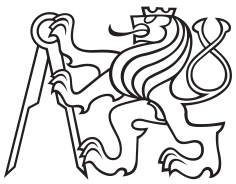
Diplomová práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra radioelektroniky

Jazykové modely pro multilingvální rozpoznávání spojité řeči

Bc. Jiří Valíček

Vedoucí: Doc. Ing. Petr Pollák, CSc.
Obor: Multimediální technika
Studijní program: Komunikace, multimédia a elektronika
Leden 2017

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Valíček** Jméno: **Jiří** Osobní číslo: **383020**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra radioelektroniky**

Studijní program: **Komunikace, multimédia a elektronika**

Studijní obor: **Multimediální technika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Jazykové modely pro multilingvální rozpoznávání spojité řeči

Název diplomové práce anglicky:

Language Models for Multilingual Continuous Speech Recognition

Pokyny pro vypracování:

1. Seznamte se s problematikou rozpoznávání spojité řeči s velkým slovníkem s užším zaměřením na jazykové modelování.
2. Vytvořete n-gramové jazykové modely použitelné v multilingválním systému rozpoznávání spojité řeči pro vybrané jazyky dostupné v databázích SpeechDat(E) a GlobalPhone. Korpusy pro tvorbu jazykových modelů hledejte ve veřejně dostupných zdrojích pro zvolené jazyky.
3. Navrhněte proceduru tvorby výslovnostního slovníku pro vytvořené jazykové modely s užším zaměřením na definici výslovnosti pro nová neznámá slova v jednotlivých jazycích.
4. Vytvořené jazykové modely a výslovnostní slovníky otestujte na testovacích textových korpusech a v dostupném rozpoznávači spojité řeči.

Seznam doporučené literatury:

[1] J. Psutka, L. Miller, J. Matousek, V. Radova: Mluvíme s počítačem česky. Academia, 2006.
[2] X. Huang, A. Acero, H.-W. Hon: Spoken Language Processing. Prentice-Hall 2001.
[3] J. Fiala: DNN-HMM Based Multilingual Recognizer of Telephone Speech. Diploma thesis. CTU FEE. Prague, 2016.
[4] D. Povey et al, The Kaldi Speech Recognition Toolkit. In Proc. of IEEE 2011 ASRU, Hawai, US, 2011. Note. Project WEB-page http://kaldi.sourceforge.net/

Jméno a pracoviště vedoucí(ho) diplomové práce:

doc. Ing. Petr Pollák CSc., katedra teorie obvodů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **21.09.2016** Termín odevzdání diplomové práce: **09.01.2017**

Platnost zadání diplomové práce: **20.02.2018**

Podpis vedoucí(ho) práce	Podpis vedoucí(ho) ústavu/katedry	Podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. | Datum převzetí zadání

Podpis studenta

Poděkování

Chtěl bych poděkovat Doc. Ing. Petru Pollákovi, CSc. za vedení práce a projevenou trpělivost a také Ing. Jiřímu Fialovi za pomoc s rozpoznávačem řeči.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 09. ledna 2017

.....

Jiří Valíček

Abstrakt

Tato práce se zabývá jazykovým modelováním v multilingválním systému rozpoznávání spojité řeči. Primárním úkolem je vytvoření n-gramových jazykových modelů z volně dostupných textových korpusů a návrh postupu pro vytvoření výslovnostních slovníků se zaměřením na definici výslovností pro neznámá slova. Zadaný úkol je vypracován pro čtyři jazyky: polština, slovenština, ruština a maďarština. Pro každý tento jazyk bylo nalezeno několik veřejně dostupných textových korpusů, ze kterých byly, pomocí balíku SRILM, vytvořeny jazykové modely. Výslovnostní slovníky byly tvořeny třemi nástroji. g2p-sk pro slovenštinu a univerzální BAS G2P a Sequitur G2P. V případě nástroje Sequitur G2P byly navíc natrénovány modely výslovnostních pravidel ze slovníků s ověřenou výslovností. Výstupem této práce je metodika, realizující celý proces zpracování textových korpusů, a její implementace, s možností rozšíření o další jazyky. Vytvořeny byly jazykové modely, výslovnostní slovníky a modely pravidel výslovností. Modely a slovníky jsou testovány v LVCSR systému a na textových korpusech. V rozpoznávání řeči byla dosažena chybovost v rozmezí 13%-41% WER v závislosti na jazyce. Při porovnání nástrojů pro tvorbu slovníku, dosáhl Sequitur G2P lepších výsledků než BAS G2P.

Klíčová slova: rozpoznávání spojité řeči, multilingvální systém, jazykový model, ngram, textový korpus, výslovnostní slovník, grafém, foném

Vedoucí: Doc. Ing. Petr Pollák, CSc.
Fakulta elektrotechnická,
Technická 2,
166 27 Praha 6

Abstract

This thesis deals with language modelling for multilingual continuous speech recognition system. The primary objective of this thesis was to create n-gram language models using freely available resources and design a procedure to create pronunciation dictionaries with focus on new words transcription. Given task was performed on four languages: Polish, Slovak, Russian and Hungarian. For each of these languages several free resources of text corpora were found. Language model creation was done using SRILM toolkit. To create pronunciation dictionaries three tools were used: g2p-sk for Slovak language and multilingual BAS G2P and Sequitur G2P. While testing the Sequitur G2P pronunciation models were trained on verified dictionaries. The output of this thesis is method of text corpora processing and implementation of said method. Created were language models, pronunciation dictionaries and pronunciation models. Models and dictionaries are tested in LVCSR system and on text corpora. Depending on the language, obtained results were in range 13%-41% WER. In dictionary comparison Sequitur G2P performed better than BAS G2P.

Keywords: continuous speech recognition, multilanguage system, language model, ngram, text corpus, pronunciation dictionary, grapheme, phoneme

Title translation: Language Models for Multilingual Continuous Speech Recognition

Obsah

Obrázky

Tabulky

Kapitola 1

Úvod

Díky technickému pokroku posledních let pronikají do společnosti technologie, masově známé především z děl vědecké fikce. Jedná se o zařízení, jejichž představení proběhlo před mnoha lety ve filmech jako je například 2001: Vesmírná odysea nebo Star Trek. Představa autorů těchto děl často pracuje s umělou inteligencí ve formě počítače nebo robota. Pro komunikaci používají tyto umělé inteligence ve většině případů řeč, ta je nejsnazší a nejběžnější formou lidské komunikace.

Z pohledu dnešní techniky je možné tuto komunikaci rozdělit do dvou úloh, syntéza řeči a rozpoznávání řeči. Rozpoznávání řeči můžeme dále rozdělit na získání informace z proneseného textu a na pochopení informace. Je zřejmé, že před rozpoznáváním smyslu promluvy a případným vykonáním příkazu, je nutné získat vyřčenou informaci, například v textové podobě.

Systémy rozpoznávání řeči jsou realizovány již od druhé poloviny minulého století. V prvních pokusech se jednalo o rozpoznávání mezi několika jednotlivými slovy. Postupem času a techniky byla úloha ztěžována zvětšováním rozpoznávaného slovníku a přirozeností promluv. To znamená, že se rozpoznávání přesunulo z jednotlivých slov do vět a neohraničených promluv. Rozpoznávání takových promluv je nejobecnější úloha. V těchto promluvách se smazávají mezery mezi slovy, nemusí být jasné, kde se nachází konec věty a často se v nich objevují i tzv. neřečové události. Takto zobecněnou úlohu nazýváme rozpoznávání spojité řeči s velkým slovníkem.

Rozpoznávání řeči je úloha závislá na jazyce a v případě vytvoření rozpoznávače pro nový jazyk, musíme získat značné množství dat pro daný jazyk. Tento problém se snaží zjednodušit multilingvální systémy. Ty dokáží pracovat s akustickými daty více jazyků, a rozšíří tak trénovací množinu

dat na základě podobnosti mezi jazyky. Podobnosti se hledají na úrovni subslovních jednotek, fonémů. V rozpoznávání řeči pracuje s fonémy akustický model, který modeluje zvukovou podobu jazyka[9].

Další částí systému rozpoznávání řeči je model jazykový. Ten slouží k pokrytí pravidel tvorby vět a definici slovníku daného jazyka. Oba modely je nutné vytvořit před samotným rozpoznáváním. Na rozdíl od akustického modelování není možné v jazykovém modelu použít data jiných jazyků. Z pohledu jazykového modelování, nepřináší multilingvální systém zjednodušení při sběru dat nebo trénování. Pro vytvoření jazykového modelu musíme použít rozsáhlý textový korpus v daném jazyce. Korpus musí být dále zpracován, jedním ze způsobů zpracování je n-gramové modelování. N-gramové modely jsou soubory pravděpodobností výskytu posloupností slov.

Při použití jazykového modelu v systému rozpoznávání řeči je nutné předložit také výslovnostní slovník všech slov použitých v modelu. Slovník poté slouží k převodu informace mezi akustickým a jazykovým modelem. Úkolem této práce je vytvoření a otestování n-gramových jazykových modelů a připravit jejich výslovnostní slovníky pro různé jazyky.

Práce je členěna do šesti kapitol. Ve druhé kapitole je popsána teoretická část této práce. Nachází se zde popis rozpoznávání řeči, akustické analýzy, akustického modelu, jazykového modelu a kritérií hodnocení. V následující kapitole jsou uvedeny principy aplikace jazykového modelu v multilingválním systému a schéma postupu této práce včetně detailnějšího popisu jednotlivých částí. Ve čtvrté kapitole se nalézají popisy nástrojů, korpusů, vytvořeného algoritmu a konkrétní postupy implementace. V páté kapitole jsou předvedeny všechny vytvořené jazykové modely a jsou diskutovány dosažené výsledky. V poslední kapitole je shrnutí celé práce.

Kapitola 2

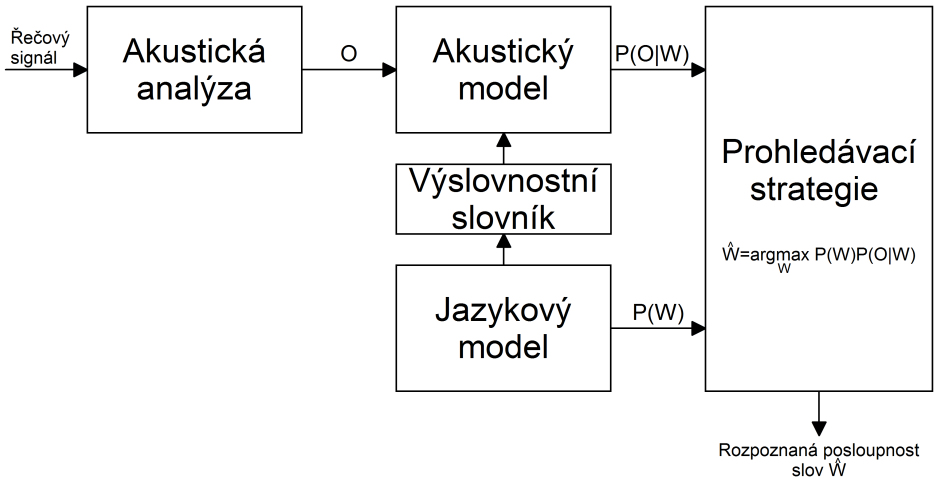
Rozpoznávání spojité řeči

Princip systému pro rozpoznávání řeči, který byl použit při prvních pokusech ve druhé polovině minulého století, má svoje místo i v konkrétních aplikacích dnešního světa. Například při jednoduchém rozpoznávání několika příkazů pro ovládání jednoúčelových zařízení je stále možné použít tyto metody, které jsou založené na porovnávání se vzory. Důležitější úloha je ovšem rozpoznávání spojitých promluv, které používají rozsáhlé slovníky. Takovýto úkol již není možné realizovat stejným systémem, a to z důvodu výpočetní náročnosti. Důvodem nynějších úspěchů systémů pro rozpoznávání řeči je použití statistických metod při trénování akustického i jazykového modelu.

Rozpoznávání řeči lze rozložit do následujících úkolů

1. Akustická analýza
2. Natrénování akustického modelu
3. Vytvoření jazykového modelu
4. Vyhledání nejpravděpodobnější posloupnosti slov

Zmíněné úkoly se poté aplikují při rozpoznávání řeči ve smyslu ilustrovaném blokovým schématem.



Obrázek 2.1: Blokové schéma systému rozpoznávaní řeči

Prohledávací strategie neboli dekódování, má za úkol vyhledat takovou posloupnost slov $\hat{W}$, která maximalizuje podmíněnou pravděpodobnost $P(W|O)$. Pokud na vstupu je vektor příznaků z řečového signálu O

$$O = \{o_1 o_2 \dots o_T\} \tag{2.1}$$

a byla řečena promluva W obsahující N slov

$$W = \{w_1 w_2 \dots w_N\}. \tag{2.2}$$

Můžeme poté s použitím Bayesova pravidla psát

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|O) = \underset{W}{\operatorname{argmax}}\, \frac{P(W)P(O|W)}{P(O)}, \tag{2.3}$$

kde $P(O|W)$ je pravděpodobnost výskytu vektoru příznaků O, pokud byla vyslovena posloupnost W, tato pravděpodobnost je v systému rozpoznávání řeči charakterizována akustickým modelem. $P(W)$ označuje model jazykový. Pravděpodobnost $P(O)$ není funkcí W, z toho důvodu můžeme použít zápis

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W, O) = \underset{W}{\operatorname{argmax}}\, P(W)P(O|W). \tag{2.4}$$

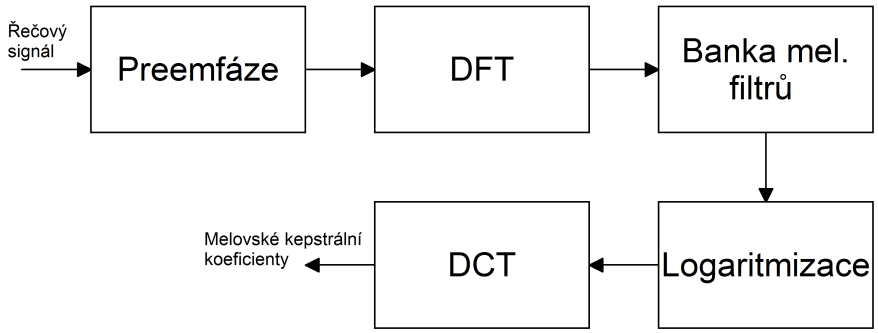
Z této rovnice vyplývá, že odhad posloupnosti slov $\hat{W}$ je závislý pouze na pravděpodobnostech $P(W)$ a $P(O|W)$, které jsou na sobě nezávislé a odpovídají jazykovému a akustickému modelu. Je tedy možné připravit modely nezávisle na sobě[23].

2.1 Akustická analýza

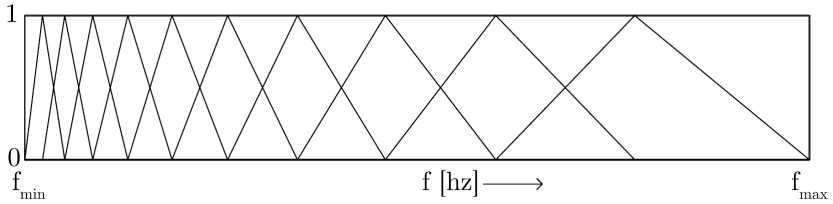
Jak bylo uvedeno dříve, první úkol v rozpoznávači řeči je analýza řečového signálu. Jako předpoklad při zpracování řeči se bere pomalá změna řečového signálu, kdy považujeme řečové úseky o délce zhruba 20ms za stacionární.

Samotná analýza se skládá z filtru dolní propusti, pro potlačení vyšších frekvencí. Dále z analogově-digitálního převodu signálu, po kterém mohou být aplikovány metody pro zlepšení parametrů signálu, což může být například odstranění šumu nebo kompenzace poklesu energie na vyšších frekvencích. Takto upravený signál je segmentován dle zmíněného předpokladu o stacionaritě. Při realizaci segmentace se používají váhovací okna s překryvem. V praxi se často používá Hammingovo okno s překryvem 50%.

Hlavním úkolem analýzy je převod řečových segmentů na vektory příznaků. To se provádí pomocí metod krátkodobé spektrální analýzy. Nejjednodušší metodou je převod na spektrální koeficienty pomocí Fourierovy transformace. Tato metoda ovšem nepřináší znatelné snížení redundance oproti vzorkům signálu. Při realizaci se nejčastěji používá melovských kepstrálních koeficientů (MFCC) nebo perceptivní lineární prediktivní analýza (PLP)[4].



Obrázek 2.2: Blokové schéma výpočtu MFCC



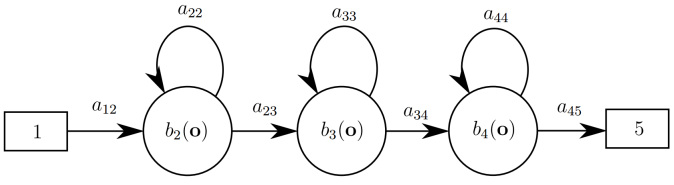
Obrázek 2.3: Banka melovských filtrů Zdroj:[9]

Při výpočtu mel-frekvenčních kepstrálních koeficientů se uvažuje nelinearita vnímání zvuků lidských uchem. Konkrétně dochází ke kompenzaci logaritmického vnímání frekvencí. Toho je dosaženo převodem frekvencí do melovské frekvenční škály

Vstupem při výpočtu MFCC je segmentovaný řečový signál, který je dále upraven preemfází a převeden do amplitudového spektra. K této krátkodobé Fourierově transformaci se nejčastěji používá FFT. Spektrum je poté filtrováno bankou pásmových filtrů rozložených rovnoměrně po frekvenční ose v melovské škále. Výstup filtrů je dále logaritmován, tím je omezena dynamika signálu ve stejném smyslu jako u lidského sluchu. Konečným krokem je diskrétní kosinová transformace[23].

2.2 Akustický model

Akustický model poskytuje odhad pravděpodobnosti $P(O|W)$, což lze vyjádřit jako pravděpodobnost pozorování vektoru O, pokud byla řečena posloupnost slov W. Při sestavování akustického modelu je nutné mít v paměti reálné podmínky použití. Při trénování mohou být k dispozici data nahraná čtením ve studiu a řešenou úlohou rozpoznávání řeči řidiče v jedoucím voze. Vidíme, že parametry prostředí i samotné promluvy budou velmi rozdílné, proto je potřeba jisté flexibility modelu. Dále je požadováno, aby byl model přesný, dokázal rozlišit mezi podobnými slovy s různým významem. Posledním parametrem je účinnost, kdy musí model vracet výsledky v přiměřeném čase. Standardem v dnešních systémech jsou skryté Markovovy modely (Hidden Markov Model - HMM).



Obrázek 2.4: Schéma levopravého pětistavového modelu Zdroj:[3]

HMM jsou pravděpodobnostní konečné automaty, generující posloupnost vektorů pozorování $O = \{o_1 o_2 \dots o_T\}$, skrytou posloupností stavů. Tento proces se podobá realizace řeči u lidí. Je možné slyšet řeč, ale stavy vedoucí na tento výsledek pozorovat možné není. V každém časovém kroku je při změně stavu modelu z s_i na s_j použit soubor předem daných pravděpodobností a_{ij} a je podle výstupní pravděpodobnosti $b_j(o_j)$ generován vektor pozorování o_t.

Struktura HMM dovoluje modelování celých slov nebo také jednotek, ze kterých se slova skládají. Tím máme na mysli monofony, difony a trifony. V případě trénování HMM je potřeba nahrávek několika realizací pro natrénování každé jednotky, což jeden z důvodů, proč se v při realizacích volí subslovní jednotky. Nejčastěji se pracuje s trifony, které poskytují kontext před i za fonémem[3].

2.3 Jazykový model

Poslední částí podílející se na rozpoznávání řeči je jazykový model. Ten slouží k charakteristice rozpoznávaných promluv, specifikuje použitou řeč. Řeč můžeme určit z pohledu použitého jazyka, například český nebo anglický jazykový model. Můžeme se také stejně jako u akustického modelu setkat s modelem specifickým pro mluvčího nebo s modelem obecným. Jazykový model se také může lišit v rozpoznávání různých témat, protože slovní zásoba použitá například na lékařské přednášce se bude lišit od zásoby použité při komentování sportovního utkání. Dle těchto parametrů vytváříme jazykové modely, aby co nejpřesněji odpovídaly myšlenému použití.

Základním modelem je obecný jazykový model, který by měl charakterizovat pravidla skládání vět v daném jazyce a specifikovat použitý slovník. Existují specifické situace, které mění pravděpodobnost použití slovních spojení. Pokud je naše úloha velmi odlišná od obecného rozpoznávání v daném jazyce, je vhodné upravit obecný model nebo vytvořit jiný model, který bude přesněji postihovat dané promluvy. Takový model již není obecný, ale jedná se o model tematicky specifický. V této práci se budeme zabývat vytvářením obecných jazykových modelů.

Jazykový model určující apriorní pravděpodobnost $P(W)$ všem posloupnostem slov je nazývaný stochastický jazykový model. Pravděpodobnost $P(W)$ posloupnosti W, čítající K slov je určena jako

$$\begin{aligned} P(W) = P(w_1^K) &= P(w_1 w_2 w_3 \dots w_K) = \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)\dots P(w_K|w_1 w_2 \dots w_K - 1) = \\ &= P(w_1)P(w_2|w_1^1)P(w_3|w_1^2)\dots P(w_K|w_1^{K-1}) = \\ &= \prod_{i=1}^{K} P(w_i|w_1^{i-1}) \end{aligned} \tag{2.5}$$

a v případě části této posloupnosti $w_1 w_2 ... w_k (k < K)$, platí

$$\begin{aligned} P(w_1^k) &= P(w_1^{k-1})P(w_k|w_1^{k-1}) = \\ &= P(w_1)P(w_2|w_1^1)P(w_3)(w_1^2)\dots P(w_k|w_1^{k-1}), \\ k &= 2, \dots, K. \end{aligned} \tag{2.6}$$

Při tomto rozkladu vidíme, že pravděpodobnost výskytu slova w_i je podmíněna pouze historií posloupností slov $w_1 \dots w_{i-2} w_{i-1}$. Tento rozklad je vhodný pro implementaci. Ovšem vypočítávat apriorní pravděpodobnosti $P(w_1^k)$ všech posloupností délky K by bylo náročné. Proto se v praxi používá aproximace, při níž se historie zkracuje na posledních n slov.

Modely získané touto aproximací se nazývají n-gramové modely. N-gramem nazýváme posloupnost n slov získaných z trénovacího korpusu. Nejčastěji se v praktické aplikaci setkáváme s unigramy, bigramy a trigramy. Jsou to posloupnosti o délce jedno, dvě a tři slova. Pro postihnutí všech závislostí jazyka by bylo vhodnější použít n-gramy s $n > 3$, avšak takové modely jsou velmi rozsáhlé a práce s nimi je složitá. Z principu tvorby n-gramových modelů vyplývá, že je jejich použití méně vhodné pro jazyky, které mají volnější pravidla pro skládání vět. Tento fakt je však kompenzován především snadným způsobem vytváření. Z tohoto důvodu se tyto modely standardně používají i pro takové jazyky a budou použity v této práci.

Při použití n-gramového modelu je podmíněná pravděpodobnost $P(w_k|w_1^{k-1})$ slova w_k závislá na historii o délce $n-1$, tzn. můžeme aproximovat $P(w_k|w_1^{k-1}) \approx P(w_k|w_{k-n+1}^{k-1})$ a platí, že

$$P(w_1^k) \approx \prod_{i=1}^{k} P(w_i|w_{k-i+1}^{i-1}). \tag{2.7}$$

Pro odhad pravděpodobnosti se používá relativní četnost výskytu v trénovacím korpusu. Pravděpodobnost $P(w_k|w_{k-2}w_{k-1})$ můžeme odhadnout v případě trigramového modelu jako

$$\bar{P}(w_k|w_{k-2}w_{k-1}) = \frac{N(w_{k-2}, w_{k-1}, w_k)}{N(w_{k-2}, w_{k-1})}, \tag{2.8}$$

kde $N(w_{k-2}, w_{k-1}, w_k)$ je četnost trigramu w_{k-2}, w_{k-1}, w_k a $N(w_{k-2}, w_{k-1})$ je četnost bigramu w_{k-2}, w_{k-1} v trénovacích datech[23].

2.4 Hodnocení jazykových modelů

Vytvořené n-gramové jazykové modely a jejich slovníky se mohou lišit v mnoha parametrech. Základním parametrem n-gramových modelů je stupeň modelu. Dalším je velikost, kterou lze chápat ve smyslu počtu unigramů nebo celkovou velikost. Shodně nastavené parametry můžeme také aplikovat na jiný korpus stejného jazyka, a získat tak jiný model. Při použití jazykového modelu v rozpoznávači je i jeho výslovnostní slovník. Slovník může být vytvořen různými způsoby, a tím může dojít ke změně úspěšnosti rozpoznávání.

Nevýhodou srovnávání párů model-slovník v rozpoznávání je časová náročnost a nutnost dostupného rozpoznávače. Jazykové modely je možné porovnávat i odděleně, a to předpovědí slov v neznámém textu. Standardem tohoto hodnocení je perplexita[23].

Perplexitu vyjadřujeme následovně

$$PP = \frac{1}{\sqrt[K]{\overline{P}(w_1 w_2 \dots w_K)}}, \tag{2.9}$$

kde jmenovatelem je odhad apriorní pravděpodobnosti posloupnosti slov W, o délce K, normalizovaná vzhledem k počtu slov.

Význam rozměru perplexity je průměrný počet slov, mezi kterými se akustický model rozhoduje v rozpoznávání, použijeme-li daný jazykový model. Perplexitou je možné porovnávat dva různé jazykové modely na jednom korpusu nebo obráceně dva korpusy s použitím jednoho modelu. Při srovnávání modelů pomocí perplexity, jsou podstatné až velké změny. Změny v řádu

jednotek procent nepřináší stejný rozdíl při rozpoznávání, proto nelze stavět perplexitu na úroveň WER[23].

Úspěšnost rozpoznávání je tedy nejdůležitější parametr pro srovnání párů model-slovník. Nejčastěji se uvádí jako WER(Word Error Rate)[25], alternativně jako $ACC = 1 - WER$, a zahrnuje v sobě chyby vznikající náhradou slov S, odstraněním slov D a vložením slov I. WER lze vyjádřit jako

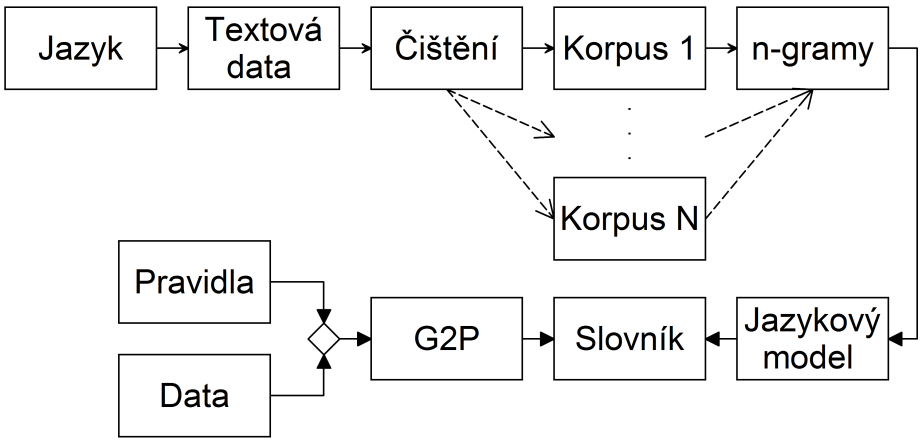
$$WER = 100 \times \frac{S + D + I}{N}[\%]. \tag{2.10}$$

Kapitola 3

Jazykový model v multilingválním systému

Jedním z úkolů této práce je vytvoření jazykových modelů použitelných v multilingválním rozpoznávači řeči. S ohledem na tento požadavek byla navržena metodika zpracování textových korpusů tak, aby byla možná implementace pro více jazyků. V této práci jsou implementovány čtyři jazyky, ale vytvořené skripty je možné dle daného postupu rozšířit o další.

Metodika realizovaná v této práci je ilustrována následujícím blokovým schématem.



Obrázek 3.1: Blokové schéma postupu práce

3.1 Korpus a jeho čištění

K vytvoření jazykového modelu je nutné získat značné množství textu, který bude co nejlépe odpovídat mluvené podobě daného jazyka. Z tohoto pohledu jsou ideální ručně psané přepisy promluv. Získání těchto přepisů je ovšem složité. Zvláště pak pro jazyky, jež jsou minoritní z pohledu počtu mluvčích. Většina zdrojů takto připravených dat podléhá licencím, které nedovolují jejich volné šíření a použití. Navíc i tyto zdroje jsou omezené svou velikostí a k postihnutí celého jazyka by nestačily.

Řešením tohoto problému jsou volně dostupné korpusy vytvořené ze zdrojů nepodléhajícím placeným licencím. Jedná se většinou o novinové články, knihy, internetové stránky atd. Výhodou těchto zdrojů je, kromě dostupnosti, také rozsah v řádu mnoha milionů slov. Vzhledem k původu těchto textů je nutné počítat s přítomností nežádoucích dat. Za takové lze považovat přítomnost slov a vět v jiných jazycích, nealfanumerické znaky, nesmyslná slova a podobné. Protože výsledkem zpracování těchto dat je jazykový model s velikostí omezenou na nejčastější slova, řádově stovky tisíc unigramů, uvažuje se, že tato slova budou ve výrazné menšině oproti platným slovům. Přesto je ovšem nutné vstupní korpusy upravit a nevhodná slova a věty z nich odstranit. Dalším problémem jsou číslice, nejlepším řešením by bylo jejich přepsání do slovní podoby. Pokud bychom je pouze odstranili, vznikla by chyba ve slovosledu, která by se v závislosti na četnosti mohla silně projevit. Pokud nechceme řešit významy problematických částí textu, je vhodné odstranit celou větu. Výsledkem zpracování korpusu složeného z vět je soubor obsahující jednu větu na řádek, bez interpunkce a pouze v malých písmenech abecedy daného jazyka. Některé zdroje jsou dostupné ve formě již napočítaných n-gramů ze zdrojových textů. V takovém případě je nutné použít pouze takové n-gramy, které obsahují pouze znaky dané abecedy.

Při práci s texty různých jazyků je potřeba brát ohled na použitou znakovou sadu těchto textů, ale také na znakovou sadu užívanou nástroji při jejich zpracovávání. V unixových systémech přejímá velké množství nástrojů informaci o kódování z `locale`, kde jsou uložené informace o jazyku nebo také právě o kódové sadě. Není to však pravidlem, a pro některé nástroje je nutné zvolit znakovou sadu zvlášť nebo specifikovat přejmutí znakové sady z locale ručně.

Při čištění textu použitím obecných pravidel lze odstranit velkou většinu neplatných slov. Některá další neplatná slova se ovšem mohou dále objevovat. Mohou to být například gramatické chyby, překlepy nebo slova vzniklá z posloupnosti slov vynecháním mezery. Dalšími slovy, která není možné snadno vyfiltrovat, jsou slova odpovídající skládání slov, ale bez reálného významu. Pro češtinu by příkladem takového slova mohlo být slovo *lovalín*. U všech

těchto neplatných slov se předpokládá, že jejich četnost je oproti platným slovům malá, a proto se jich ve velikostně omezených modelech objeví pouze minimum a s nízkou pravděpodobností. Další snahy o jejich odstranění nejsou vzhledem k poměru náročnosti a přínosu relevantní.

V případě, že objem jednoho textového korpusu není dostatečný, je možné spojit více korpusů před vytvořením výsledného jazykového modelu.

3.2 Výslovnostní slovník

V rozpoznávání řeči slouží výslovnostní slovník ke spojení akustického a jazykového modelu. Multilingvální rozpoznávač se od jednojazyčného systému liší v přístupu k fonémům v trénování akustického modelu. Aby bylo možné provést srovnání fonémů mezi jazyky, je nutné sjednotit používaný zápis výslovnosti fonémů ve výslovnostním slovníku. Což je soubor alespoň dvojicí slov, zapsaných na jednotlivé řádky. Kde první je grafický zápis slova a další jsou jeho fonetické zápisy. Více fonetických zápisů slova odpovídá různým možnostem výslovnosti. Takto vytvořený slovník slouží k přenosu informace mezi akustickým a jazykovým modelem při dekódování. Seznam slov ve výslovnostním slovníku musí pokrývat všechna slova obsažená v jazykovém modelu. Pro fonetický zápis se používají fonetické abecedy.

V práci [9] byla vytvořena pravidla, která slouží k převodu fonémů do jednotné abecedy XSAMPA. Poté byly identifikovány fonémy, které jsou společné pro dva nebo více jazyků a fonémy, a které jsou unikátní pro jednotlivé jazyky. Touto analýzou bylo zjištěno, že v průměru je foném sdílen dvěma jazyky. Díky tomuto zjištění je možné rozšířit trénovací množinu pro konkrétní fonémy o jejich výskyty z jiných jazyků. Bylo zjištěno, že polský jazyk je možné nejlépe pokrýt ze společných fonémů. Naproti tomu maďarský jazyk má nejvíce unikátních fonémů. Tato pravidla byla aplikována i na výslovnostní slovníky vytvořené v této práci.

```
disciplína      d i s ts i p l i: n a
školstve        S k o l s t v e
vyhlásený       v i h l a: s e n i:
tradičného      t r a J\ i tS J e: h o
súhlasom        s u: h l a s o m
obrovskej       o b r o u_^ s k e i_^
slovenky        s l o v e N k i
```

Obrázek 3.2: Příklad slovenského výslovnostního slovníku

3.2.1 Vytvoření výslovnostních slovníků

Prvním krok k vytvoření výslovnostního slovníku je získání seznamu unigramů použitých v jazykovém modelu nebo v trénovacím korpusu. K takovému seznamu je vytvořen fonetický přepis jednotlivých slov. Pro vytvoření fonetických přepisů je možné principiálně použít tři přístupy.

V ideálním případě by měl být nejprve seznam slov porovnán s manuálně vytvořeným slovníkem, který obsahuje všechna slova daného jazyka a jejich výslovnosti. Protože takový slovník není obecně k dispozici, můžeme použít slovníky jiných modelů daného jazyka a shodným slovům přiřadit fonetický zápis z nich. Na slova neznámá poté použijeme následující způsoby.

1. Manuální přepis

   Při manuálním přepisu by byla zbylá slova po prvním kroku ručně zkontrolována a jejich přepisy zapsány. Tento přístup je obecně nejsprávnější, ale nejnáročnější na čas a peníze.

2. Výslovnostní pravidla příslušného jazyka

   Pro některé jazyky je možné použít výslovnostní pravidla a neznámým slovům přiřadit výslovnost na základě posloupnosti grafémů. Nevýhodou tohoto způsobu může být náročnost pravidel daného jazyka a výjimky. Výjimky je nutné postihnout ve slovníku.

3. Výslovnost dle podobnosti (Pronunciation by analogy[19])

   U tohoto přístupu je snaha o napodobení člověka, kdy na základě znalosti výslovnosti podobných slov, dokáže systém odhadnout výslovnost slov neznámých. Jedná se o modelování výslovností na základě trénovacích dat. Tato data získáme předchozími dvěma způsoby přístupy.

Výběr konkrétního způsobu přepisu je závislý na daném jazyce, velikosti slovníku i na množství dat, která již máme. Pro některé jazyky je možné poměrně snadno definovat výslovnostní pravidla, jež postihnou velkou většinu slov. Pokud je ovšem k dispozici pro daný jazyk ověřený slovník o dostatečné velikosti, můžeme jej použít k natrénování modelu výslovnosti[19].

3.3 Fonetické abecedy

Jedná se o fonetické inventáře, dovolující vytvoření přesného zápisu výslovnosti promluvy v daném jazyce. Pokud máme univerzální abecedu je možné

zapsat výslovnost jakéhokoliv jazyka, případně poté provést přesné srovnání výslovností napříč jazyky. V příloze této práce jsou uvedeny tabulky s fonetickými abecedami jednotlivých jazyků zpracovávaných v této práci.

3.3.1 International Phonetic Alphabet (IPA)

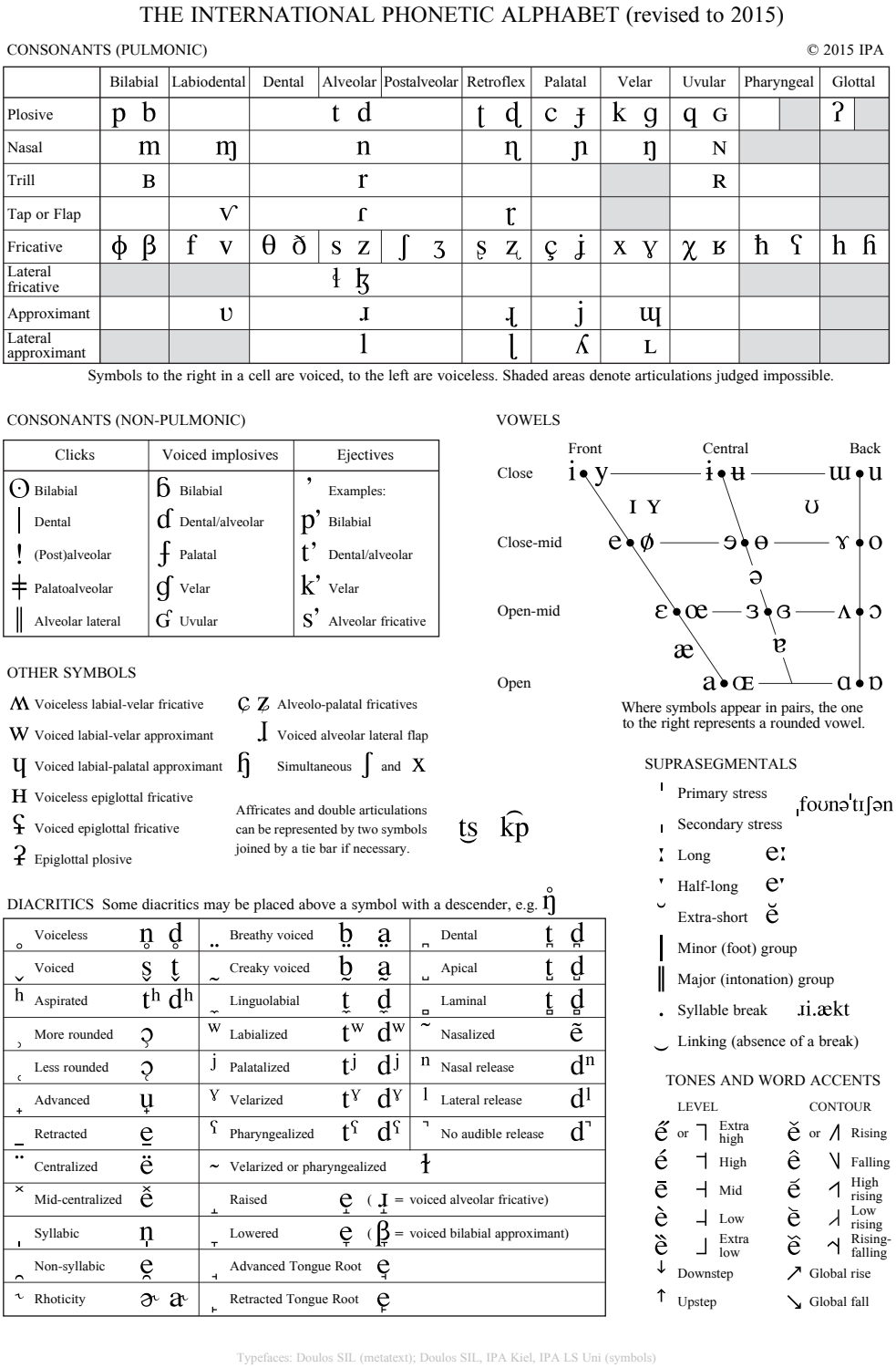
IPA byla vytvořena mezinárodní fonetickou asociací v roce 1888 a naposledy byla aktualizována v roce 2015. IPA je standardem v zápisu výslovností a obsahuje symboly pro zápis všech zvuků, všech jazyků na světe. K zápisu se používají písmena latinky, řečtiny a speciální symboly.

3.3.2 Speech Assessment Methods Phonetic Alphabet (SAMPA)

Jednalo se o abecedu vyvíjenou nejprve mezi lety 1987-1989 jen pro šest evropských jazyků (Dánština, Holandština, Angličtina, Francouzština, Němčina a Italština). Později byly přidávány další jazyky, tak aby bylo použití SAMPA univerzální. Hlavní důvod pro vývoj této abecedy, byla její aplikace v počítačích, protože na rozdíl od abecedy IPA se k zápisu používají pouze znaky ASCII. Přes snahu o standardizaci bohužel vznikly abecedy pro jednotlivé jazyky a tak k úplnému sjednocení nedošlo[27].

3.3.3 Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA)

Jedná se o variaci SAMPA abecedy, která je postavena jako pravidla k přepisu IPA abecedy, tak aby byla možná univerzální interpretace. Tato abeceda byla vytvořena roku 1995 J. C. Wellsem a používá ASCII znaky jako k zápisu výslovností[29].



Obrázek 3.3: International Phonetic Alphabet (IPA).
Zdroj: internationalphoneticassociation.org

Kapitola 4

Použité korpusy a nástroje

Pro vytvoření jazykových modelů a jejich slovníků bylo nalezeno množství textových korpusů. Konkrétní korpusy i nástroje pro práci s nimi jsou popsány v této kapitole. Dále jsou zde uvedeny kroky pro přidání nového jazyka do vytvořeného algoritmu.

4.1 Textové korpusy

Nalezené textové korpusy jsou bez tematického zaměření a mají velkým objem textů, tudíž jsou velmi vhodné pro sestavení obecného jazykového modelu.

4.1.1 Europarl

Tento korpus byl vytvořen na základě jednání evropského parlamentu z let 1996-2011. Jedná se o korpus, který obsahuje texty dvaceti jedna jazyků států Evropské unie. V této práci byl Europarl korpus použit pro modely polského a maďarského jazyka [17].

4.1.2 Slovenský národní korpus

Základním korpusem pro slovenský jazyk byl Slovenský národní korpus, který se začal vytvářet v roce 2002 v Jazykovednom ústave Ľ. Štúra Slovenskej

akadémie vied. Tento korpus obsahuje texty různých žánrů a původů za posledních přibližně šedesát let. Korpus je rozdělen do subkorpusů dle žánru, období, regionů a dalších parametrů. Pro tuto práci byl zvolen druhý největší subkorpus s názvem prim-7.0-public-sane, který na rozdíl od největšího subkorpusu neobsahuje texty s nesprávnou diakritikou, před rokem 1955, z oblastí mimo Slovensko a z lingvistických časopisů[2].

4.1.3 Maďarský Webcorpus

Jedná se o korpus sesbíraný z osmnácti milionů internetových stránek z domény .hu. Tento korpus byl vytvořen v roce 2003 jako část projektu WordSword pod Budapest Institute of Technology Media Research and Education Center. Pro potřeby této práce byl použit korpus jmenující se 4%, který je oproti plnému korpusu pročištěn od velké většiny neplatného textu. Odstraněny byly například opakující se texty, texty nemaďarské, stránky s texty bez diakritiky a stránky s velkým množstvím chyb[10][18].

4.1.4 Polský národní korpus

Polský korpus byl sestaven spojením iniciativ subjektů: Institut počítačových věd při Polské Akademii Věd, Institut Polského Jazyka při Polské Akademii Věd, Polského Vědeckého Vydavatelství PWN a Oddělení Výpočetní a Korpusové Lingvistiky při Univerzitě v Łódźi. Zdrojem textů pro tento korpus byly texty různých žánrů knih, novin, časopisů, přepisů konverzací a internetových stránek. Korpus byl dokončen v roce 2012[12].

4.1.5 MultiUN

Korpus MultiUN je složen z dokumentů nacházejících se na webových stránkách Organizace spojených národů v letech 2000-2009. Tento korpus byl uveřejněn v roce 2010 a obsahuje texty pro sedm následujících jazyků.

1. anglický jazyk
2. francouzských jazyk
3. španělský jazyk
4. arabský jazyk

5. ruský jazyk
6. čínský jazyk
7. německý jazyk

V této práci byla použita pouze ruská část korpusu[5].

4.1.6 Common crawl korpus

V práci [1] byly pomocí dat získaných organizací Common Crawl[7] sestaveny korpusy novinových článků z let 2007-2014 pro šest následujících jazyků.

1. anglický jazyk
2. francouzských jazyk
3. český jazyk
4. finský jazyk
5. ruský jazyk
6. německý jazyk

K dispozici jsou korpusy pro jednotlivé roky i jazyky. Pro potřeby této práce byly použity ruské články z let 2014 a 2015. Tyto dva roky tvoří většinu dat sesbíraných pro ruskou část korpusu.

4.1.7 Ruský národní korpus

Ruský národní korpus zachycuje širokou škálu ruských textů od poloviny osmnáctého století až do současnosti. Je rozdělen do několika částí podle žánru, původu nebo použití. Nachází se v něm například paralelní korpusy, korpusy dialektů nebo poezie a další. Byl uveřejněn v roce 2003[13].

4.2 Databáze SpeechDat-E

Jedná se o databázi telefonních nahrávek pro pět jazyků střední a východní Evropy. Nahrávky byly vytvořeny pro češtinu, polštinu, maďarštinu, slovenštinu od 1000 mluvčích (1052 pro češtinu) a pro 2500 ruských mluvčí. K nahrávkám byly vytvořeny také přepisy. Promluvy v této databázi jsou foneticky bohaté, nelze je tedy přímo považovat za standardní. Součástí tohoto projektu bylo také vytvoření ověřených výslovnostních slovníků. Ty obsahují všechna slova objevující se v databázi a k jejich zápisu byla použita SAMPA[11].

4.3 SRILM

Tento balík je vyvíjen od roku 1997 ve SRI International[26]. Obsahuje nástroje k vytváření, práci a testování jazykových modelů. Postup jeho instalace a závislosti jsou popsány v práci [28].

Jednou z částí zadání této práce je vytvoření obecných jazykových modelů z volně dostupných zdrojů. Z tohoto důvodu bylo rozhodnuto, že algoritmus musí umět pracovat, jak se soubory vět, tak s n-gramy. Dalším požadavkem bylo, aby algoritmus uměl pracovat s velkým množstvím dat, tudíž byl použit následující postup. Nejprve byly pomocí nástroje `ngram-count` spočteny a seřazeny n-gramy každého korpusu zvlášť. Do zvláštního souboru se zároveň ukládaly cesty k jednotlivým souborům s n-gramy.

```
ngram-count -read in -no-sos -no-eos -sort -order NUM -write out
```

- `ngram-count`: název nástroje
- `-read in`: parametr pro čtení souboru s n-gramy, v případě souboru vět je místo `-read in` použito `-text in`, kde in je vstupní soubor
- `-no-sos, -no-eos`: vypnutí tokenů pro začátky a konce vět
- `-sort`: seřazení n-gramů podle abecedy
- `-order NUM`: stupeň vypočtených n-gramů
- `-write out`: zápis n-gramů do souboru out

Pro vytvoření modelů s daným počtem unigramů byl nejprve získán seznamu všech unigramů. Ten byl poté seřazen podle četnosti výskytu unigramů.

Prvních 60 000, 180 000 a 340 000 řádků z tohoto souboru bylo vybráno a uloženo jako slovník pro další použití.

Soubor s cestami byl následně využit v nástroji `merge-batch-counts`, který postupně kombinuje všechny soubory s n-gramy až vznikne jeden. Zároveň byly v tomto kroku sečteny stejné n-gramy. Tím došlo ke zvýraznění obecných slov.

Soubor se všemi n-gramy byl dále použit k výpočtu parametru vyhlazování. Důvodem použití všech n-gramů je aplikace Good-Turingova vyhlazování, které využívá informace o n-gramech objevujících se pouze jeden krát[23].

```
ngram-count -read in -gt1 out1 -gt2 out2
```

- `ngram-count:` název nástroje
- `-read in:` parametr pro čtení souboru s n-gramy
- `-gt1 out1, -gt2 out2:` v případě nespecifikování výstupního modelu, jsou ze vstupního souboru spočteny parametry pro vyhlazování

Model byl vytvořen ze souboru všech n-gramů, který byl omezen slovníky na požadované unigramy. Model je tak složen pouze z určených slov a jejich kombinací vyskytujících se v korpusech. Vyhlazení bylo provedeno podle vypočtených vyhlazovacích parametrů.

```
ngram-count -vocab slovnik -limit-vocab -read in -order NUM
     -gt1 out1 -gt2 out2 -lm model
```

- `ngram-count:` název nástroje
- `-vocab slovnik -limit-vocab:` omezení modelu pouze na slova obsažená ve slovnik
- `-read in:` parametr pro čtení souboru s n-gramy
- `-order NUM:` stupeň vypočtených n-gramů
- `-gt1 out1, -gt2 out2:` nyní je model specifikován, proto se soubory berou jako vstup
- `-lm model:` soubor pro zápis modelu

V případě zmenšování velikosti modelu je v nástroji `ngram` přítomný parametr `-prune`, který omezuje počty n-gramů stupně dva a více. Unigramy jsou zachovány všechny, ale ostatní stupně jsou omezeny, tak aby nedošlo ke změně perplexity větší než je definovaný parametr. Je tak možné výrazně zmenšit model při malé ztrátě informace.

```
ngram -lm in -prune level -order NUM -write-lm out
```

- `ngram`: název nástroje
- `-lm in`: parametr pro vstup jazykového modelu
- `-prune level`: úroveň povolené změny
- `-order NUM`: stupeň vypočtených n-gramů
- `-write-lm out`: výstupní soubor se zmenšeným modelem

4.4 Výslovnostní slovník

Pro vytvoření slovníku je třeba seznam slov používaných jazykovým modelem. Pokud máme model bez seznamu, lze jej získat použitím nástrojů pro práci s jazykovými modely. V našem případě byl k dispozici, protože jím byla omezována velikost výsledných modelů.

Pro zjednodušení práce se slovníky, bylo při vytváření využito průniku množiny slov v jazykových modelech pro daný jazyk. Protože největší model obsahuje všechna slova modelů menších, je možné uplatnit výslovnostní slovník největšího modelu i pro ostatní modely.

4.4.1 g2p-sk

Jedná se o nástroj sestavený Jozefem Ivaneckým[14]. Slouží k přepisu grafického zápisu do zápisu fonetického pro slovenský jazyk. Byl vytvořen na základě pravidel slovenské výslovnosti. Celkově bylo vytvořeno 257 pravidel. Tento nástroj je dostupný ve standardních repozitářích pro Ubuntu a Debian. Jeho nevýhodou je rychlost překladu - např. slovník o velikosti 60 000 slov byl přeložen za 2 hodiny.

```
g2p-sk [-dl] [-color] [-help] [ofile <file_name>] [<input_file>]
```

- `-dl:[0-5]:` úroveň debugování (standardní je 1)
- `-color:` barevný výstup
- `-help:` vypsání nápovědy
- `-ofile:` <file_name> zápis do souboru (standardně STDOUT)
- `-stats:` zobrazení statistiky

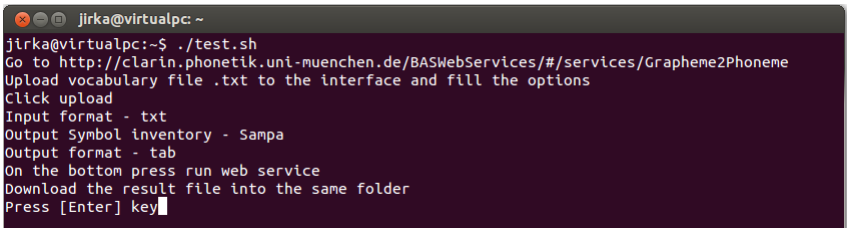
4.4.2 BAS G2P

Nástroj BAS G2P byl vyvinut na Ludwig-Maximilians-Universität München[24] a je dostupný přes webové rozhraní nebo webovou službu. V algoritmu vytvořeném k této práci se používá webové rozhraní. Jedná se o jeden z nástrojů pro úlohu zpracování řeči, jenž byl touto skupinou uveřejněn[15]. Jeho výhodou je dostupnost mnoha jazyků - v době psaní práce sedmnáct (pro anglický jazyk a švýcarskou němčinu lze navíc použít více variant), rychlost překladu - slovník o velikosti 340 000 slov je přeložen během několika minut a volná dostupnost pro nekomerční aplikace. Nástroj je založen na rozhodovacích stromech, které byly natrénovány na slovnících o velikosti řádově desítky tisíc slov pro každý z jazyků. [16]

Albánština	Nizozemština	Angličtina - Amerika
Estonština	Gruzínština	Angličtina - Austrálie
Finština	Francouzština	Angličtina - Velké Británie
Němčina	Maďarština	Angličtina - Nový Zéland
Italština	Polština	Švýcarská němčina(6 dialektů)
Rumunština	Ruština	Haitská kreolština
Slovenština	Španělština	

Tabulka 4.1: BAS G2P - seznam jazyků

Jak bylo uvedeno výše, v této práci bylo použito webové rozhraní tohoto nástroje. Při práci s vytvořeným algoritmem je uživatel vyzván k nahrání seznamu slov přes internetový prohlížeč. Po vstupu na odkazovanou stránku uživatel vloží do vyznačeného místa soubor obsahující seznam slov. Soubor musí mít jednu z povolených koncovek a musí být kódován UTF-8, oboje zajišťuje připravený skript. Poté uživatel stiskne upload, zvolí z roletek možnosti dle výstupu terminálu a stiskne Run Web Service, čímž začne běžet překlad. Po dokončení překladu je třeba, aby uživatel uložil slovník do stejné složky jako jsou seznamy slov a v terminálu stiskne tlačítko Enter[6].



Obrázek 4.1: BAS G2P - výzva k nahrání seznamu

4.4.3 Sequitur G2P

Jako poslední způsob generování výslovnostního slovníku byl zvolen Sequitur G2P. Jedná se o nástroj pracující na principu trénování modelů výslovnosti z výslovnostních slovníků[19]. Princip Sequitur G2P je stejný jako BAS G2P. Rozdíl je v tom, že Sequitur byl použit v kombinaci se slovníky s ověřenou výslovností. Oproti tomu u BAS G2P není kvalita ani původ většiny trénovacích dat znám.

Malá část výslovnostních slovníků byla před trénováním oddělena, aby bylo možné provést testování generování výslovností. Dle pokusů v uvedené práci byl zvolen počet iterací trénování rovno šesti.

Při trénování modelu je nejprve vytvořen unigramový model a s každou další iterací je zvednut stupeň modelu o jeden. Po dokončení iterace je možné model otestovat na testovací části slovníku. Pokud jsme spokojeni s dosaženou úspěšností, aplikujeme model na seznam slov.

```
g2p.py --encoding=UTF-8 -train in --devel 5% --write-model out-1
```

- `-encoding=UTF-8:` použité kódování UTF-8
- `-train in:` čtení trénovacích dat ze souboru in
- `-devel 5%:` použití 5% trénovacích dat jako held-out
- `-write-model out-1:` uložení natrénovaného modelu do out-1

```
g2p.py --encoding=UTF-8 --train in --ramp-up --model out-1
       --devel 5% --write-model out-2
```

- `--ramp-up:` zvednutí stupně modelu o jedna
- `--model out-1` vstupní model vytvořený v minulém kroku

```
g2p.py --encoding=UTF-8 --model out-N --test testLex
```

- `--test testLex:` testování vytvořeného modelu na testovacím slovníku testLex

```
g2p.py --encoding=UTF-8 --model out-N --apply words
```

- `--apply words:` aplikování vytvořeného modelu na soubor slov words

4.4.4 Adaptace výstupů g2p nástrojů do X-SAMPA

V této práci bylo pracováno s nástroji k vytvoření výslovnostních slovníků, jejichž výstupy byly ve formátu SAMPA. Rozpoznávání bylo realizováno v rozpoznávači, který pracuje s XSAMPA abecedou. V práci [9] jsou uvedena pravidla pro převod mezi SAMPA a XSAMPA abecedou. Protože výstup užitých nástrojů neodpovídal požadované abecedě fonémů, byly výslovnostní slovníky upravovány. Tyto úpravy nazýváme mapováním.

Slovenský jazyk

Nástroj g2p-sk se od rozpoznávače lišil pouze v deseti fonémech. Přepis samohlásek a,e,i,o,u a dvojhlásky ie je zřejmý. V případě ostatních fonémů došlo k zápisu fonému jeho zjednodušeným ekvivalentem využívaným v rozpoznávači. *f_v* bylo nahrazeno jako *w*, *G* jako *x* a *h* jako *h*. Při použití Sequitur slovníků nebylo nutné žádných změn.

g2p	**XSAMPA**	**g2p**	**XSAMPA**
A	a	E	e
O	o	I	i
y	i	U	u
G	x	f_v	w
i_e	i_^e	h\	h

Tabulka 4.2: Adaptace slovenský jazyk

Maďarský jazyk

V případě maďarského jazyka bylo nutné doplnit foném *N*, a to tam kde se po *n* vyskytuje *k* nebo *g*. Dále byl zaměněn foném *F* na místo *m* a *n*, pokud byly následovány *f* nebo *v*. Foném *b j:* byl nahrazen krátkou variantou, a ta poté za *C*. Nakonec byly

g2p	XSAMPA	g2p	XSAMPA
b j:	C	b j	C
n k	N k	n g	N g
d’	J\	d’:	J\:
n f	F f	m f	F f
n v	F v	m v	F v
a	A	O	A
t’	c	t’:	c:
t: t’:	c:	t t’:	c

Tabulka 4.3: Adaptace maďarský jazyk

g2p	XSAMPA	g2p	XSAMPA
g_j	g j	k_j	k j
n_j	n j	p_j	p j
dz_j	tz\	ts_j	ts\
x_j	x j	I	1
z_j	z\	s_j	s\
n k	N k	n g	N g

Tabulka 4.4: Adaptace polský jazyk

nahrazeny fonémy *t t’:* a *t’* za *c*, a stejně pro dlouho variantu za *c:*. Akustický model použitý v rozpoznávači byl oproti původní verzi[9] zjednodušen a neobsahuje fonémy *g:* a *h*. V případě Sequitur slovníku bylo třeba pouze mapovat *g:* na *g*, aby odpovídal zjednodušenému akustickému modelu.

Polský jazyk

V případě polského jazyka byl doplněn foném *N*, pro *n k* nebo *n g*. Dále byl upraven zápis měkčených souhlásek a *I* bylo změněno na *1*. Pro slovník Sequitur nebylo uplatněno žádné mapování.

Ruský jazyk

Rozpoznávač ruského jazyka byl realizován s akustickým modelem, který zanedbává měkčené fonémy *b’ f’ g’ k’ l’ p’ r’ v’ tS’* a také samohlásky s přízvukem označované pomocí ". Dále foném *h* byl zapsán zjednodušenou variantou *x*. Ostatní náhrady byly pouze změny zápisu. Slovník Sequitur obsahoval měkčené fonémy i samohlásky s přízvukem, oboje bylo mapováno dle tabulky.

g2p	XSAMPA	g2p	XSAMPA
h	x	v’	v
z’	z\	r’	r
t’	c	p’	p
x’	C	l’	l
m’	F	k’	k
d’	J\	g’	g
n’	N	f’	f
s’	s\	b’	b
S’	S:	tS’	tS
a"	a	e"	e
i"	i	1"	1
o"	o	u"	u

Tabulka 4.5: Adaptace ruský jazyk

4.5 Použití algoritmu

Algoritmus byl realizován několika skripty, a tak aby práce s ním byla co nejjednodušší. Při volání hlavního skriptu jsou zadávány čtyři parametry: jazyk, volba části zpracování, vstupní a výstupní soubor.

```
lm-multil.sh LANG [1-4] in out
```

- `LANG:` zvolení jazyka - HU,RU,PL nebo SK
- `[1-4]:` volba části zpracování
- `in, out:` vstupní a výstupní složka

1. Vytvoření souboru čistých unigramů

   Vstupem je složka obsahující připravené korpusy. Algoritmus umí pracovat se souborem vět i s n-gramy, případně s jejich kombinací. Z jednotlivých korpusů jsou spočítány unigramy a spojeny do jednoho seznamu. Ten je čištěn tak, aby obsahoval pouze slova složená z písmena daného jazyka.

2. Vytvoření modelů a jejich seznamů slov

   Vstupní i výstupní složka je stejná jako u 1. Zde je pomocí seznamu povolených slov vytvořen soubor s unigramy, který je seřazen podle četností. Poté dojde k vybrání prvních 60 000, 180 000 a 340 000 slov z těchto unigramů. Vyhlazovací metodou je Good-Turing[23], jejíž parametry se počítají zvlášť z celého korpusu. Nakonec jsou aplikovány GT parametry a omezené slovníky pro vytvoření konkrétních jazykových modelů.

3. Zmenšení počtu bigramů

   Vstupní i výstupní složka je stejná jako u předchozích. Ke zmenšení je použit nástroj `ngram` s parametrem `-prune 1e-8`. Zmenšené modely jsou uloženy ve stejné složce a ve jméně jsou označeny parametrem oříznutí.

4. Příprava výslovnostních slovníků

   Zde je vstupem soubor se seznamem slov, pro který má být vytvořen výslovnostní slovník. Výstupní složka zůstává stejná. V případě slovenského jazyka je využit g2p-sk, který probíhá bez dalšího uživatelova vstupu. Pro ostatní jazyky je uživatel veden k nahrání seznamu slov do webového rozhraní BAS G2P a k jeho nastavení. Nakonec je provedeno mapování a seřazení slovníku podle abecedy.

Dle výchozího nastavení, dojde k vytvoření šesti bigramových jazykových modelů. V případě požadování jiného stupně modelů, je možné toto snadno změnit. Stejně snadná je také změna parametru zmenšení modelu a změna velikosti slovníků k omezení modelů. Pro Sequitur G2P je připraven skript, který ulehčuje trénování modelů výslovnosti.

4.6 Úprava algoritmu pro přidání dalších jazyků

Přidání dalšího jazyka do vytvořeného algoritmu je realizováno v několika krocích. Nejjednodušší je přidání jazyka podporovaného BAS G2P, případně jazyka, pro který máme k dispozici ověřený slovník. Prvně je nutné najít vhodný textový korpus. V této práci byly nalezeny korpusy, které jsou více jazyčné a mohou být zpracovány.

V dalším kroku je nutné korpus upravit podle popisu v 3.1 Korpus a jeho čištění - UTF-8, jedna věta na řádek, bez interpunkce, malá písmena. V případě, že vstupním korpusem jsou n-gramy, je potřeba zkontrolovat jejich formát, a pokud neodpovídá, provést potřebné úpravy. Odstranění interpunkce, převod na malá písmena atd. Posledním krokem je uložení upravených korpusů do jedné složky.

Dále musí být do první části hlavního skriptu doplněna abeceda daného jazyka, tak aby byly odstraňovány věty s nepovolenými znaky. Následně zkopírovat do čtvrté části postup vytvoření výslovnostního slovníku, kde budou změněny názvy skriptů pro kontrolu fonému a mapování. Tato část musí být upravena podle zvoleného způsobu tvorby výslovnostního slovníku.

Nakonec se musí vytvořit zmíněné skripty. Pokud není známé, jaké bude prováděno mapování, bude mapovací skript zatím vynechán. Skript pro kontrolu použitých fonémů musí vracet fonémy, které rozpoznávač nezná. V posledním kroku je nutné experimentálně doplnit mapování fonémů.

Kapitola 5

Experimentální část

Tato kapitola popisuje provedená testování jazykových modelů a výslovnostních slovníků. Testování bylo provedeno na textových korpusech a promluvách SpeechDat-E databáze[11] shodné s prací [9], tak aby bylo možné srovnání dosažených výsledků. Primárně jsou zde prezentovány výsledky dosažené pro slovenský, polský, maďarský a ruský jazyk. Jazykové modely pro český jazyk vytvořeny nebyly, protože byly vytvořeny v rámci předchozí práce [22] z Českého národního korpusu. Pro porovnání byly české modely otestovány stejným způsobem jako modely ostatních jazyků. Jazykové modely jsou hodnoceny na dosažených perplexitách, počtu neznámých slov a úspěšností v rozpoznávání spojité řeči.

Akustická analýza byla prováděna MFCC s 25ms dlouhým Poveyho oknem. Dále byla aplikována banka 24 filtrů v rozsahu 100-3800Hz. Ponecháno bylo 13 koeficientů. Vypočteny byly dynamické koeficienty delta a delta-delta. Nakonec bylo použito LDA a MLLT transformace. K akustickému modelování byl zvolen SGMM model trifonů s diskriminativním přetrénováním bMMI. Celý rozpoznávač byl realizovaný nástroji Kaldi[21] a je detailně popsaný v práci [9].

5.1 Úprava korpusů

Pro jednotné zpracování bylo prvním krokem převedení korpusů do kódování UTF-8. K tomu byl použit `iconv`, standardně dostupný v linuxových distribucích. Dále byly korpusy čištěny, tak aby bylo ztraceno co nejméně dat před samotným zpracováním. Což spočívalo v zápisu jedné věty na řádek, odstranění interpunkce, uvozovek atd. Nakonec byl text převeden do malých písmen jednotlivých abeced. V případě korpusů XML formátu byly odstraněny XML značky. Dále byl korpus upravován jako po zápisu jedné věty na řádek. Finální čistění provádí algoritmus samostatně. Do výsledných modelů jsou zařazeny jen věty, které jsou složeny pouze z definovaných abeced jednotlivých jazyků.

- Ruský jazyk - абвгдеёжзийклмнопрстуфхцчшщъыьэюя
- Polský jazyk - aąbcćdeęfghijklłmnńoóprsśtuwyzźż
- Maďarský jazyk - aábcdeéfghiíjklmnoóöőpqrstuúüűvwxyz
- Slovenský jazyk - aáäbcčdďeéfghiíjklĺľmnňoóôpqrŕsštťuúvwxyýzž

Tímto omezením bylo dosaženo přítomnosti pouze slov, kterým dokáží nástroje vytvořit překlad. Zároveň bylo takto zajištěno, že pokud by byl jako vstup algoritmu použit nefiltrovaný nebo nedostatečně filtrovaný korpus nedostanou se do jazykového modelu problémová slova. Pokud by byl použit nefiltrovaný korpus jako vstup pro algoritmus, bylo by velké množství vět smazáno.

K definici formátu textu bylo použito nástrojů standardně dostupných v linuxových distribucích.

- Sed
- Perl
- Grep

Při práci s těmito nástroji se používají regulární výrazy. Díky nim lze obecně popsat vzory textu určeného k odstranění nebo ponechání. Výše byla zmíněna filtrace textu obsahující znaky neodpovídající abecedě daného jazyka. Představme si soubor, který je již částečně upravený, tak že obsahuje jednu větu bez diakritiky na každém řádku. Naším úkolem je nyní pro slovenský text odstranění takových vět, kde se vyskytují neplatné znaky. To lze provést například následujícím způsobem s nástrojem sed.

Protože sed pracuje po jednotlivých řádcích je definice regulárního výrazu níže následující: pro soubor `in` smaž řádky obsahující jiné znaky než písmena slovenské abecedy a mezeru a výsledek zapiš do souboru `out`.

```
sed "/[^aáäbcčdďeéfghiíjklĺľmnňoóôpqrŕsštťuúvwxyýzž ]/d" in > out
```

- `in > out:` názvy vstupního a výstupního souboru, kde znak > značí přesměrování STDOUT do souboru
- `sed:` název použitého nástroje
- `"/vzor/d":` odstranění řádku obsahující specifikovaný vzor
- `[^ab]:` hranaté závorky se používají k výčtu znaků a znak ^ slouží k negaci

5.2 Vytvořené jazykové modely

V této práci bylo vytvořeno dvacet čtyři jazykových modelů pro čtyři jazyky: polština, čeština, ruština a maďarština. Kompletní seznam a velikosti jednotlivých modelů uvádí Tabulka 5.1 v počtech unigramů a bigramů. Pro srovnání jsou uvedeny také velikosti českých obecných jazykových modelů. Z těchto dat vyplývá, že jazykové modely vytvořené v této práci jsou v průměru větší než české modely. Největší jsou modely slovenského jazyka a nejmenší modely jsou pro ruštinu. Při oříznutím počtu bigramů modelu s parametrem `1e-8`, dojde k přibližně troj- až čtyřnásobnému snížení bigramů. Dále je možné říct, že korpusy použité pro polské jazykové modely obsahovaly velké množství bigramů, které se objevily s malou četností. Tento fakt způsobil, že modely s oříznutím jsou nejmenší pro polštinu. Důvodem této skutečnosti by mohla být nedostatečná provázanost a velikost korpusu polštiny, což by mohlo způsobit méně přesnou předpověď slov polskými modely.

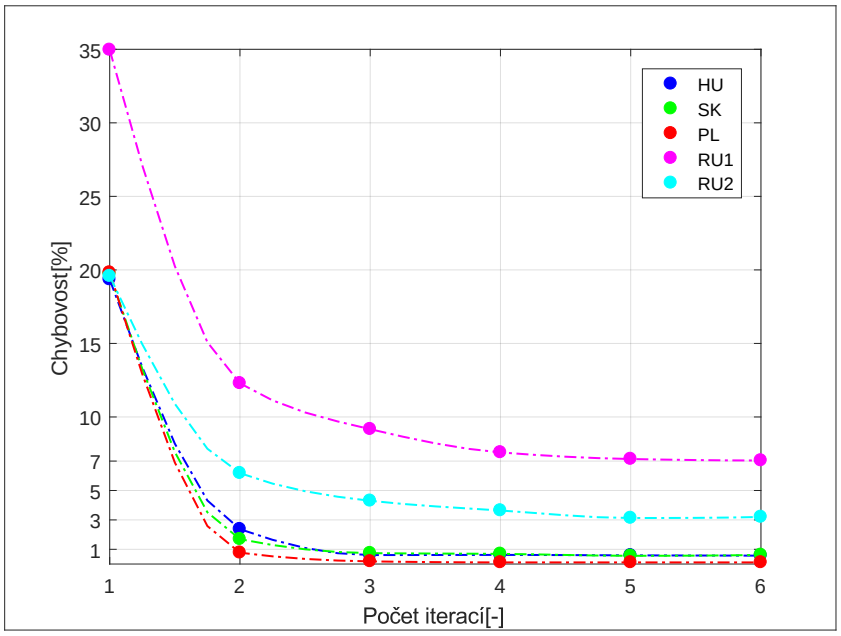
Model	**unigramy**[-]	**bigramy**[-]
SK-60000-2-1e-8	60 000	11 585 451
SK-60000-2	60 000	48 203 864
SK-180000-2-1e-8	180 000	15 550 476
SK-180000-2	180 000	75 096 963
SK-340000-2-1e-8	340 000	17 121 879
SK-340000-2	340 000	87 625 774
HU-60000-2-1e-8	60 000	10 937 468
HU-60000-2	60 000	34 070 911
HU-180000-2-1e-8	180 000	15 212 418
HU-180000-2	180 000	52 133 637
HU-340000-2-1e-8	340 000	17 380 667
HU-340000-2	340 000	61 534 247
PL-60000-2-1e-8	60 000	7 505 221
PL-60000-2	60 000	29 320 636
PL-180000-2-1e-8	180 000	9 848 182
PL-180000-2	180 000	42 161 772
PL-340000-2-1e-8	340 000	11 020 787
PL-340000-2	340 000	47 351 207
RU-60000-2-1e-8	60 000	9 478 316
RU-60000-2	60 000	26 699 120
RU-180000-2-1e-8	180 000	12 333 030
RU-180000-2	180 000	37 345 012
RU-340000-2-1e-8	340 000	13 399 200
RU-340000-2	340 000	41 504 742
CZ-60000-2	*60 000*	*27 305 979*
CZ-180000-2	*180 000*	*39 047 471*
CZ-340000-2	*340 000*	*43 210 293*

Tabulka 5.1: Srovnání velikostí modelů

5.3 Výslovnostní slovníky

Vytvořeny byly dva výslovnostní slovníky pro každý ze čtyř zpracovávaných jazyků. Principiálně se pracovalo se dvěma metodami generování výslovnosti. Prvním způsobem byla definice pravidel výslovnosti, tento způsob používá g2p-sk. Druhým způsobem bylo trénování modelů výslovnosti na základě jiných slovníků výslovnosti, tak pracují BAS G2P a Sequitur G2P. Rozdílem v jejich použití byly trénovací slovníky. Pro BAS G2P nejsou vstupní data známá pro většinu jazyků, v případě Sequitur G2P bylo k trénování modelů použito slovníků s ověřenou výslovností z databáze SpeechDat-E[11].

Při trénování modelů Sequitur G2P byly aplikovány parametry experimentálně určené v práci [19]. Testování úspěšnosti bylo provedeno na tisícovce slov v každém jazyce. Na následujícím grafu je uveden vývoj chybovosti při určení fonému v závislosti na stupni modelu. Pro maďarštinu, slovenštinu a polštinu bylo dosaženo chybovosti ~0.5%[1]. Pro ruštinu jsou uvedeny dva průběhy RU1 a RU2. Průběh označený jako RU1(purpurová) byl dosažen při trénování na slovníku obsahujícím i přízvučné samohlásky a měkčené souhlásky. V tomto případě byla dosažena chybovost ~7%. Průběh RU2(azurová) zobrazuje chybovost pro modely trénované po mapování trénovacího slovníku a dosahuje ~3%. Z grafu lze vyvodit, že zjednodušení zápisu výslovnosti zlepšilo chybovost přibližně na polovinu původní hodnoty. Výsledná úspěšnost je v porovnání s ostatními jazyky stále nízká a je způsobena změnou výslovnosti samohlásek v závislosti na okolním kontextu a přízvuku daného slova. Okolí samohlásky lze modelovat dobře, dle chybovosti ostatních jazyků. Lze tedy předpokládat, že problémové je modelování polohy přízvuku ve slově.



Obrázek 5.1: Chybovost určení fonémů

[1] Tabulka B.1 v příloze

Iterace	HU[%]	PL[%]	SK[%]	RU1[%]	RU2[%]
1	78.96	70.70	70.10	99.70	80.24
2	20.14	5.60	12.10	68.10	38.92
3	3.91	1.30	6.10	46.64	26.48
4	3.91	0.80	5.70	38.21	22.17
5	3.61	0.80	4.60	35.51	19.46
6	3.31	0.70	5.10	35.01	20.06

Tabulka 5.2: Chybovost modelování výslovnosti slova

V Tabulce 5.2 je zaznamenána chybovost generování výslovnosti celých slov v závislosti na počtu iterací trénování.

5.4 Testování Jazykových modelů

K testování a porovnání vytvořených jazykových modelů je použito tří kritérií. Perplexita PPL, počet neznámých slov OOV, ty se počítají z testovacího textového korpusu, a chybovost v rozpoznávači WER. Protože byly pro každý jazyk vytvořeny dva výslovnostní slovníky, je v tabulkách uvedeno WER1 pro slovník BAS G2P nebo g2p-sk a WER2 pro slovník Sequitur G2P. Pro porovnání modelů vytvořených v této práci jsou v tabulkách odděleně uvedeny také výsledky dosažené 0gramovými modely. Jedná se o modely vytvořené z trénovacích dat, tudíž pokrývají celý korpus, kromě ruského modelu, kde bylo jedenáct neznámých slov. Tyto modely neovlivňují výsledek rozpoznávání a výsledná úspěšnost je proto závislá pouze na akustickém modelu a výslovnostním slovníku[9].

5.4.1 Český jazyk

Pro porovnání výsledků dosažených modely vytvořenými v této práci, byly otestovány také české obecné jazykové modely vytvořené v práci [22]. Můžeme vidět, že mají nízké počty neznámých slov i perplexitu. V porovnání s 0gramovým modelem dosahují také podstatně vyšší úspěšnosti v rozpoznávání. Výslovnostní slovník českých modelů byl vytvořen na základě výslovnosti hlásek a nebyl dále kontrolován. Z dosažených výsledků lze proto říct, že pro češtinu je tento způsob generování výslovností dostačující.

5.4.2 Slovenský jazyk

Pro slovenské jazykové modely bylo dosaženo nejlepších výsledků ze všech jazyků. Perplexita a počet neznámých slov ukazuje na velkou podobnost trénovacího a testo-

Model	PPL	OOV[%]	WER[%]
CZ-60000-2	*1 463*	*10.42*	*29.03*
CZ-180000-2	*2 021*	*3.67*	*18.71*
CZ-340000-2	*2 621*	*1.65*	*15.83*
CZ-fiala-0	*14 802*	*0*	*27.50*

Tabulka 5.3: Výsledky modelů českého jazyka

vacího korpusu a velmi dobré zachycení pravidel slovenského jazyka. To lze přisoudit množství dat, které Slovenský národní korpus obsahuje. Při srovnání slovníku vytvořeného g2p-sk podle pravidel slovenského jazyka(WER1) vychází chybovost vyšší o ~2,5% než v případě výslovnosti generované Sequitur G2P výslovnostními modely(WER2). Tento fakt ukazuje na nesprávně definovaná pravidla v nástroji g2p-sk. Také lze díky těmto hodnotám říct, že slovenština patří mezi jazyky s dobře modelovatelnou výslovností. Úspěšnost zmenšených jazykových modelů je velmi podobná úplným modelům.

Model	PPL	OOV[%]	WER1[%]	WER2[%]
SK-60000-2-1e-8	2 493	9.19	27.93	25.62
SK-60000-2	2 471	9.19	27.31	25.45
SK-180000-2-1e-8	3 053	3.34	18.99	16.21
SK-180000-2	2 941	3.34	18.24	15.75
SK-340000-2-1e-8	3 397	1.46	16.73	13.63
SK-340000-2	1 315	1.46	15.84	13.08
Model	**PPL**	**OOV[%]**	**WER[%]**	**-**
SK-fiala-0	*7 885*	*0*	*24.77*	-

Tabulka 5.4: Výsledky modelů slovenského jazyka

5.4.3 Maďarský jazyk

Na základě perplexity a počtu neznámých slov, můžeme konstatovat, že maďarské modely poměrně dobře postihují testovací korpus. Počet OOV je oproti českým nebo slovenským vyšší, ale stále dosahuje dobrých hodnot. Pokud porovnáme úspěšnost v rozpoznávání proti 0gramovému modelu, vidíme, že srovnatelné jsou až modely s 340 000 unigramy. Menší modely mají díky vyššímu OOV úspěšnost malou. Ze srovnání použitých slovníků lze vyvodit, že model získaný z BAS G2P pro maďarštinu je podobně kvalitní jako model vytvořený v této práci v Sequitur G2P. Rozdíl mezi slovníky je ~1,5%. Pokud srovnáme dosažené úspěšnosti rozpoznávání mezi plnými a zmenšenými jazykovými modely, je vidět, že více než trojnásobné zmenšení počtu bigramů se projeví snížením úspěšnosti o ~1,5%. V porovnání s českými modely lze za důvod vyšších hodnot WER označit nesoulad slovníků v modelu a v promluvách, na který odkazuje hodnota OOV.

Model	**PPL**	**OOV**[%]	**WER1**[%]	**WER2**[%]
HU-60000-2-1e-8	1 949	13.04	33.74	32.69
HU-60000-2	1 849	13.04	33.33	32.26
HU-180000-2-1e-8	2 605	7.11	24.77	23.23
HU-180000-2	882	7.11	24.16	22.45
HU-340000-2-1e-8	2 945	5.21	22.14	20.40
HU-340000-2	2 536	5.21	21.17	19.31
Model	**PPL**	**OOV**[%]	**WER**[%]	**-**
HU-fiala-0	*7 550*	*0*	*18.94*	-

Tabulka 5.5: Výsledky modelů maďarského jazyka

5.4.4 Polský jazyk

Jazykové modely polského jazyka dosáhly nejhorších výsledků v testování na textovém korpusu. Mají nejvyšší perplexitu i počet neznámých slov. Úvaha, diskutována v 5.2 Vytvořené jazykové modely, se těmito daty potvrzuje. Výsledkem nedostatečně velkého korpusu jsou modely, které mají horší schopnost předpovědi slov. Hodnota OOV největšího polského jazykového modelu odpovídá hodnotám menších modelů ostatních jazyků. Úspěšnosti v rozpoznávání odpovídají předpokladům testování na textovém korpusu. Porovnat můžeme polské modely 340 000 unigramů s maďarskými modely 180 000 unigramů, kdy OOV je ~7% a dosažená úspěšnost WER ~27%(PL) a ~23%(HU). Úspěšnost použitých slovníků se liší výrazněji než v případě maďarštiny. Výslovnostní model natrénovaný z ověřených výslovnostních slovníků se projevuje snížením chybovosti o ~6%. Úplný a zmenšený jazykový model dosahují přibližně shodných výsledků, rozdíl menší než 0,5%. Tato hodnota je nižší než u ostatních jazyků a dále potvrzuje nesoulad jazykového modelu s obsahem promluv. Ze srovnání s 0gramovým modelem vychází nově vytvořený model hůře o ~7,5%.

Model	**PPL**	**OOV**[%]	**WER1**[%]	**WER2**[%]
PL-60000-2-1e-8	5 274	25.67	60.96	55.73
PL-60000-2	5 071	25.67	60.96	55.97
PL-180000-2-1e-8	8 833	12.14	41.31	34.75
PL-180000-2	8 305	12.14	41.07	34.60
PL-340000-2-1e-8	11 506	7.26	34.70	27.35
PL-340000-2	10 775	7.26	34.51	26.98
Model	**PPL**	**OOV**[%]	**WER**[%]	**-**
PL-fiala-0	*7 453*	0	*19.51*	-

Tabulka 5.6: Výsledky modelů polského jazyka

5.4.5 Ruský jazyk

Zjištěná perplexita i počet slov mimo slovník jsou nadprůměrné. V případě ruských modelů, však nedošlo k velkému poklesu bigramů u zmenšených modelů jako tomu bylo u polských modelů. Z toho se dá usoudit, že kvalita textového korpusu ruského jazyka byla vyšší. V porovnání PPL a OOV s polskými modely, dosahují ruské modely přibližně výsledků modelů o stupeň větších. Z těchto hodnot bylo usouzeno, že v rozpoznávání by měly být úspěšnější než polské modely. Úspěšnost dosažená v rozpoznávání je ovšem nejhorší ze zpracovávaných jazyků. Se slovníkem BAS G2P byla chybovost 55-62%. Použitím slovníku Sequitur byla úspěšnost zvýšena o ~12%. To je nejvyšší dosažený rozdíl při porovnání slovníků. Jak bylo diskutováno v 5.3 Výslovnostní slovníky, pro ruský jazyk měly výslovnostní modely nejvyšší chybovost, a to ~3% na foném. Průměrné slovo v ruském korpusu, zjištěno z testovací části ruského výslovnostního slovníku, má délku ~8 znaků. Výsledkem je, že celkově se alespoň jeden špatně určený foném vyskytl ve 20% slov testovací části slovníku. Porovnáme-li chybovost určení výslovnosti celých slov s ostatními jazyky,

- PL špatně určeno 0,7% slov, průměrná délka slova ~7 znaků
- HU špatně určeno 3,3% slov, průměrná délka slova ~10 znaků
- SK špatně určeno 5,1% slov, průměrná délka slova ~9 znaků

vidíme, že chybovost rozpoznávání byla velmi ovlivněna chybami ve výslovnostních slovnících. Pro dosažení lepších výsledků by bylo vhodné použití nástroje generující výslovnosti na základě definice pravidel, spíše než trénováním výslovnostních modelů. Rozdíl dosažených výsledků vytvořených modelů oproti 0gramovému modelu je také způsoben zmenšenou fonetickou sadou rozpoznávače. Podle výsledků v práci [9] tím dojde také ke zhoršení úspěšnosti. To bylo také dokázáno otestováním 0gramového modelu, který má o ~4,5% vyšší chybovost než v práci [9].

Model	**PPL**	**OOV**[%]	**WER1**[%]	**WER2**[%]
RU-60000-2-1e-8	4 350	17.44	61.80	49.70
RU-60000-2	4 211	17.44	60.69	49.35
RU-180000-2-1e-8	6 565	9.11	56.06	43.03
RU-180000-2	3 112	9.11	54.98	42.12
RU-340000-2-1e-8	7 972	6.19	55.53	42.07
RU-340000-2	7 504	6.19	54.48	40.92
Model	**PPL**	**OOV**[%]	**WER**[%]	**-**
RU-fiala-0	*9 652*	*0.07*	*33.56*	-

Tabulka 5.7: Výsledky modelů ruského jazyka

Kapitola 6

Závěr

Tato práce je dílčím řešením multilingválního rozpoznávače spojité řeči. Jejím konkrétním přínosem je řešení problému s chybějícími obecnými jazykovými modely pro některé jazyky. V rámci této práce byly z volně dostupných zdrojů vytvořeny obecné jazykové modely čtyř jazyků přítomných ve SpeechDat-E databázi[11]: slovenština, polština, maďarština a ruština. Pro češtinu, která je také součástí SpeechDat-E, byly obecné jazykové modely již vytvořeny dříve v práci [22]. Pro každý z jazyků bylo nalezeno několik textových korpusů, které byly čištěny. Nakonec proběhlo zpracování balíkem nástrojů SRILM[26]. Bylo vytvořeno šest jazykových modelů v jednotlivých jazycích. Jeden jazykový model o velikosti 340 000 unigramů, poté omezením počtu unigramů na 180 000 a 60 000 byly vytvořeny další dva modely. Poslední tři modely vznikly ořízntutím počtu bigramů z každého z modelů. Celkově tak bylo v této práci vytvořeno 24 jazykových modelů. Otestovány byly kromě modelů vytvořených v této práci také modely českého jazyka a 0gramové modely vytvořené v práci [9].

Následně byly použity tři nástroje pro vytvoření výslovnostních slovníků: g2p-sk[14], BAS G2P[15] a Sequitur G2P[19]. g2p-sk je nástroj postavený na definici pravidel slovenského jazyka. BAS G2P je online nástroj obsahující natrénované modely výslovností pro několik jazyků. V této práci byl použit BAS G2P pro polštinu, maďarštinu a ruštinu. Sequitur G2P slouží k natrénování vlastních modelů výslovnostních pravidel. Jako trénovací data pro Sequitur G2P byly vloženy ověřené výslovnostní slovníky při databázi SpeechDat-E.

Kromě jazykových modelů, výslovnostních modelů a výslovnostních slovníků je výstupem této práce také metodika zpracování textových korpusů a skripty její implementace. Dále je možné dle uvedeného postupu rozšířit skripty o další jazyky.

Dosažené výsledky

Pro každý z jazyků byly vytvořeny dva výslovnostní slovníky, jeden pomocí Sequitur G2P a druhý použitím g2p-sk nebo BAS G2P. Pro Sequitur bylo také provedeno zhodnocení přínosu počtu iterací trénování na snížení chyby generování výslovnosti. Trénováním Sequitur G2P vzniklo šest modelů pro každý z jazyků, které s různou přesností zachycují pravidla trénovacích slovníků. V případě maďarského, slovenského a polského jazyka bylo s modelem třetího stupně dosaženo chybovosti pod 1%. Pro ruský jazyk byly srovnány dvě varianty slovníků, první obsahovala úplnou fonetickou sadu ruského jazyka a u druhé byla sada zmenšena o měkčené souhlásky a přízvučné samohlásky. V rozpoznávání řeči jsou otestovány pouze výslovnostní modely nejvyššího, šestého, stupně. Aby bylo možné použít pár jazykový model a výslovnostní slovník v rozpoznávání řeči, byl pro každý jazyk realizován jednoduchý nástroj pro mapování fonémů dle pravidel v práci [9].

Hodnocení vytvořených modelů a slovníků je provedeno na textových korpusech a v systému rozpoznávání spojité řeči. Rozpoznávačem řeči byl systém popsaný v práci [9]. Dosažené výsledky jsou také srovnány se 0gramovými jazykovými modely vytvořenými v práci [9] a obecnými modely práce [22]. Jazykové modely práce [9] byly vytvořeny z přepisů promluv, na kterých bylo provedeno testování. Zároveň byly pro tyto promluvy vytvořeny ověřené výslovnostní slovníky. Použitím takto vytvořené kombinace 0gramových jazykových modelů a ověřených slovníků ukazuje na vlastnosti akustické části rozpoznávače.

Dosažená úspěšnost obecných jazykových modelů byla negativně ovlivněna složením promluv databáze SpeechDat-E. Databáze byla tvořena foneticky bohatými větami, které neodpovídají běžné formě jazyka. Jazyk modelovaný všemi jazykovými modely se dále liší z důvodu čísel ve větách. V případě SpeechDat-E databáze jsou v promluvách ponechány čísla a v přepisech jsou přepsány do slovní podoby, což také ovlivnilo výpočet průměrné délky slov na testovacím korpusu, kdy byla slova v průměru delší než 8 znaků. Dá se předpokládat, že tento zápis čísla v textovém korpusu není příliš častý a častěji budou čísla zapsána formou číslovek. Protože věty s číslovkami byly odstraňovány z textového korpusu při vytváření jazykových modelů, je schopnost práce modelu v takovýchto větách zhoršena.

Pro jednotlivé jazyky bylo dosaženo následujících úspěšností: slovenský jazyk ~13%, maďarský jazyk ~19%, polský jazyk ~27% a ruský jazyk ~41%.

Porovnáním výsledků dosažených použitím BAS G2P a Sequitur G2P v polštině, maďarštině a ruštině lze říct, že se online nástroj BAS G2P pro generování výslovností těchto jazyků příliš nehodí. Srovnatelného výsledku bylo dosaženo pouze pro maďarský jazyk, kde byl rozdíl ~1,5%. U polského a ruského jazyka byla úspěšnost o 6% a 12% nižší než Sequitur G2P. Rozdíl mezi g2p-sk a Sequitur G2P byl ~2,5% ve prospěch nástroje Sequitur. Tento údaj ukazuje na nesprávnou definici pravidel slovenské výslovnosti v g2p-sk. Vzhledem k době trvání vytvoření slovníku pomocí g2p-sk a dosažené chybovosti, nelze jeho použití doporučit. Jelikož jsou dostupné ověřené výslovnostní slovníky databáze SpeechDat-E, je vhodnější využít výslovnostních modelů trénovaných pomocí Sequitur G2P u všech zpracovávaných jazyků.

Pro ruský jazyk bylo zjištěno, že výslovnost generovaná výslovnostními modely má velkou chybovost, 3% na foném a 20% při určení výslovnosti celého slova. Pro ostatní jazyky byla chybovost určení celého slova následující: PL ~1%, HU ~3% a SK ~5%. Z tohoto srovnání vidíme, že chybovost určení jednoho fonému v ruštině je srovnatelná s chybovostí určení celého slova u ostatních jazyků. Důvodem této chybovosti je hlavně změna výslovnosti samohlásek v ruském jazyce. Na základě této analýzy lze vyvodit, že aplikace výslovnostních modelů pro ruský jazyk je méně vhodná, než pro ostatní zpracovávané jazyky.

Výsledky dosažené jazykovými modely se zmenšeným počtem bigramů ukazují, při srovnání s úplnými modely, výhodnost tohoto způsobu přípravy. Modely jsou přibližně trojnásobně menší a jejich úspěšnost je v průměru pouze o ~1% horší než u úplných modelů.

Budoucí práce

Při analyzování dosažených výsledků byly vyvozeny některé závěry, které by mohly být zkoumány v dalších pracích. Z pohledu analyzovaných korpusů se zdá, že pro modelování polského jazyka by bylo vhodné mít k dispozici větší korpus. Při srovnání způsobů generování výslovnosti byla zjištěna vysoká chybovost v případě ruského jazyka a bylo doporučeno použití definování pravidel výslovnosti namísto modelování. Posledním námětem je zvýšení počtu unigramů modelů a zmenšení počtu bigramů, tím by bylo sníženo OOV při zachování rozumné velikosti jazykového modelu.

Příloha A

Literatura

[1] O. Bojar, R. Chatterjee, Ch. Federmann, B. Haddow, M. Huck, Ch. Hokamp, P. Koehn, V. Logacheva, Ch. Monz, M. Negri, M. Post, C. Scarton, L. Specia, M. Turchi. *Findings of the 2015 Workshop on Statistical Machine Translation.* In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, 2015. Association for Computational Linguistics.

[2] Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2015. *Slovenský národný korpus – prim-7.0-public-all.* `http://korpus.juls.savba.sk`, [cit.2017-01-03].

[3] Aleš Brich. *Analýza robustnosti moderních rozpoznávačů řeči na bázi TANDEM architektury.* Master's thesis, České vysoké učení technické v Praze, 2016.

[4] Kamil Chalupníček. *Rozpoznávání diktované řeči pro medicínské aplikace.* Master's thesis, Vysoké učení technické v Brně, 2004.

[5] A. Eisele, Y. Chen. *MultiUN: A Multilingual Corpus from United Nation Documents.* In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 2868–2872, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[6] CLARIN-D. *Short guide on how to use BAS G2P.* `http://de.clarin.eu/en/g2p-en`, [cit.2017-01-03].

[7] Common Crawl. *Common Crawl corpus.* `https://commoncrawl.org/`, [cit.2017-01-03].

[8] Filozofická fakulta Univerzity Karlovy v Praze. *Český národní korpus.* `http://korpus.cz/`, [cit.2017-01-03].

[9] Jiří Fiala. *DNN-HMM Based Multilingual Recognizer of Telephone Speech.* Master's thesis, Czech technical university in Prague, 2016.

[10] P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, V. Trón. *Creating open language resources for Hungarian.* In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*, pages 203–210, Lisbon, Portugal, 2004.

[11] H. Heuvel, J. Boudy, Z. Bakcsi, J. Cernocky, V. Galunov, J. Kochanina, W. Majewski, P. Pollak, M. Rusko, J. Sadowski, P. Staroniewicz, H. S. Tropf. *SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed.* In *Proceedings of 7th European conference on speech communication and technology (EUROSPEECH)*, Aalborg, Denmark, 2001.

[12] Institute of Computer Science, Polish Academy of Sciences. *National Corpus of Polish.* `http://nkjp.pl/`, [cit.2017-01-03].

[13] Russian Academy of Sciences Institute of Russian language. *Russian National Corpus.* `http://www.ruscorpora.ru/en/index.html`, [cit.2017-01-03].

[14] J. Ivanecký. *Analysis of Rule Based Phonetic Transcription Technique Applied to the Slovak Language.* In *Computer Treatment of Slavic and East European Languages*, pages 130–136, Bratislava, Slovakia, 2005. Vydavateľstvo Slovenskej akadémie vied.

[15] T. Kisler, U. D. Reichel, F. Schiel, Ch. Draxler, B. Jackl, N. Pörner. *BAS Speech Science Web Services - an Update of Current Developments.* In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA).

[16] U. D. Reichel, T. Kisler. *Language-independent grapheme-phoneme conversion and word stress assignment as a web service.* In *Elektronische Sprachsignalverarbeitung, Studientexte zur Sprachkommunikation*, pages 42–49, Dresden, Germany, 2014. TUDpress.

[17] P. Koehn. *Europarl: A Parallel Corpus for Statistical Machine Translation.* In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

[18] A. Kornai, P. Halácsy, V. Nagy, Cs. Oravecz, V. Trón, D. Varga. *Web-based frequency dictionaries for medium density languages.* In *Proceedings of the Second International Workshop on Web as Corpus*, pages 1–8, Trento, Italy, 2006. Association for Computational Linguistics.

[19] M. Bisani, H. Ney. *Joint-sequence Models for Grapheme-to-phoneme Conversion.* Speech Communication, 50(5):434–451, 2008.

[20] Eugen Pauliny. *Slabičné [r], [l] v slovenčine.* Slovo a slovesnost, (4):307–310, 1977.

[21] Daniel Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, K. Veselý. *The Kaldi Speech Recognition Toolkit.* In *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Waikoloa Village, Hawaii, 2011. IEEE Signal Processing Society.

[22] V. Procházka, P. Pollák, J. Žďánský, J. Nouza. *Performance of Czech Speech Recognition with Language Models Created from Public Resources.* Radioengineering, 40(4):1002–1008, 2011.

[23] J. Psutka, L. Müller, J. Matoušek, V. Radová. *Mluvíme s počítačem česky.* Academia, 2006.

[24] U. D. Reichel. *PermA and Balloon: Tools for string alignment and text processing.* In *Interspeech, Thirteenth Annual Conference of the International Speech Communication Association*, pages 346–349, Portland, Oregon, 2012. International Speech Communication Association (ISCA).

[25] S. Renals, H. Shimodaira. *Automatic Speech Recognition*. `http://www.inf.ed.ac.uk/teaching/courses/asr/`, [cit.2017-01-03]. The University of Edinburgh.

[26] SRI International - R&D for Government and Business. *The SRI Language Modeling Toolkit*. `http://www.speech.sri.com/projects/srilm/`, [cit.2017-01-03].

[27] UCL Psychology & Language Sciences. *SAMPA - computer readable phonetic alphabet*. `http://www.phon.ucl.ac.uk/home/sampa/`, [cit.2017-01-03].

[28] Jiří Valíček. *Jazykové modely pro rozpoznávání řeči v různých tematických oblastech*. Bachelor's thesis, České vysoké učení technické v Praze, 2014.

[29] J. C. Wells. *Computer-coding the IPA: A proposed extension of SAMPA*. Speech, Hearing and Language, Work in Progress, 8:271–289, 1994.

[30] Ábe Kráľ. *Pravidlá slovenskej výslovnosti*. Slovenské pedagogické nakladateľstvo, 1988.

Příloha B

Přílohy

num	SAMPA	IPA	X-SAMPA	num	SAMPA	IPA	X-SAMPA
1	i	i	i	20	Z	ʒ	Z
2	I	ɨ	1	21	s’	ɕ	s\
3	e	e	e	22	z’	ʑ	z\
4	a	a	a	23	x	x	x
5	o	o	o	24	ts	ts	ts
6	u	u	u	25	dz	dz	dz
7	e ~	ẽ	e~	26	tS	tʃ	tS
8	o~	õ	o~	27	dZ	dʒ	dZ
9	p	p	p	28	ts’	tɕ	ts\
10	b	b	b	29	dz’	dʑ	dz\
11	t	t	t	30	m	m	m
12	d	d	d	31	n	n	n
13	k	k	k	32	n’	ɲ	J
14	g	g	g	33	N	ŋ	N
15	f	f	f	34	l	l	l
16	v	v	v	35	r	r	r
17	s	s	s	36	w	w	w
18	z	z	z	37	j	j	j
19	S	ʃ	S				

Obrázek B.1: Polská XSAMPA Zdroj:[9]

num	SAMPA	IPA	X-SAMPA	num	SAMPA	IPA	X-SAMPA
1	i	i	i	27	J	ɲ	J
2	e	e	e	28	v	v	v
3	a	a	a	29	u_^	w	w
4	o	o	o	30	i_^	j̯	j_r
5	u	u	u	31	j	j	j
6	i:	iː	i:	32	p	p	p
7	e:	eː	e:	33	b	b	b
8	a:	aː	a:	34	t	t	t
9	o:	oː	o:	35	c	c	c
10	u:	uː	u:	36	d	d	d
11	i_^a	i̯a	i_^a				
12	i_^e	i̯e	i_^e	37	k	k	k
13	i_^u	i̯u	i_^u	38	g	g	g
14	u_^o	u̯o	u_^o	39	f	f	f
15	r	r	r	40	w	w	w
16	r=	r̩	r=	41	s	s	s
17	r=:	r̩ː	r=:	42	z	z	z
18	l	l	l	43	S	ʃ	S
19	l=	l̩	l=	44	Z	ʒ	Z
20	l=:	l̩ː	l=:	45	x	x	x
21	L	ʎ	L	46	h	h	h
22	m	m	m	47	ts	ts	ts
23	F	ɱ	F	48	tS	tʃ	tS
24	n	n	n	49	dz	dz	dz
25	N\	n	n	50	dZ	dʒ	dZ
26	N	ŋ	N				

Obrázek B.2: Slovenská XSAMPA Zdroj:[9]

Iterace	**HU**[%]	**PL**[%]	**SK**[%]	**RU1**[%]	**RU2**[%]
1	19.35	19.82	19.75	34.96	19.58
2	2.37	0.77	1.69	12.29	6.18
3	0.61	0.18	0.74	9.17	4.30
4	0.62	0.11	0.69	7.59	3.64
5	0.59	0.11	0.56	7.14	3.13
6	0.57	0.11	0.62	7.04	3.20

Tabulka B.1: Chybovost modelování fonému

num	SAMPA	IPA	X-SAMPA	num	SAMPA	IPA	X-SAMPA
1	i	i	i	35	tS	tʃ	tS
2	i:	iː	i:	36	tS:	tʃː	tS:
3	E	ɛ	E	37	dZ	dʒ	dZ
4	e:	eː	e:	38	f	f	f
5	O	ɑ	A	39	f:	fː	f:
6	A:	aː	a:	40	v	v	v
7	o	o	o	41	v:	vː	v:
8	o:	oː	o:	42	s	s	s
9	2	ø	2	43	s:	sː	s:
10	u	u	u	44	z	z	z
11	u:	uː	u:	45	z:	zː	z:
12	y	y	y	46	S	ʃ	S
13	y:	yː	y:	47	S:	ʃː	S:
14	2:	øː	2:	48	Z	ʒ	Z
15	p	p	p	49	Z:	ʒː	Z:
16	p:	pː	p:	50	m	m	m
17	b	b	b	51	m:	mː	m:
18	b:	bː	b:	52	n	n	n
19	t	t	t	53	n:	nː	n:
20	d	d	d	54	J	ɲ	J
21	t’	c	c	55	J:	ɲː	J:
22	d’	ɟ	J\	56	r	r	r
23	t’:	cː	c:	57	r:	rː	r:
24	t:	tː	t:	58	l	l	l
25	d’:	ɟː	J\:	59	l:	lː	l:
26	d:	dː	d:	60	j	j	j
27	k	k	k	61	j:	jː	j:
28	k:	kː	k:	62	h	h	h
29	g	g	g	63	h:	hː	h:
30	g:	gː	g:	64	x	x	x
31	ts	ts	ts	65	F	ɱ	F
32	dz	dz	dz	66	N	ŋ	N
33	ts:	tsː	ts:	67	h\	ɦ	h\
34	dz:	dzː	dz:	68	x’	ç	C

Obrázek B.3: Maďarská XSAMPA Zdroj:[9]

num	SAMPA	IPA	X-SAMPA	num	SAMPA	IPA	X-SAMPA
1	"1	ˈi	"1	26	tS’	tʃʲ	tS’
2	"a	ˈa	"a	27	t-S’	tʃʲ	tS’
3	"e	ˈe	"e	28	t-s	ts	ts
4	"i	ˈi	"i	29	f	f	f
5	"o	ˈo	"o	30	f’	fʲ	f’
6	"u	ˈu	"u	31	v	v	v
7	i	i	i	32	v’	vʲ	v’
8	1	ɨ	1	33	s	s	s
9	e	e	e	34	s’	ɕ	s\
10	a	a	a	35	z	z	z
11	o	o	o	36	z’	ʑ	z\
12	u	u	u	37	S	ʃ	S
13	p	p	p	38	S’:	ʃʲ ː	S’:
14	p’	pʲ	p’	39	Z	ʒ	Z
15	b	b	b	40	x	x	x
16	b’	bʲ	b’	41	x’	ç	C
17	t	t	t	42	m	m	m
18	t’	c	c	43	m’	mʲ	F
19	d	d	d	44	n	n	n
20	d’	ɟ	J\	45	n’	ɲ	N
21	k	k	k	46	l	l	l
22	k’	kʲ	k’	47	l’	lʲ	l’
23	g	g	g	48	r	r	r
24	g’	gʲ	g’	49	r’	rʲ	r’
25	ts	ts	ts	50	j	j	j

Obrázek B.4: Ruská XSAMPA Zdroj:[9]

Příloha C

Seznam zkratek

Zkratka	**Význam**
SRILM	Stanford Research Institute Language Modeling
G2P	Grapheme to Phoneme
LVCSR	Large Vocabulary Continuous Speech Recognition
WER	Word Error Rate
MFCC	Mel-Frequency Cepstral Coefficients
PLP	Perceptuální Lineárně Prediktivní Analýza řEči
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
HMM	Hidden Markov Model
PPL	Perplexita
XSAMPA	The Extended Speech Assessment Methods Phonetic Alphabet
IPA	International Phonetic Alphabet
SAMPA	Speech Assessment Methods Phonetic Alphabet
ASCII	American Standard Code for Information Interchange
LDA	Linear Discriminant Analysis
MLLT	Maximum Likelihood Linear Transform
SGMM	Subspace Gaussian Mixture Model
MMI	Maximal Mutual Information
XML	Extensible Markup Language
OOV	Out of Vocabulary