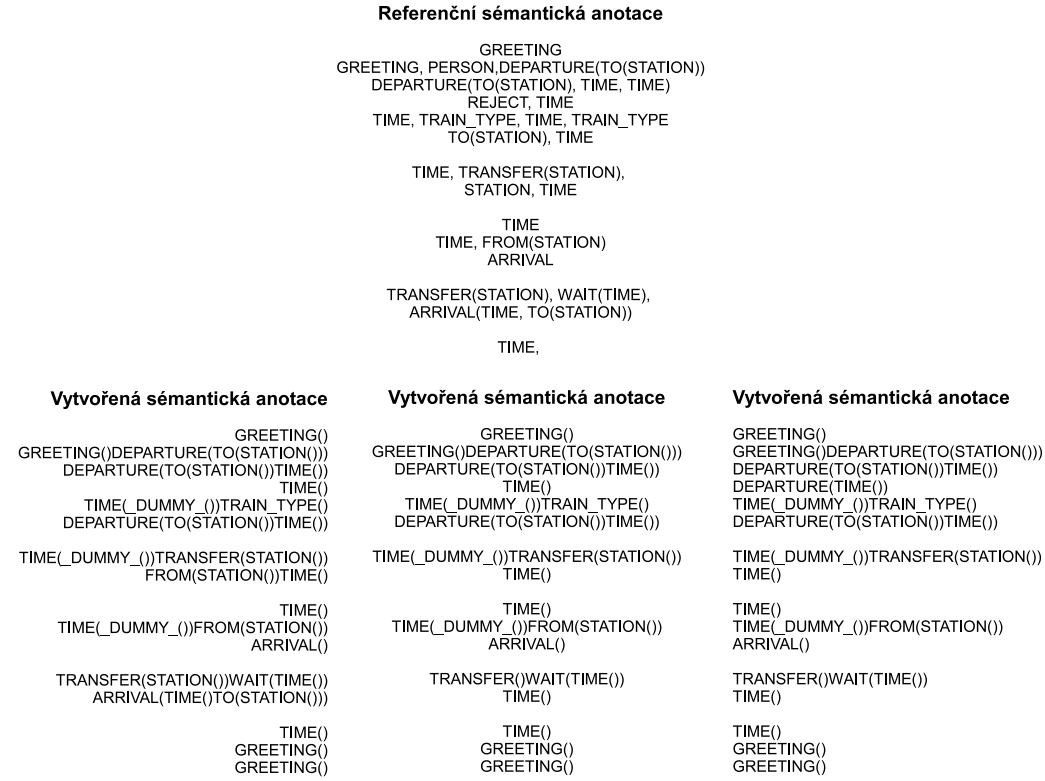


Obrázek A.3: Ukázka překladu zdrojového textu v základní podobě pro všechny tři porovnávané frázové tabulky.

Obrázek A.4: Ukázka překladu zdrojového textu při použití třídního jazykového modelu pro všechny tři porovnávané frázové tabulky.



Obrázek A.5: Ukázka překladu zdrojového textu při použití pozpracování a třídního jazykového modelu pro všechny tři porovnávané frázové tabulky.



Obrázek A.6: Ukázka referenční a HVS parserem vytvořené sémantické anotace odpovídající různým podobám cílového textu na Obrázku A.2.

HPH	MPH	MLPH
GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())	GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())	GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())
TIME(_DUMMY_())TRANSFER(STATION() STATION())TIME()	TIME(_DUMMY_())TRANSFER(STATION() STATION())TIME()	TIME(_DUMMY_())TRANSFER(STATION()) TIME()
TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()	TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()	TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()
TRANSFER(STATION()) ARRIVAL(TIME()TO(STATION()))	TRANSFER(STATION()) ARRIVAL(TIME()TO(STATION()))	TRANSFER(STATION()) ARRIVAL(TIME()TO(STATION()))
TIME() GREETING() GREETING()	TIME() GREETING() GREETING()	TIME() GREETING() GREETING()

HPH	MPH	MLPH
GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())	GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())	GREETING() GREETING()DEPARTURE(TO(STATION())) DEPARTURE(TO(STATION())TIME() TIME() TIME(_DUMMY_())TRAIN_TYPE() DEPARTURE(TO(STATION())TIME())
TIME(_DUMMY_())TRANSFER(STATION()) TIME()	TIME(_DUMMY_())TRANSFER(STATION()) TIME()	TIME(_DUMMY_())TRANSFER(STATION()) TIME()
TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()	TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()	TIME() TIME(_DUMMY_())FROM(STATION()) ARRIVAL()
TRANSFER()TIME()	TRANSFER()TIME()	TRANSFER()TIME()
TIME() GREETING() GREETING()	TIME() GREETING() GREETING()	TIME() GREETING() GREETING()

Obrázek A.7: Ukázka HVS parserem vytvořené sémantické anotace odpovídající překladům vytvořeným všemi třemi porovnávanými frázovými tabulkami pro základní podobu zdrojového textu (nahoře) a při použití třídního jazykového modelu (dole).

Zdrojový text	Cílový text
informace	informace
dobrý_den tady person prosit já chtít ptát_se_někoho vy nevědět kdy jet vlak do station	dobrý den tady person prosím já jsem se chtěla zeptat nevíte v kolik jede vlak do station
do station dnes jako vy myslet brzy teď	do station dneska jako myslíte nejbližší teďko
ne někdy k_ke večer	ne někdy k večeru
k_ke večer dnes v_ve šestnáct hodina_kolik dvacet pět rychlík nebo osmnáct hodina_kolik padesát tři rychlík	k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík
potom do station také někdy večer	potom do station taky někdy večer
osmnáct hodina_kolik padesát devět s_se přestup v_ve station station město dvacet jedna_hodina čtyřicet sedm	osmnáct padesát devět s přestupem v station station město dvacet jedna čtyřicet sedm
osmnáct hodina_kolik dvacet osm	osmnáct dvacet osm
v_ve osmnáct hodina_kolik padesát devět z_ze station	v osmnáct padesát devět z station
bude tam kdy	je to tam v kolik
s_se přestup v_ve station tam bude půl hodina_jak_dlouho pro přestup bude v_ve dvacet jedna_hodina čtyřicet sedm v_ve station	s přestupem v station tam je půl hodiny na přestup je to ve dvacet jedna čtyřicet sedm v station
dvacet jedna_hodina čtyřicet sedm děkovat moc_hodně	dvacet jedna čtyrycet sedm děkuju moc
na_shledanou	nashledanou
na_shledanou	nashledanou

Obrázek A.8: Ukázka zdrojového a cílového textu při překladu ze znakované češtiny do češtiny.

HPH	MPH	MLPH
informace	informace	informace
dobrý den tady person prosím já jsem se chtěla zeptat nevíte v kolik jede vlak do station	dobrý den tady person **prosim** já *jsem se* chtěla zeptat nevíte v kolik jede vlak do station	dobrý den tady person **prosim** já *jsem se* chtěla zeptat nevíte v kolik jede vlak do station
do station dneska jako myslíte **brzo teď**	do station dneska jako myslíte nejbližší teďko	do station dneska jako myslíte **brzo teď**
ne někdy k večeru	ne někdy k večeru	ne někdy k večeru
k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík	k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík	k večeru dneska v šestnáct dvacet pět rychlík nebo osmnáct padesát tři rychlík
potom do station taky někdy večer	potom do station taky někdy večer	potom do station taky někdy večer
osumnáct padesát devět s přestupem v station station město dvacet jedna čtyřicet sedm	osmnáct padesát devět s přestupem v station station **města** dvacet jedna čtyřicet sedm	osmnáct padesát devět s přestupem v station *station* **městě jednadvacátou** čtyřicet sedm
osmnáct dvacet osm	osmnáct dvacet osm	osmnáct šestnáct dvacet osm
v osmnáct padesát devět z station	v osmnáct padesát devět z station	v osmnáct padesát devět z station
je *to* tam v kolik	je *to* tam v kolik	je *to* tam v kolik
s přestupem v station tam je půl hodiny na přestup je *to* ve dvacet jedna čtyřicet sedm v station	s přestupem v station tam je půl hodiny na přestup je *to* ve dvacet jedna čtyřicet sedm v station	s přestupem v station tam je půl hodiny hodinu na přestup je *to* **jednadvacátý** *jedna* čtyřicet sedm v station
dvacet jedna **čtyřicet** sedm děkuji moc	dvacet jedna **čtyřicet** sedm děkuji moc	**jednadvacátou** *jedna* **čtyřicet** sedm děkuji moc
nashledanou	nashledanou	nashledanou
nashledanou	nashledanou	nashledanou
6 chyb	**6 chyb**	**16 chyb**

Obrázek A.9: Ukázka překladu ve směru znakovaná čeština – čeština při použití pozpracování a třídního jazykového modelu pro všechny tři porovnávané frázové tabulky.

Literatura

[Allen 97] J. Allen & M. Core. *DAMSL: Dialog Act Markup in Several Layer.* http://www.cs.rochester.edu/research /cisd/resources/damsl, 1997.

[Ayan 06] N.F. Ayan & B.J. Dorr. *Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT.* In Proc. 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (COLING/ACL), pages 9–16, Sydney, Australia, July 2006.

[Badler 00] N. Badler, R. Bindiganavale, J. Allbeck, W. Schuler, L. Zhao, S. Lee, H. Shin & M. Palmer. *Parameterized Action Representation and Natural Language Instructions for Dynamic Behavior Modification of Embodied Agents.* In AAAI Spring Symposium, 2000.

[Banerjee 05] S. Banerjee & A. Lavie. *Meteor: An automatic metric for MT evaluation with improved correlation with human judgments.* In Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, Ann Arbor, Michigan, 2005.

[Bauer 99] B. Bauer, S. Nießen & H. Hienz. *Towards an automatic Sign Language translation system.* In Proc. of the Int. Workshop on Physicality and Tangibility in Interaction, Siena, Italy, 1999.

[Baum 72] L. E. Baum. *An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process.* Inequalities, vol. 3, pages 1–8, 1972.

[Berger 96] A.L. Berger, P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.R. Gillett, A.S. Kehler & R.L. Mercer. *Language Translation Apparatus and Method of Using Context-based Translation Models.* Rapport technique, United States Patent 5510981, April 1996.

[Birch 06] A. Birch, Ch. Callison-Burch & M. Osborne. *Constraining the phrase-based, joint probability statistical translation model.* In Proceedings on the Workshop on Statistical Machine Translation, 2006.

[Bímová 02] P. Bímová. *Jazyk znakový – jazyk přirozený.* ČDS, vol. 10, pages 100–103, 2002.

[Bojar 06] O. Bojar & Z. Žabokrtský. *CzEng: Czech-English Parallel Corpus, Release version 0.5.* Prague Bulletin of Mathematical Linguistics, vol. 86, pages 59–62, 2006.

[Bojar 08] O. Bojar. *Exploiting Linguistic Data in Machine Translation*. PhD thesis, MFF UK, Prague, Czech Republic, 2008.

[Bos 94] J. Bos, E. Mastenbroek, S. McGlashan, S. Millies & M. Pinkal. *A Compositional DRS-based Formalism for NLP Applications*. Rapport technique, Universitaet des Saarlandes, 1994.

[Brown 88] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, R.L. Mercer & P.S. Roossin. *A Statistical Approach to Language Translation*. In 12th Int. Conf. on Computational Linguistics (COLING), pages 71–76, Budapest, Hungary, August 1988.

[Brown 90] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer & P.S. Roossin. *A Statistical Approach to Machine Translation*. Computational Linguistics, vol. 16, pages 79–85, 1990.

[Brown 93] P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra & R.L. Mercer. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, vol. 19, pages 263–311, 1993.

[Bungeroth 04] J. Bungeroth & H. Ney. *Statistical sign language translation*. In LREC 2004, Workshop proceedings: Representation and Processing of Sign Languages, pages 105–108, Lisbon, Portugal, May 2004.

[Bungeroth 08] J. Bungeroth, D. Stein, P. Dreuw, H. Ney, S. Morrissey, A. Way & L. van Zijl. *The ATIS Sign Language Corpus*. In International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, May 2008.

[Callison-Burch 06] C. Callison-Burch, M. Osborne & P. Koehn. *Reevaluating the Role of BLEU in Machine Translation Research*. In Proc. 11th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL), pages 249–256, Trento, Italy, April 2006.

[Callison-Burch 07] Ch. Callison-Burch, C. Fordyce, P. Koehn, Ch. Monz & J. Schroeder. *(Meta-) Evaluation of Machine Translation*. In ACL Workshop on Statistical Machine Translation 2007, 2007.

[Carletta 96] J. Carletta. *Assessing Agreement on Classification Tasks: The Kappa Statistic*. Computational Linguistics, vol. 22, pages 249–254, 1996.

[Chiang 05] D. Chiang. *A hierarchical phrase-based model for statistical machine translation*. In Proceedings of the Association for Computational Linguistics (ACL), pages 263–270, 2005.

[Chiang 07] D. Chiang. *Hierarchical phrase-based translation*. Computational Linguistics, vol. 33, 2007.

[Cohen 60] J. Cohen. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, vol. 20, no. 1, pages 37–46, 1960.

[Collins 02] M. Collins. *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms*. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 1–8, Philadelphia, Pennsylvania, 2002.

[Collins 05] M. Collins, P. Koehn & I. Kučerová. *Clause restructuring for statistical machine translation*. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 531–540, Ann Arbor, Michigan, June 2005.

[Darroch 72] J. N. Darroch & D. Ratcliff. *Generalized iterative scaling for log-linear models*. Annals of Mathematical Statistics, vol. 43, pages 1470–1480, 1972.

[Dempster 77] A. E Dempster, N. M. Laird & D. B Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, vol. 39(B), pages 1–38, 1977.

[Denero 06] J. Denero, D. Gillick, J. Zhang & D. Klein. *Why generative phrase models underperform surface heuristics*. In Proceedings of NAACL Workshop on Statistical Machine Translation, pages 31–38, 2006.

[Deng 08] Y. Deng, J. Xu & Y. Gao. *Phrase Table Training for Precision and Recall: What Makes a Good Phrase and a Good Phrase Pair?* In Proceedings of ACL-08: HLT, pages 81–88, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[Doddington 02] G. Doddington. *Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics*. In Proc. ARPA Workshop on Human Language Technology, 2002.

[Dorr 98] B. J. Dorr, Jordan P. W. & Benoit J. W. *A Survey of Current Paradigms in Machine Translation*. Rapport technique CS-TR-3961, 1998.

[Duda 00] R.O. Duda, P.E. Hart & D.G. Stork. Pattern classification. John Wiley and Sons, New York, NY, 2nd edition edition, 2000.

[Eisner 03] J. Eisner. *Learning Non-Isomorphic Tree Mappings for Machine Translation*. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 205–208, Sapporo, Japan, 2003. Association for Computational Linguistics.

[Elliott 00] R. Elliott, J.R.W. Glauert, J.R. Kennaway & I. Marshall. *The Development of Language Processing Support for the ViSiCAST Project*. In Assets 2000. 4th International ACM SIGCAPH Conference on Assistive Technologies, New York, 2000.

[Fraser 05] A. Fraser & D. Marcu. *ISI's Participation in the Romanian-English Alignment Task*. In Proceedings of the ACL Workshop on Building and Using Parallel Texts, pages 91–94, Ann Arbor, Michigan, 2005.

[Fraser 07] A. Fraser & D. Marcu. *Measuring Word Alignment Quality for Statistical Machine Translation*. Computational Linguistics, vol. 33, pages 293–303, September 2007.

[Gale 93] W.A. Gale & K.W. Church. *A program for aligning sentences in bilingual corpora*. Computational Linguistics, vol. 19, no. 1, pages 75–90, 1993.

[Giménez 07] J. Giménez & L. Márquez. *Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems*. In Proceedings of the ACL-2007 Workshop

on Statistcal Machine Translation (WMT-07), Prague, Czech Republic, 2007.

[Hajič 02] J. Hajič, M. Čmejrek, B. Dorr, Y. Ding, J. Eisner, D. Gildea, T. Koo, K. Parton, G. Penn, D. Radev & O. Rambow. *Natural Language Generation in the Context of Machine Translation*. Rapport technique, Johns Hopkins University, Center for Speech and Language Processing, 2002.

[Hajič 06] J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský & M. Ševčíková Razímová. *Prague Dependency Treebank 2.0*. Linguistic Data Consortium, Cat. No. LDC2006T01, 2006.

[Held 62] M. Held & R.M. Karp. *A Dynamic Programming Approach to Sequencing Problems*. Journal of the Society of Industrial and Applied Mathematics (SIAM), vol. 10, no. 1, pages 196–210, 1962.

[Hemphill 90] C.T. Hemphill, J.J. Godfrey & G.R. Doddington. *The ATIS Spoken Language Systems pilot corpus*. In Proceedings of DARPA Speech and Natural Language Workshop, pages 96–101, Hidden Valley, PA., 1990.

[Hoidekr 06] J. Hoidekr, J. V. Psutka, A. Pražák & J. Psutka. *Benefit of a class-based language model for real-time closed-captioning of TV ice-hockey commentaries*. In Proceedings of LREC 2006, pages 2064–2067, Paris, France, 2006. ELRA.

[Hronová 02] A. Hronová. *Poznáváme český znakový jazyk III. (Tvoření tázacích vět)*. Speciální pedagogika, vol. 12, no. 2, pages 113–123, 2002.

[Hrubý 03] J. Hrubý. *Boj o přijetí zákona o znakové řeči v ČR*. http://www.cktzj.com/certifikace/prednaska_20032004_zr_soubory/frame.htm, 2003.

[Huenerfauth 03] M. Huenerfauth. *A Survey and Critique of American Sign Language Natural Language Generation and Machine Translation Systems*. Rapport technique, Computer and Information Science, University of Pennsylvania, 2003.

[Huenerfauth 04a] M. Huenerfauth. *A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation*. In Proceedings of the Student Workshop at the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004), Boston, MA, USA, 2004.

[Huenerfauth 04b] M. Huenerfauth. *Spatial Representation of Classifier Predicates for Machine Translation into American Sign Language*. In Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal, 2004.

[Huenerfauth 06] M. Huenerfauth. *Generating American Sign Language Classifier Predicates For English-To-ASL Machine Translation*. PhD thesis, University of Pennsylvania, 2006.

[Ittycheriah 05] A. Ittycheriah & S. Roukos. *A maximum entropy word aligner for Arabic-English machine translation.* In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 89–96, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.

[Jelinek 98] F. Jelinek. Statistical methods for speech recognition. MIT Press, Cambridge, MA, 1998.

[Jelínek 03] L. Jelínek. *Porozumění telefonickému dotazu pro automatickou informační službu.* PhD thesis, University of West Bohemia, Pilsen, Czech Republic, 2003.

[Jelínek 04] L. Jelínek & L. Šmídl. *Spontaneous speech understanding in train timetable inquiry processing based on n-gram language models and finite state transducers.* In The 8th world multi-conference on systemics, cybernetics and informatics, volume VI, pages 444–449, Orlando, 2004. International Institute of Informatics and Systemics.

[Joshi 87] A. Joshi. *An introduction to tree adjoining grammars.* Mathematics of Language, 1987.

[Jurčíček 05] F. Jurčíček, J. Zahradil & L. Jelínek. *A human-human train timetable dialogue corpus.* In Proceedings of EUROSPEECH, Lisboa, Portugal, 2005.

[Jurčíček 07] F. Jurčíček. *Statistical approach to the semantic analysis of spoken dialogues.* PhD thesis, University of West Bohemia, Plzeň, 2007.

[Jurčíček 08] F. Jurčíček, J. Švec & L. Müller. *Extension of HVS semantic parser by allowing left-right branching.* In ICASSP 2008, International Conference on Acoustics, Speech, and Signal Processing, Las Vegas, USA, 2008.

[Kanis 06] J. Kanis, J. Zahradil, F. Jurčíček & L. Müller. *Czech-sign speech corpus for semantic based machine translation.* Lecture Notes in Artificial Intelligence, vol. 4188, pages 613–620, 2006.

[Kaplan 82] R. M. Kaplan & J. Bresnan. The mental representation of grammatical relations, chapitre Lexical-functional grammar: A formal system for grammatical representation, pages 173–281. Cambridge: The MIT Press, 1982.

[Klein 98] P. Klein. *Computing the Edit-Distance between Unrooted Ordered Trees.* In Bilardi, G. et al. (Ed.) Proceedings of the 6th Annual European Symposium, numéro 1461, pages 91–102, Venice, Italy, 1998. Springer-Verlag, Berlin.

[Knight 99] K. Knight. *Decoding Complexity in Word-Replacement Translation Models.* Computational Linguistics, vol. 25, no. 4, pages 607–615, December 1999.

[Koehn 03] P. Koehn, F.J. Och & D. Marcu. *Statistical Phrase-Based Translation.* In Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), page 127–133, Edmonton, Canada, May/June 2003.

[Koehn 04] P. Koehn. *Statistical Significance Tests for Machine Translation Evaluation.* In Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP), pages 388–395, Barcelona, Spain, July 2004.

[Koehn 07] P. Koehn, H. Hoang, A. Birch, Ch. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, Ch. Moran, R. Zens, Ch. Dyer, O. Bojar, A. Constantin & E. Herbst. *Moses: Open Source Toolkit for Statistical Machine Translation.* In Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic, June 2007.

[Kumar 04] S. Kumar & W. Byrne. *Minimum Bayes-Risk Decoding for Statistical Machine Translation.* In Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), pages 169–176, Boston, MA, May 2004.

[Lacoste-Julien 06] S. Lacoste-Julien, B. Taskar, D. Klein & M. I. Jordan. *Word alignment via quadratic assignment.* In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 112–119, New York, New York, 2006. Association for Computational Linguistics.

[Langer 04] J. Langer, V. Ptáček & K. Dvořák. *Znaková zásoba českého znakového jazyka.* Univerzita Palackého v Olomouci, 2004.

[Lavecchia 08] C. Lavecchia, D. Langlois & K. Smaïli. *Phrase-Based Machine Translation based on Simulated Annealing.* In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 2008. ELRA.

[Levenshtein 66] V. I. Levenshtein. *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, vol. 10, pages 707–710, 1966.

[Liang 06] P. Liang, B. Taskar & D. Klein. *Alignment by agreement.* In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, pages 104–111, New York, New York, 2006. Association for Computational Linguistics.

[Liddell 89] S. Liddell & R. Johnson. *American Sign Language: The Phonological Base.* Sign Language Studies, vol. 64, no. 3rd edition, pages 195–277, 1989. Washington, DC: Gallaudet University Press. 2000.

[Liu 05] Y. Liu, Q. Liu & S. Lin. *Log-linear models for word alignment.* In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 459–466, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[Lopez 08] A. Lopez. *Statistical Machine Translation.* Computing Surveys 40(3), pages 1–49, 2008.

[Macurová 96] A Macurová. *Proč a jak zapisovat znaky českého znakového jazyka.* Speciální pedagogika, vol. 6, no. 1, pages 5–19, 1996.

[Macurová 98] A. Macurová. *Naše řeč?* Naše řeč, vol. 81, pages 179–188, 1998.

[Macurová 01a] A. Macurová. *Poznáváme Český znakový jazyk.* Speciální pedagogika, vol. 11, no. 2, pages 69–75, 2001.

[Macurová 01b] A. Macurová & P. Bímová. *Poznáváme český znakový jazyk II. (Slovesa a jejich typy).* Speciální pedagogika, vol. 11, no. 5, pages 285–295, 2001.

[Macurová 03] A. Macurová. *Poznáváme ČZJ IV. (Vyjádření času).* Speciální pedagogika, vol. 13, no. 2, pages 89–98, 2003.

[Marcu 02] D. Marcu & W. Wong. *A Phrase-Based, Joint Probability Model for Statistical Machine Translation.* In Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP), pages 133–139, Philadelphia, PA, July 2002.

[Marshall 01] I. Marshall & E. Safar. *Extraction of Semantic Representations from Syntactic CMU Link Grammar linkages.* In G. Angelova et al., editeur, Recent Advances in Natural Language Processing (RANLP), pages 154–159, Tzigov Chark Bulgaria, Sept. 2001.

[Marshall 02] I. Marshall & E. Safar. *Sign Language Generation using HPSG.* In Proceedings of the 9th International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-2002, 2002.

[Matusov 06] E. Matusov, R. Zens, D. Vilar, A. Mauser, M. Popovic & H. Ney. *The RWTH Machine Translation System.* In Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, pages 31–36, Barcelona, Spain, 2006.

[Moore 05] R. C. Moore. *A discriminative framework for bilingual word alignment.* In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 81–88, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.

[Moore 06] R. C. Moore, W. Yih & A. Bode. *Improved discriminative bilingual word alignment.* In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 513–520, Sydney, Australia, 2006. Association for Computational Linguistics.

[Moore 07] R. C. Moore & Ch. Quirk. *An Iteratively-Trained Segmentation-Free Phrase Translation Model for Statistical Machine Translation.* In Proceedings of the Second Workshop on Statistical Machine Translation, pages 112–119, Prague, Czech Republic, June 2007.

[Morrissey 05] S. Morrissey & A. Way. *An Example-Based Approach to Translating Sign Language.* In 2nd International Workshop on Example-Based Machine Translation - At MT Summit X, 2005.

[Morrissey 07] S. Morrissey, A. Way, D. Stein, J. Bungeroth & H. Ney. *Combining Data-Driven MY Systems for Improved Sign Language Translation.* In Proceedings of the Machine Translation Summit XI Copenhagen, Denmark, 2007.

[Morrissey 08] S. Morrissey. *Data-Driven Machine Translation for Sign Languages*. PhD thesis, Dublin City University, Dublin, Ireland, 2008.

[Motejzíková 03] J. Motejzíková. *Poznáváme ČZJ V. (Specifické znaky)*. Speciální pedagogika, vol. 13, no. 3, pages 218–226, 2003.

[Nelder 65] J.A. Nelder & R. Mead. *A Simplex Method for Function Minimization*. The Computer Journal, vol. 7, pages 308–313, 1965.

[Och 99] F.J. Och, C. Tillmann & H. Ney. *Improved Alignment Models for Statistical Machine Translation*. In Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP), page 20–28, College Park, MD, June 1999.

[Och 02a] F.J. Och. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. PhD thesis, Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University, Aachen, Germany, October 2002.

[Och 02b] F.J. Och & H. Ney. *Discriminative Training and Maximum Entropy Models for Statistical Machine Translation*. In Proc. 40th Annual Meeting of the Assoc. for Computational Linguistics (ACL), pages 295–302, Philadelphia, PA, July 2002.

[Och 03a] F.J. Och. *Minimum Error Rate Training in Statistical Machine Translation*. In Proc. 41st Annual Meeting of the Assoc. for Computational Linguistics (ACL), Sapporo, Japan, July 2003.

[Och 03b] F.J. Och & H. Ney. *A Systematic Comparison of Various Statistical Alignment Models*. Computational Linguistics, vol. 29, no. 1, pages 19–51, March 2003.

[Papineni 98] K.A. Papineni, S. Roukos & R.T. Ward. *Maximum Likelihood and Discriminative Training of Direct Translation Models*. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), pages 189–192, Seattle, WA, May 1998.

[Papineni 01] K. Papineni, S. Roukos, T. Ward & W. Zhu. *Bleu: a method for automatic evaluation of machine translation*. Rapport technique, IBM Research Division, Thomas J. Watson Research Center, 2001.

[Pollard 94] C. Pollard & I.A. Sag. Head-driven phrase structure grammar. The University of Chicago Press, Chicago, 1994.

[Press 02] W.H. Press, S.A. Teukolsky, W.T. Vetterling & B.P. Flannery. Numerical recipes in c++. Cambridge University Press, Cambridge, UK, 2002.

[Prillwitz 89] S. Prillwitz, R. Leven, H. Zienert, T. Hanke & J. Henning. *Hamburg Notation System for Sign Languages - An Introductory Guide*. International Studies on Sign Language and the Communication of the Deaf, vol. 5, 1989.

[Psutka 04] J. Psutka, P. Ircing, J. Hajič, V. Radová, J. Psutka, W. Byrne & S. Gustman. *Issues in annotation of the Czech spontaneous speech corpus*

in the MALACH project. In Fourth international conference on language resources and evaluation (LREC04), pages 607–610, Lisbon, 2004. European Language Resources Association.

[Psutka 06] J. Psutka, L. Müller, J. Matoušek & V. Radová. Mluvíme s počítačem česky. Academia, Prague, 2006.

[Riezler 05] Stefan Riezler & J. T. Maxwell. *On Some Pitfalls in Automatic Evaluation and Significance Testing for MT*. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 57–64, Ann Arbor, Michigan, June 2005.

[Safar 02] E. Safar & I. Marshall. *Sign Language Translation via DRT and HPSG*. In A. Gelbukh, editeur, Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, CICLing, Lecture Notes in Computer Science 2276, Mexico, 17-23 February 2002. Springer Verlag.

[Shapiro 91] L. Shapiro & A.B. Stephens. *Boostrap Percolation, the Schröder Numbers, and the N-Kings Problem*. SIAM Journal on Discrete Mathematics, vol. 4, no. 2, pages 275–280, May 1991.

[Sleator 91] D. Sleator & D Temperley. *Parsing English with a Link Grammar*. Rapport technique, Carnegie Mellon University, 1991.

[Smith 06] D.A. Smith & J. Eisner. *Minimum Risk Annealing for Training Log-Linear Models*. In Proc. 21st Int. Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics (COLING/ACL), pages 787–794, Sydney, Australia, July 2006.

[Snover 06] M. Snover, B. Dorr, R. Schwartz, L. Micciulla & J. Makhoul. *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proc. Conf. of the Assoc. for Machine Translation in the Americas (AMTA), pages 223–231, Cambridge, MA, 2006.

[Speers 01] d'A.L Speers. *Representation of American Sign Language for Machine Translation*. PhD thesis, Department of Linguistics, Georgetown University, 2001.

[Stein 06] D. Stein, J. Bungeroth & H. Ney. *Morpho-Syntax Based Statistical Methods for Sign Language Translation*. In Conference of the European Association for Machine Translation (EAMT), pages 169–177, Oslo, Norway, June 2006.

[Stroppa 06] N. Stroppa & A. Way. *MaTrEx: DCU Machine Translation System for IWSLT 2006*. In Proceedings of the International Workshop on Spoken Language Translation, pages 31–36, Kyoto, Japan, 2006.

[Suszczanska 02] N. Suszczanska, P. Szmal & J. Francik. *Translating Polish Texts into Sign Language in the TGT System*. In 20th IASTED International Multi-Conference Applied Informatics AI 2002, pages 282–287, Innsbruck, Austria, 2002.

[Suszczańska 99] N. Suszczańska. *Computational Grammar of Syntax Groups*. Cybernetics and Systems Analysis, vol. 29(6), pages 166–175, 1999.

[Taskar 05] B. Taskar, S. Lacoste-Julien & D. Klein. *A discriminative matching approach to word alignment*. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 73–80, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics.

[Tillmann 97a] C. Tillmann & H. Ney. *Word trigger and the EM algorithm*. In Proceedings of the Conference on Computational Natural Language Learning, pages 117–124, Madrid, Spain, 1997.

[Tillmann 97b] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga & H. Sawaf. *Accelerated DP based Search for Statistical Translation*. In Proceedings of the 5 th European Conference on Speech Communication and Technology, pages 2667–2670, 1997.

[Tillmann 01] C. Tillmann. *Word Re-Ordering and Dynamic Programming based Search Algorithms for Statistical Machine Translation*. PhD thesis, Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University, Aachen, Germany, May 2001.

[Tillmann 03] CH. Tillmann & F. Xia. *A Phrase-based Unigram Model for Statistical Machine Translation*. In Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, 2003.

[Udupa 06] R. Udupa & H.K. Maji. *Computational Complexity of Statistical Machine Translation*. In Proc. 11th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL), pages 25–32, Trento, Italy, April 2006.

[Čmejrek 04] M. Čmejrek, J. Cuřín, J. Havelka, J. Hajič & V. Kuboň. *Prague Czech-English Dependency Treebank. Syntactically Annotated Resources for Machine Translation*. In Proceedings of LREC 2004, Lisbon, May 2004.

[Čmejrek 06] M. Čmejrek. *Using Dependency Tree Structure for Czech–English Machine Translation*. PhD thesis, ÚFAL, MFF UK, Prague, Czech Republic, 2006.

[Veale 98] T. Veale, A. Conway & B. Collins. *The challenges of cross-modal translation: English to sign language translation in the ZARDOZ system*. Machine Translation, vol. 13, pages 81–106, 1998.

[Venugopal 03] A. Venugopal, S. Vogel & A. Waibel. *Effective Phrase Extraction from Alignment Models*. In Proceedings of ACL 2003, Sapporo, Japan, 2003.

[Vilar 06] D. Vilar, M. Popović & H. Ney. *AER: Do we need to "improve" our alignments?* In Proc. Int. Workshop on Spoken Language Translation (IWSLT), pages 205–212, Kyoto, Japan, November 2006.

[Vogel 96] S. Vogel, H. Ney & C. Tillmann. *HMM-Based Word Alignment in Statistical Translation*. In 16th Int. Conf. on Computational Linguistics (COLING), pages 836–841, Copenhagen, Denmark, August 1996.

[Švec 07] J. Švec. *Sémantická analýza promluv systému NÁDRAŽÍ.* Diplomová práce, Západočeská univerzita v Plzni, 2007.

[Vysuček 04] P. Vysuček. *Poznáváme ČZJ VI. (Specifické znaky).* Speciální pedagogika, vol. 14, no. 1, pages 16–27, 2004.

[Walker 01] M. Walker & R. Passonneau. *DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems.* In In Proc. of ACL, 2001.

[Wu 95] D. Wu. *Stochastic Inversion Transduction Grammars, with Application to Segmentation, Bracketing, and Alignment of Parallel Corpora.* In Proc. 14th Int. Joint Conf. on Artificial Intelligence (IJCAI), pages 1328–1334, Montreal, Canada, August 1995.

[Wu 96] D. Wu. *A Polynomial-Time Algorithm for Statistical Machine Translation.* In Proc. 34th Annual Meeting of the Assoc. for Computational Linguistics (ACL), pages 152–158, Santa Cruz, CA, June 1996.

[Wu 97] D. Wu. *Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora.* Computational Linguistics, vol. 23, no. 3, pages 377–403, September 1997.

[Wu 98] D. Wu & H. Wong. *Machine Translation with a Stochastic Grammatical Channel.* In Proceedings of the 17th international conference on Computational linguistics, pages 1408–1415, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics.

[Yamada 01] K. Yamada & K. Knight. *A Syntax-based Statistical Translation Model.* In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 523–530, Toulouse, France, 2001. Association for Computational Linguistics.

[Zens 02] R. Zens, F.J. Och & H. Ney. *Phrase-Based Statistical Machine Translation.* In Proc. M. Jarke, J. Koehler, G. Lakemeyer, editors, 25th German Conf. on Artificial Intelligence (KI2002), volume 2479 of *Lecture Notes in Artificial Intelligence (LNAI)*, page 18–32, Aachen, Germany, September 2002. Springer Verlag.

[Zens 08] R. Zens. *Phrase-based Statistical Machine Translation: Models, Search, Training.* PhD thesis, Lehrstuhl für Informatik 6, Computer Science Department, RWTH Aachen University, Aachen, Germany, February 2008.

[Zettlemoyer 07] L. Zettlemoyer & R. C. Moore. *Selective Phrase Pair Extraction for Improved Statistical Machine Translation.* In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics, pages 209–212, Rochester, New York, 2007. Association for Computational Linguistics.

[Zhang 03] Y. Zhang, S. Vogel & A. Waibel. *Integrated phrase segmentation and alignment algorithm for statistical machine translation.* In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE'03), 2003.

[Zhang 04] Y. Zhang & S. Vogel. *Measuring Confidence Intervals for the Machine Translation Evaluation Metrics*. In Proc. Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI), pages 85–94, Baltimore, MD, October 2004.

[Zhao 00] Liwei Zhao, Karin Kipper, William Schuler, Christian Vogler, Norman I. Badler & Martha Palmer. *A Machine Translation System from English to American Sign Language*. In AMTA, pages 54–67, 2000.

[Zitouni 03] I. Zitouni, K. Smaïli & J.-P. Haton. *Statistical language modeling based on variable-length sequences*. Computer Speech and Language, vol. 17, pages 27–41, 2003.

Seznam publikovaných prací

Publikace v angličtině

1. KANIS, J., MÜLLER, L.: Advances in Czech – Signed Speech Translation. In Lecture Notes in Artificial Intelligence. 2009. (*přijato*)

2. KANIS, J., KRŇOUL, Z.: Interactive HamNoSys Notation Editor for Signed Speech Annotation. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Paris : ELRA, 2008. s. 88-93. ISBN 2-9517408-4-0.

3. KRŇOUL, Z., KANIS, J., ŽELEZNÝ, M., MÜLLER, L.: Czech Text-to-Sign Speech Synthesizer. In Lecture Notes in Computer Science. 2008, sv.4892, č.1, s.180-191, ISSN 0302-9743.

4. KANIS, J., MÜLLER, L.: Automatic Czech - Sign Speech Translation. In Lecture Notes in Artificial Intelligence. 2007, sv.4629, s.488-495, ISSN 0302-9743.

5. KANIS, J., ZAHRADIL, J., JURČÍČEK, F., MÜLLER, L.: Czech-Sign Speech Corpus for Semantic Based Machine Translation. In Lecture Notes in Artificial Intelligence. Berlin : Springer , 2006, sv.4188, s.613-620, ISSN 0302-9743.

6. KRŇOUL, Z., ŽELEZNÝ, M., MÜLLER, L., KANIS, J.: Training of Coarticulation Models Using Dominance Functions and Visual Unit Selection Methods for Audio-Visual Speech Synthesis. In Interspeech. Bonn : ISCA, 2006, roč.2006, č.1, s.585-588, ISSN 1990-9772.

7. KRŇOUL, Z., KANIS, J., ŽELEZNÝ, M., MÜLLER, L., CÍSAŘ, P.: 3D Symbol Base Translation and Synthesis of Czech Sign Speech. In Proceedings of the 11th international conference "Speech and computer"SPECOM'2006 . St.Petersburg : Anatolya Publisher, 2006. s. 530-535. ISBN 5-7452-0074-X.

8. PRAŽÁK, A., PSUTKA, J., HOIDEKR, J., KANIS, J., MÜLLER, L., PSUTKA, J.: Automatic Online Subtitling of the Czech Parliament Meetings. In Lecture Notes in Artificial Intelligence. Berlin : Springer, 2006, sv.4188, s.501-508, ISSN 0302-9743.

9. PRAŽÁK, A., PSUTKA, J., HOIDEKR, J., KANIS, J., MÜLLER, L., PSUTKA, J.: Adaptive Language Model in Automatic Online Subtitling. In Proceedings of the Second IASTED International Conference on Computational Intelligence. Anaheim : ACTA Press, 2006. s. 479-483. ISBN 0-88986-602-3.

10. KANIS, J., MÜLLER, L.: Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization. In Lecture Notes in Artificial Intelligence. Berlin : Springer, 2005, sv.3658, s.132-139, ISSN 0302-9743.

11. KANIS, J., MÜLLER, L.: Using Lemmatization Technique for Automatic Diacritics Restoration. In SPECOM 2005 Proceedings. Moscow : Moscow State Linguistic University, 2005. s. 255-258. ISBN 5-7452-0110-X.

12. KANIS, J., ZELINKA, J., MÜLLER, L.: Automatic Numbers Normalization in Inflectional Languages. In SPECOM 2005 proceedings. Moscow : Moscow State Linguistic University, 2005. s. 663-666. ISBN 5-7452-0110-X.

13. CÍSAŘ, P., ŽELEZNÝ, M., KRŇOUL, Z., KANIS, J., ZELINKA, J., MÜLLER, L.: Design and Recording of Czech Speech Corpus for Audio-Visual Continuous Speech Recognition. In Proceedings of the Auditory-Visual Speech Processing International Conference 2005. Vancouver Island : AVSP2005, 2005. s. 1-4. ISBN 1 876346 53 1.

14. ZELINKA, J., KANIS, J., MÜLLER, L.: Automatic Transcription of Numerals in Inflectional Languages. In Lecture Notes in Artificial Intelligence. Berlin : Springer, 2005, sv.3658, s.326-333, ISSN 0302-9743.

15. KANIS, J., MÜLLER, L.: Using the Lemmatization Technique for Phonetic Transcription in Text-to-Speech System. In Lecture Notes in Artificial Intelligence. Berlin : Springer, 2004, sv.3206, s.355-361, ISSN 0302-9743.

Publikace v češtině

1. KANIS, J. : Využití metod automatického zpracování přirozeného jazyka v TTS a TTSL systémech. Odborná práce ke státní doktorské zkoušce. Západočeská univerzita, Fakulta aplikovaných věd, Plzeň, 2005.

2. KANIS, J. : Použití lemmatizátoru v systému fonetické transkripce. Diplomová práce. Západočeská univerzita, Fakulta aplikovaných věd, Plzeň, 2003.

Citace[1]

Publikace číslo 4 byla citována v:

- CAMPR, P., HRÚZ, M., TROJANOVÁ, J.: Collection and Preprocessing of Czech Sign Language Corpus for Sign Language Recognition. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Paris : ELRA, 2008. s. 1-4. ISBN 2-9517408-4-0.

Publikace číslo 9 byla citována v:

- LJAŠEVSKAJA, O. N., SIČINAVA, D. V., KOBRICOV, B. P.: Avtomatizacija postroenija slovarja na materiale massiva neslovarnyx slovoform. Internet-matematika 2007 : Sb. Rabot učastnikov konkursa nauč. proektov po inform. poisku. Ekaterinburg : Izd-vo Ural. un-ta, 2007. s. 118–125. (*rusky*)

- LJAŠEVSKAJA, O. N.: K probleme lemmatizacii neslovarnyx slovoform (angl. Toward the Lemmatization of Word Forms Absent from the Dictionary). Trudy meždunarodnoj konferencii Dialog 2007

[1] Zdroj citací: scholar.google.com

Publikace číslo 11 byla citována v:

- MATOUŠEK, J., TIHELKA, D., ROMPORTL, J.: Current State of Czech Text-to-Speech System ARTIC. In Lecture Notes in Artificial Intelligence. Berlin : Springer, 2006, sv.4188, s.439-446, ISSN 0302-9743.

Publikace číslo 13 byla citována v:

- KOLÁŘ, J. Automatic Segmentation of Speech into Sentence-like Units: disertační práce. Plzeň: Západočeská univerzita v Plzni, 2008. xiv + 168 s.