TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Multilingvální systémy rozpoznávání řeči a jejich efektivní učení

Disertační práce

Studijní program: P2612 – Elektrotechnika a informatika
Studijní obor: 2612V045 – Technická kybernetika

Autor práce: **Ing. Radek Šafařík**
Vedoucí práce: prof. Ing. Jan Nouza, CSc.

Liberec 2020

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Multilingual speech recognition systems and their effective learning

Dissertation

Study programme: P2612 – Electrotechnics and informatics
Study branch: 2612V045 – Technical cybernetics

Author: **Ing. Radek Šafařík**
Supervisor: prof. Ing. Jan Nouza, CSc.

Liberec 2020

Prohlášení

Byl jsem seznámen s tím, že na mou disertační práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé disertační práce pro vnitřní potřebu TUL.

Užiji-li disertační práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Disertační práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé disertační práce a konzultantem.

Současně čestně prohlašuji, že texty tištěné verze práce a elektronické verze práce vložené do IS STAG se shodují.

1. 12. 2020 Ing. Radek Šafařík

Abstract

The diseratation thesis deals with creation of automatic speech recognition systems (ASR) and with effective adaptation of already existing system to a new language. Today's ASR systems have modular stucture where individual moduls can be considered as language dependent or independent. The main goal of this thesis is research and development of methods that automate and make the development of language dependent modules effective as much as possible using free available data from the internet, machine learning methods and similarities between langages. It is accompanied by documented application and testing of the methods on the major Slavic languages. The work is associated with research projects dealing with development of broadcast monitoring systems for Slavic languages.

In the first part, basic concepts and the state of the art are described with focus on individual moduls and parts of the development of ASR systems. It is followed by description of Slavic languages with respect to ASR. The main part of the work is divided into two parts. The first one deals with lingvistic-lexical aspects of the development and the second one deals with acoustic-phonetic aspects.

The lingvistic-lexical part deals with the development of a text corpus, a pronunciation lexicon and a language model. Principles and procedures for effective gathering and processing of text data obtained from the internet are described here. The text data needs to be cleaned from unwanted elements, normalized and langauge filtering should be applied. In case of a language using non-latin alphabet, it is appropriate to make an alphabet conversion. Cyrilic-to-latin alphabet conversion was designed for this purpose. Then, words are chosen from the corpus to create the lexicon and statistical language model is computed.

The acoustic-phonetic part deals with the development of a phonetic inventory, creation of pronunciation for words in lexicon and the development of an acoustic model (AM). First, principles of a selection of phonemes for a new language and approaches for the creation of pronunciations are described. Next, approaches for gathering acoustis data from the internet and their processing for creation of an AM are described. Three AM training schemes are described. First supervised approach uses recordings with phonetic anotations from which the AM is trained.

Second lightly-supervised approach uses recordings together with some accompanying text which might contain parts of the speech in the recordings. The recordings

are transribed by an existing speech recognizer and any match between the output and the accompanying text is being searched. Matching parts are cut and added to the train set. All recordings are iteratively processed and more training data are gathered. In the case when the development of a system for a new language, acoustic data from another language can be used in multilingual system for gathering data for the target langauge.

Third unsuprvised approach uses several different ASR systems to create phonetic annotations for recordings without any related text. Recordings are transribed with all systems and if their outputs match the output is used as its phonetic annotation.

To test all created systems, standardized test sets were created from real data. Final versions of the systems were tested on the test sets to evaluate their usability in the broadcast monitoring tasks. Most of the systems achieved results below 20% of Word Error Rate.

As last, proposed methods where applied to another three europen languages. The development was performed mostly automatically using only free available data from the internet. The systems achieved results below 22% of Word Error Rate after few months of development.

Keywords

automatic speech recognition, language modeling, acoustic modeling, machine learning, multilingual systems, GMM, DNN, Slavic languages

Poděkování

Rád bych poděkoval panu prof. Ing. Janu Nouzovi, CSc. za jeho pomoc, čas a profesionální vedení během doktorského studia a při tvorbě této práce. Dále bych rád poděkoval všem kolegům na pracovišti, kteří přispěli ke tvorbě této práce. A v neposlední řadě bych rád poděkoval své rodině a přátelům za podporu po celou dobu studia.

Obsah

Seznam zkratek

ACC	Vyhodnocovací metrika správnosti (Accuracy)
AM	Akustický model (Acoustic Model)
ASR	Automatické rozpoznávání řeči (Automatic Speech Recognition)
BE	Běloruština
BG	Bulharština
BS	Bosenština
CNR	Černohorština
CZ	Čeština
DNN	Hluboké neuronové sítě (Deep Neural Networks)
DOM	Objektový model dokumentu (Document Object Model)
ERR	Vyhodnocovací metrika chybovosti (Error Rate)
ES	Španělština
FM	Fakulta Mechatroniky, informatiky a mezioborových studií
G2P	Převod grafémů na fonémy (Grapheme to Phoneme)
GMM	Vícemodální gaussovský model (Gaussian Mixture Model)
HTK	Nástroj pro tvorbu skrytých markovských modelů (Hidden Markov Model Toolkit)
HTML	Značkovací jazyk por webové stránky (Hypertext Markup Language)
HR	Chorvatština
IPA	Mezinárodní fonetická abeceda (International Phonetic Alphabet)
LID	Identifikace jazyka (Language Identification)
LM	Jazykový model (Language Model)
LV	Lotyština
LVCSR	Rozpoznávání spojité řeči s velkou slovní zásobou (Large Vocabulary Continuous Speech Recognition)
MFCC	Kepstrální příznaky řeči (Mel-Frequency Cepstral Coefficients)
MK	Makedonština
ML-ASR	Multilingvální systém rozpoznávání řeči
OOL	Části promluvy v cizím jazyce (Out of Language)
OOV	Slova mimo slovník (Out of Vocabulary)
PL	Polština
PREC	Vyhodnocovací metrika přesnosti (Precision)
REC	Vyhodnocovací metrika senzitivity (Recall)
ReLU	Aktivační funkce neuronových sítí (Rectified Linear Unit)
RU	Ruština
SK	Slovenština
SL	Slovinština
SQ	Albánština
SR	Srbština
TUL	Technická univerzita v Liberci
UK	Ukrajinština
WER	Metrika vyhodnocování přesnosti rozpoznávání (Word Error Rate)
XML	Rozšiřitelný značkovací jazyk (Extensible Markup Language)
YLD	Vyhodnocovací metrika výtěžnosti (Yield)

Seznam obrázků

Seznam tabulek

Úvod

Systémy automatického rozpoznávání řeči (angl. Automatic Speech Recognition systems, ASR) slouží k převodu signálu mluvené řeči do podoby vhodné pro další zpracování počítačovými programy. V případě diktovacích, přepisovacích či překladových systémů má výstup podobu textu, ale například u různých hlasově ovládaných aplikací to mohou být interní symboly či příkazy, které jsou pak převáděny na příslušné akce.

Počátky výzkumu v této oblasti sahají do 60. let 20. století a jsou spojeny s rozvojem prvních výkonnějších počítačů. Během dalších dvou dekád došlo k rychlému rozvoji metod efektivní reprezentace řečového signálu pomocí spektrálních (a později kepstrálních) příznaků. Slova se podařilo dekomponovat do malého počtu stavebních jednotek (odvozených od hlásek) a ty reprezentovat pomocí matematických modelů (nejčastěji to byly skryté Markovovy modely) a celé věty pak poskládat na základě pravděpodobnostních přístupů založených nejčastěji na n-gramových modelech [1][2][3].

Díky tomu bylo možné již na začátku 90. let představit první komerční programy určené zejména pro diktování do počítače. Ty ještě spoléhaly na vstřícný přístup ze strany uživatele. Avšak s rozvojem dalších robustnějších metod se použití rozšířilo i na oblast automatického přepisu televizních a rozhlasových zpráv, přepis jednání (např. v parlamentu) a posléze i na méně kvalitní záznamy např. telefonních hovorů [2][3].

Ve stejné době se objevily i první dialogové systémy, v nichž byla, kromě rozpoznávání, použita i hlasová syntéza. V současné době se se systémy rozpoznávání řeči setkáváme v mnoha mobilních aplikacích, třeba v rámci hlasového vyhledávání (např. VoiceSearch od Googlu), u hlasových asistentek (např. Siri od firmy Apple) nebo konverzačních a chatovacích programech (např. Alexa od Amazonu). Většina těchto aplikací výrazně šetří čas uživatele, neboť hlasová interakce je mnohem rychlejší a přirozenější než práce s klávesnicí, myší a obrazovkou. Pro osoby s některými typy tělesného postižení se navíc jedná o jedinou možnost, jak používat moderní techniku [2][3].

Automatické zpracování mluvené (i textové) podoby jazyka má ale jednu specifickou vlastnost - je závislé právě na daném jazyku: na jeho písmu a kódování, hláskovém inventáři, na slovníku a výslovnosti, na syntaxi a gramatice, a v nepo-

slední řadě i na společenském a historickém kontextu. Z těchto důvodů byly první systémy rozpoznávání řeči vyvíjeny vždy pro konkrétní jazyk, čemuž se přizpůsobovaly i použité metody a přístupy. Kromě techniků a programátorů byli součástí výzkumných týmů také experti na lingvistiku a fonetiku. Takový výzkum a vývoj byl finančně náročný a v začátcích si ho mohly dovolit pouze velké firmy (např. IBM či Microsoft), či významné akademické instituce, a většinou se soustředil pouze na velké světové jazyky (zejména angličtinu, francouzštinu, španělštinu, japonštinu, apod.)[3].

Teprve později se podařilo lépe vymezit a oddělit části systému závislé na konkrétním jazyku od těch nezávislých, což následně umožnilo efektivnější přenos poznatků a algoritmů (a později dokonce i natrénovaných modelů) do dalších jazyků. Přesto i dodnes platí, že každý jazyk má své specifické vlastnosti, které ovlivňují např. nezbytnou velikost slovníku, převod mezi psanou a vyslovovanou formou slov, vazby mezi větnými členy, formátování, apod.

Je ovšem také pravda, že současné informační technologie a zejména existence internetu s obrovským množstvím veřejně přístupných (textových a mluvených) dat, umožňují, aby se jazykově závislé moduly učily přímo z těchto dat. Moderní metody strojového učení tak do velké míry umožňují nahradit práci jazykových expertů a významným způsobem zkrátit dobu vývoje systému určeného pro konkrétní jazyk. Což zároveň znamená, že lze těmito technologiemi velmi rychle pokrýt i menší jazyky.

Zaměření práce

Tato práce se zabývá výzkumem a implementací metod, které umožňují rychlý vývoj jazykově závislých modulů pro systémy automatického rozpoznávání řeči. Vznikla na pracovišti (Laboratoř počítačového zpracování řeči na Technické univerzitě v Liberci[1]), kde již od poloviny 90. let pracuje tým, který se touto problematikou zabývá. Během dvou desítek let zde byly vytvořeny všechny základní moduly sloužící pro sestavení a provozování systému ASR, který může být nasazen jak v on-line, tak i off-line režimu a hodí se pro zpracování dat buď ze souboru na disku, nebo přímo z mikrofonu či dokonce z internetového streamu (např. televizního či rozhlasového vysílání).

V době, kdy jsem do týmu přišel, byl již systém schopen pracovat s mluvenou češtinou a slovenštinou, a probíhaly práce na zvládnutí polštiny a chorvatštiny. Zatímco vývoj češtiny (a samozřejmě celého rozpoznávacího řetězce) trval více než 10 let, slovenštinu se podařilo zpracovat cca za 3 roky, a u dalších jazyků se už vývoj základní verze pohyboval kolem jednoho roku. V té době byl vytyčen cíl zvládnout během několika let všechny slovanské jazyky s tím, že vývoj každého z nich by neměl trvat více než několik měsíců.

[1] https://www.ite.tul.cz/speechlab/

Bylo proto nutné seznámit se se všemi těmito jazyky, najít jejich společné a zároveň i odlišné rysy, sestavit pravidla pro vytváření textových korpusů, slovníků a jazykových modelů, vytvořit převodníky mezi ortografickou a fonetickou podobou slov, navrhnout a implementovat metody pro automatické získávání trénovacích dat, a to s různou mírou supervize, a v neposlední míře také vytvořit prostředí pro objektivní testování vyvinutých modulů.

Součástí mé práce byl proto návrh, ověřování a základní implementace všech výše uvedených postupů, jakož i tvorba mnoha pomocných nástrojů, bez nichž by se výzkum a vývoj neobešel, ať už se jednalo například o programy pro hromadné stahování textových a akustických dat, nástroje na jejich analýzu a automatické zpracování, tvorbu a optimalizaci fonetických inventářů a výslovnostních generátorů, moduly pro zpracování čísel a zkratek, až po finální natrénování akustických a jazykových modelů pro každý jazyk.

Zároveň je však třeba říci, že jsem se nemusel zabývat vývojem těch částí rozpoznávacího systému, které jsou jazykově nezávislé. Měl jsem tedy k dispozici již hotové moduly zpracovávající akustický signál a transformující ho na příznakové vektory a dále pak velmi efektivně pracující dekodér převádějící sekvence příznakových vektorů na textový výstup. Tyto klíčové části systému navržené jinými členy týmu jsem tak mohl využívat pro svou práci, což ji výrazně urychlilo, na druhou jsem je musel používat v takové podobě, v jaké byly naimplementovány, bez možnosti do nich zasahovat, což někdy určovalo volbu mých metod a přístupů.

V okamžiku, kdy byl splněn vytyčený úkol zaměřený na slovanské jazyky, jsem se pokusil ověřit, zda mnou navržené metody a vytvořené nástroje jsou použitelné i na jiné jazyky, zejména takové, které nejsou se slovanskými příbuzné a u nichž už nelze použít ani základní míru porozumění mluvenému a psanému slovu. Pro tuto část jsem vybral 3 jazyky z různých částí Evropy: španělštinu (jakožto příklad světového jazyka s velkými datovými zdroji), lotyštinu (příklad malého neslovanského jazyka) a albánštinu (velice specifický jazyk, který podle dostupné literatury dosud nebyl nikým řešen). Výsledky použití mnou navržených metod jsou popsány v předposlední kapitole práce.

Motivace výzkumu a vazba na praxi

Moje práce byla součástí dvou velkých výzkumných projektů řešených na pracovišti. Oba byly financovány Technologickou agenturou České republiky. Jednalo se o tyto projekty:

- TA04010199 „MULTILINMEDIA - Multilingvální platforma pro monitoring a analýzu multimédií“ (2015-2017),
- TH03010018 „DeepSpot - Multilingvální technologie pro detekci a včasné upozornění“ (2018-2021).

Hlavním cílem obou projektů bylo zvládnout přepis a následnou analýzu televizních, rozhlasových a internetových pořadů ve 13 slovanských jazycích, a to češtiny, slovenštiny, polštiny, ruštiny, ukrajinštiny, běloruštiny, slovinštiny, chorvatštiny, srbštiny, bosenštiny, černohorštiny, makedonštiny a bulharštiny. V současné době je již většina z nich předána partnerovi projektu, firmě Newton technologies, a.s., která využívá rozpoznávací systém a jeho jazykové moduly v rámci on-line monitoringu několika desítek stanic provozovaných v těchto zemích.

1 Interdisciplinární základy a východiska

Výzkum a vývoj v oblasti počítačového zpracování řeči má víceoborový charakter. Kromě technických a přírodovědných disciplín, jako jsou akustika, zpracování signálů, matematické modelování, teorie rozhodování, či strojové učení, hrají významnou roli také poznatky ze společensko-vědních oborů, zejména lingvistiky a fonetiky. U systémů, které mají pracovat ve vícejazyčném prostředí, je tato role ještě mnohem důležitější. V této kapitole proto budou krátce představeny základní poznatky, terminologie a postupy z těch oblastí, na nichž bude v dalších kapitolách stavěno.

1.1 Evropské jazyky

Většina jazyků používaných v Evropě patří do rodiny indoevropských jazyků. Ze 740 miliónů obyvatel Evropy přibližně 94 % mluví indoevropským jazykem. Největšími jazykovými skupinami jsou germánské, románské a slovanské. Těmi dohromady mluví 90 % evropské populace a každá skupina má přes 200 milionů rodilých mluvčích. Menšími skupinami jsou helénské, baltské, albánské, indoárijské či keltské jazyky [4].

Přibližně 45 miliónů obyvatel Evropy mluví neindoevropskými jazyky. Nejpočetnějšími skupinami z nich jsou uralské a turkické jazyky. Menšími jsou např. baskičtina či kavkazské jazyky. S přibývající imigrací také narůstají počty mluvčích různých asijských a afrických jazyků, z nichž nejvíce zastoupenou je arabština [4].

V průběhu minulého století převzala v Evropě angličtina status Lingua franca, a stala se tak hlavním komunikačním jazykem v různých oblastech vědy, techniky, mezinárodního obchodu či diplomacie. Dnes je používána a vyučována prakticky ve všech evropských zemích. Nicméně v zemích bývalého Sovětského svazu stále ve velké míře figuruje používání ruštiny. Navzdory rozšířenosti angličtiny, Evropská unie propaguje přístup úřední komunikace ve všech jazycích unie.

Standardním a nejrozšířenějším písmem užívaným v Evropě je dnes latinka, druhá je cyrilice používaná ve východních státech Evropy a jako třetí je řecké písmo používané v Řecku a na Kypru.

1.2 Textová podoba jazyka

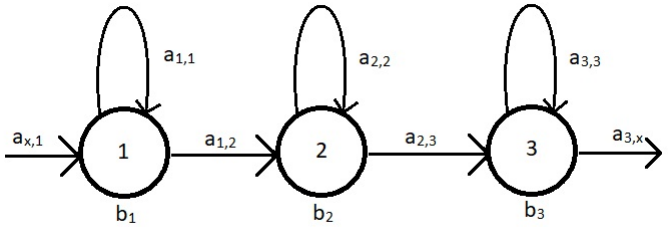
U každého jazyka rozlišujeme podle [5],[6] a [7] mluvenou a písemnou podobu. Mluvená je historicky starší a mnohem variabilnější, písemná prošla dlouhým procesem standardizace a kodifikace. Základní významovou jednotkou psané formy je *slovo*. Slova sestavená do *vět* pak vytvářejí sdělení. Slova jsou zapisována pomocí *znaků* dané *abecedy* (budeme je označovat též jako *grafémy*), které dávají slovu jeho *ortografickou* podobu. V ohebných jazycích jsou slova modifikována pomocí *morfologických* pravidel, která ze základní formy (nazývané *lemma*) vytvářejí odvozené slovní formy. Sestavování smysluplných a srozumitelných vět ze slov se řídí pravidly *gramatiky* daného jazyka.

V současné době velkého rozmachu informačních technologií je důležitým nástrojem pro analýzu a zpracování textů v daném jazyce textový *korpus*. Myslí se tím velmi rozsáhlý a dostatečně reprezentativní soubor textů shromážděných z mnoha různorodých zdrojů. Jeho statistickým zpracováním lze sestavit seznam nejčastěji používaných slov a vytvořit tak reprezentativní *slovník* (nazývaný též *lexikon*), díky kterému je možné dosáhnout požadované úrovně pokrytí psaných (a do velké míry též mluvených) textů. Statistickými nástroji lze vyjádřit též vztahy mezi slovy ve větách, a to na základě četnosti jejich výskytů v rámci za sebou jdoucích slovních sekvencí. Takto popsaný mezislovní kontext se označuje jako *jazykový model*.

1.3 Zvuková podoba jazyka

Podobně jako v psané formě, i u mluvené podoby je dle [8] a [9] základní významovou jednotkou slovo. To je sestaveno z *hlásek* daného jazyka (v této práci je budeme též označovat jako *fonémy*). Hlásky jsou zvuky tvořené činností vokálního traktu (zejména hlasivkami, nastavením ústní a případně i nosní dutiny, jazykem a rty). Hlásky se dělí na *samohlásky* (charakterizované nepřerušovaným proudem zvuku vytvořeného hlasivkami) a *souhlásky*, jejichž charakter je časové proměnný, obsahuje šumovou složku a je ovlivněn překážkami v hlasovém traktu. Hlásky lze uspořádat do skupin podle charakteru zvuku nebo místa, které zvuk určuje, např. samohlásky otevřené a uzavřené, sykavky, nosové hlásky nebo třeba explozivy. Tyto skupiny pak mohou hrát důležitou roli při trénování sdílených hláskových modelů nebo při přenosu již hotových modelů z jednoho jazyka do druhého.

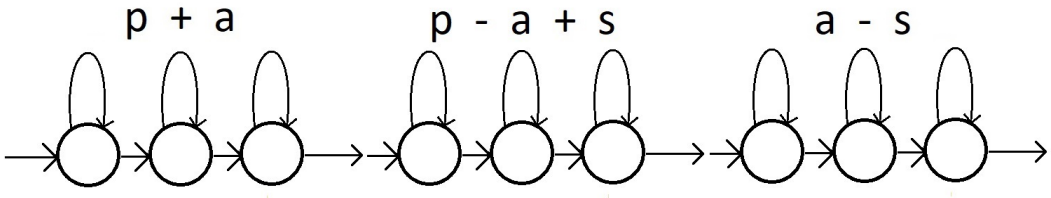
Charakter jednotlivých hlásek se dále liší v závislosti na okolním kontextu, kdy např. hláska /*a*/ zní jinak (a má tudíž i jiný spektrální průběh), vyskytuje-li se před ní (či za ní) sykavka, nosovka či explozíva. Fonetici tyto kontextové varianty označují jako *alofony*. Při počítačovém zpracování se tato variabilita řeší tím, že modely hlásek bývají vícestavové (nejčastěji třístavové) a jednotlivé varianty se modelují a trénují zvlášť jako tzv. *trifony*, tj. kontextově závislé modely s různým pravým a levým okolím [1][10].



Obrázek 1.1: Třístavový model hlásky

Zvuková podoba řeči má mnohem větší míru variability než textová. Kromě již zmíněné kontextové variability hlásek hraje (dokonce ještě větší) roli osoba řečníka a způsob, jakým mluví a jak vyslovuje. Dalším důležitým faktorem je též prostředí, kde se řeč odehrává, které může značným způsobem ovlivnit nasnímaný signál. Jediným způsobem, jak alespoň částečně eliminovat vliv těchto faktorů na systém rozpoznávání řeči, je vytvořit dostatečně robustní *akustický model* všech hlásek a jejich variant natrénovaný na velmi rozsáhlém souboru nahrávek pořízených od tisíců různých osob a v různých situacích a akustických podmínkách. V posledních dvou dekádách se nejčastěji používaným typem stal skrytý Markovův model (Hidden Markov Model, HMM). U hlásek má nejčastěji třístavovou levo-pravou strukturu (viz obrázek 1.1) se dvěma typy parametrů: a) přechodovými pravděpodobnostmi mezi stavy a b) stavovými výstupními pravděpodobnostmi. Výstupní pravděpodobnosti bývají reprezentovány buď směsmi gaussovských rozložení (Gaussian Mixture Model, GMM) nebo, v současnosti mnohem častěji hlubokými neuronovými sítěmi (Deep Neural Networks, DNN). Mluvíme pak o akustickém modelu typu GMM-HMM nebo DNN-HMM). HMM libovolného slova se sestaví jednoduchým zřetězením příslušných hláskových (trifonových) modelů (viz obrázek 1.2)[1][10].

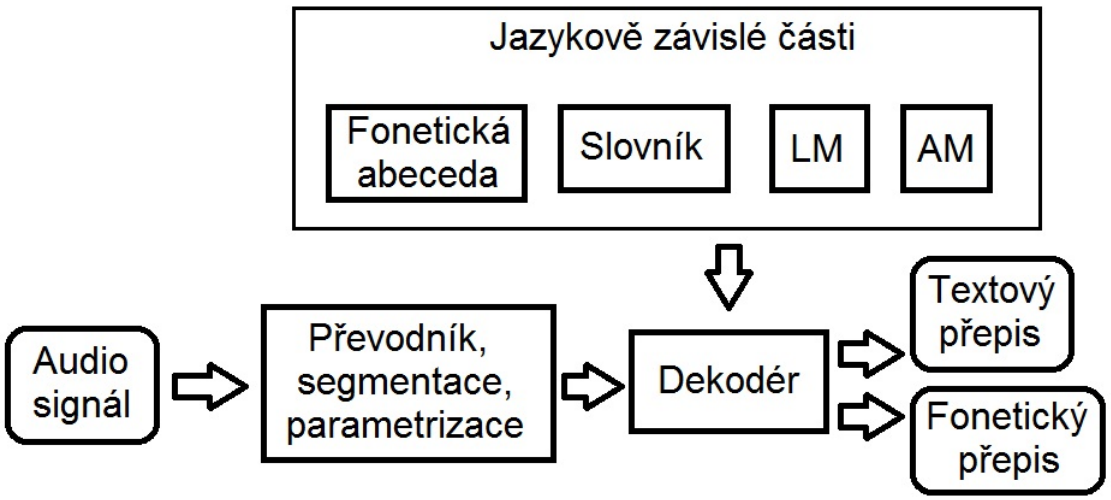
Je třeba ještě zmínit to, že v akustickém signálu nemusí být pouze řeč, ale též další zvuky, které je třeba též vzít v úvahu. Nejčastěji jsou těmito neřečovými událostmi ticho, hluk okolí, nádech, či třeba zvuk vydávaný řečníkem při váhání ('hm'). Tyto zvuky označujeme společným názvem „hluky", modelujeme je stejnou třístavovou strukturou a včleňujeme je do akustického modelu.



Obrázek 1.2: Trifónový model slova "pas"

1.4 Modulární systém rozpoznávání řeči

Systém rozpoznávání řeči v klasické podobě (a též v podobě používané v této práci) je vyobrazen na obrázku 1.3. Na jeho vstup přichází zvukový signál a na výstupu se s určitým zpožděním objevuje přepsaný text.



Obrázek 1.3: Schéma modulárního systému rozpoznání řeči

Systém se skládá z několika modulů. První z nich segmentuje signál do tzv. *rámců* (budeme používat i anglický výraz *frame*) dlouhých obvykle 25 ms, a v každém z nich je vypočteno P *spektrálních příznaků*. Ty jsou buď přímo použity při dekódování (v DNN-HMM systému), nebo jsou ještě převedeny na *kepstrální příznaky* (v GMM-HMM systému).

Dekodér postupně zpracovává sekvence těchto příznakových vektorů tak, že počítá pravděpodobnosti, s jakými by byly vygenerovány jednotlivými hláskovými variantami reprezentovanými akustickým modelem. Na základě toho pak průběžně počítá pravděpodobnosti, že se v blízkém rozmezí framů objevuje některé slovo ze slovníku. Zároveň vyhledává nejpravděpodobnější sekvence slov jdoucích za sebou, a to na základě kombinace skóre vzešlého z akustického a jazykového modelu.

Na výstup může být kromě cílového ortografického přepisu poslán i fonetický přepis, který hraje velmi důležitou roli v případě, kdy rozpoznávací systém slouží pro přípravu budoucích trénovacích dat, jak bude ukázáno v kapitole 6.

Úkolem dekodéru je najít takovou sekvenci slov W, která má největší pravděpodobnost $P(W|X)$ pro pozorovanou sekvenci příznaků X a dané modely. S využitím Bayesovy věty je výpočet pravděpodobnosti definován jako:

$$P(W|X) = \frac{p(X|W) \cdot P(W)}{p(X)} \tag{1.1}$$

kde $p(X)$ je apriorní pravděpodobnost pozorování sekvence příznakových vektorů X a $p(X|W)$ je pravděpodobnost, že pro danou sekvenci slov W bude pozorován příznakový vektor X. Posledně zmíněný člen je nazýván *akustický model*. $P(W)$ je poté apriorní pravděpodobnost pozorování sekvence slov W nezávisle na příznakovém vektoru X a je nazývána *jazykový model*.

Dekodér se pak snaží nalézt:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{p(X|W) \cdot P(W)}{p(X)} \tag{1.2}$$

Prostor všech možných slov, ze kterých se může W skládat, je definován *slovníkem*. Ten také obsahuje všechny přípustné výslovností varianty pro každé slovo.

Modul zpracování signálu a dekodér tvoří jazykově nezávislou část rozpoznávacího systému. Slovník (doplněný o výslovnosti všech slov), seznam přípustných hlásek, akustický model a jazykový model jsou naopak ty části, které jsou jazykově závislé, a pro každý jazyk musí být předem vytvořeny či natrénovány. Způsob efektivní tvorby těchto jazykově závislých modulů je hlavním tématem této práce.

Tabulka 1.1 zobrazuje základní parametry rozpoznávacího systému vyvinutého na pracovišti a použitého v rámci této práce.

Tabulka 1.1: Parametry použitého systému rozpoznávání řeči

Vzorkovací frekvence	16 kHz
Délka rámce	25 ms
Rámcová frekvence	100 hz
GMM příznaky	39 dim. MFCC
GMM počet komponent	32
DNN příznaky	39 dim. Log filter banks
DNN architektura	dopředná pětivrstvá síť
DNN počty neuronů	1024-768-768-512-512
DNN aktivační funkce	ReLU

1.5 Učící se strategie použité v této práci

První systémy rozpoznávání řeči byly do velké míry postaveny na pravidlech dodaných tvůrci systému. Jednalo se tehdy o tzv. přístup řízený expertem (angl. expert-driven approach). Vývoj takového systému byl velmi pracný, vyžadoval znalce z různých oborů a při velkém množství pravidel byl velmi nepřehledný a těžko modifikovatelný.

Současné systémy naopak získávají a průběžně zdokonalují své rozhodovací schopnosti na základě automatické analýzy dodaných dat (data-driven approach). Tento přístup je z hlediska lidské práce mnohem efektivnější, vyžaduje však obrovské množství dat, z nichž velká část musí být předem označkována. Například zvukové nahrávky určené pro trénování akustického modelu musí být doplněny buď fonetickými, nebo alespoň ortografickými přepisy toho, co je v každé nahrávce řečeno. I tuto činnost lze však do velké míry zautomatizovat, jak bude ukázáno v dalších kapitolách.

Tato práce se snaží v co největší míře používat metody a přístupy známé z oblasti strojového učení [11]. V některých případech může jít jen o využití zpětné vazby či jednoduchou optimalizaci řešení problému, v jiných případech jsou použity pokročilejší učící se algoritmy, a to především v případě akustického modelování.

Přístupy k učení systémů se dají rozdělit do tří základních kategorií: supervizované učení (supervised learning), lehce supervizované učení (lightly supervised learning), a nesupervizované učení (unsupervised learning). Všechny tři strategie jsou v této práci aplikovány, a to například:

- klasické supervizované učení při běžném trénování akustického modelu v případě, že máme k dispozici zvuková data a k nim přesné fonetické přepisy,
- lehce supervizované učení tvoří klíčovou část postupu při vyhledávání a získávání dat použitelných pro trénování, jak je popsáno zejména v 6 kapitole,
- nesupervizované učení nabízí možnost vytěžovat trénovací data, u nichž neexistují přepisy, a to v případě, že máme k dispozici několik různě nakonfigurovaných rozpoznávacích systémů, u nichž můžeme vytvořit zpětnou vazbu nutnou pro učení na základě míry jejich vzájemné shody.

Metody pracující s pravidly však stále mají své využití, např. při generování výslovnosti pro slova ve slovníku. Speciálně ve slovanských jazycích se výslovnost řídí poměrně jasnými pravidly, takže zde je tento přístup na místě.

1.6 Metriky používané pro vyhodnocování přesnosti rozpoznávání

K vyhodnocení přesnosti rozpoznávání je v této práci používána míra *WER* (Word Error Rate), která využívá metody hledání minimální vzdálenosti na úrovni slov pro zjištění počtu operací potřebných ke změně jednoho řetězce na druhý. Následně je hodnota WER vypočtena pomocí vzorce:

$$WER = \frac{S + D + I}{N} \tag{1.3}$$

Hodnoty S, I a D označují počet záměn (substitucí), vložených slov (inzercí) a odstraněných/nerozpoznaných slov (delecí). N označuje celkový počet slov v referenčním textu. Výsledná hodnota udává slovní rozdíl obou textů v procentech. Ze vztahu 1.3 také vyplývá, že hodnota WER může nabývat hodnot větších než 100 % z důvodu relativně neomezeného množství možných inzercí.

Další důležitou metrikou je míra slov mimo slovník *OOV* (Out of vocabulary). Ta udává (opět v procentech) počet slov v referenčního textu, která nejsou obsažena ve slovníku rozpoznávacího systému, a tudíž nemohou být správně rozpoznána.

V této práci je zaveden ještě parametr označovaný jako *OOL* (Out of language), který udává míru zastoupení řeči v jiném než cílovém jazyce. Tato metrika je vhodným doplňkem k OOV při analýze úspěšnosti reálných nahrávek, např. u zpravodajských pořadů různých stanic.

2 Současný stav v dané oblasti a existující řešení

V této části bude představen současný stav v oblasti výzkumu a vývoje metod zaměřených na systémy rozpoznávání řeči pracující v multilingválním prostředí. Pro efektivní vývoj takových systémů je třeba vyřešit řadu dílčích úloh, od získání dostatečného množství textových a akustických dat, přes tvorbu slovníků, definování fonetického inventáře a vygenerování výslovnosti pro každé slovo, vytvoření fonetického přepisu ke každé nahrávce, natrénování akustického a jazykového modelu, až po vytvoření prostředí pro objektivní testování hotového systému. Každá z těchto dílčích úloh představuje vlastní výzkumný problém, který bude stručně popsán v následujících sekcích. Vedle popisu existujících řešení se vždy pokusím nastínit i přístup můj (či mého pracoviště) s tím, že jeho podrobnější vysvětlení bude uvedeno v dalších kapitolách.

2.1 Pracoviště zaměřující se na multilingvální rozměr rozpoznávání řeči

Jak už bylo řečeno v úvodu, výzkumné (akademické i firemní) týmy se zpočátku věnovaly hlavně vývoji systémů určených pro svůj vlastní jazyk. Po jeho zvládnutí se pak některé z nich pustily i do dalších jazyků, buď příbuzných, nebo takových, které nabízely významný komerční potenciál. Tímto směrem se ubíral vývoj např. u průkopnických firem v oboru, jako byly IBM [12], [13] či Dragon [14]. V současné době hraje v této oblasti největší roli společnost Google, která má ve svém portfoliu ASR systémů většinu světových jazyků. Podařilo se jí to i díky tomu, že si jako první uvědomila důležitost shromažďování co největšího množství dat různorodého charakteru, tedy i textových a mluvených.

K významným týmům z řad akademických institucí patří např. francouzský tým LIMSI (soustředěný kolem Jean-Luca Gauvaina a Lori Lamel), který již na konci 20. století začal vytvářet jazykové mutace svých systémů pro různé jazyky, a to jak světové [15], tak i tzv. *jazyky s minimálními zdroji* (under-resourced či low-resourced languages). Hledání cest použitelných pro tyto minoritní jazyky je dnes důležitým výzkumným trendem – viz např. [16] a [17].

Dalším významným pracovištěm je KIT na německé Karlsruhe Universität, který je spojen zejména se jménem Tanji Schultz. Ta je považována za jednu z předních průkopnic v tomto oboru, o čemž svědčí i její disertační práce *Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen* z roku 2000 [18]. Nejznámějším příspěvkem jejího týmu je vytvoření standardizované databáze nahrávek v mnoha jazycích nazvané Globalphone [19] – více v sekci 2.7.

Snad je možné říci, že i tým na TUL se dnes řadí ke známějším skupinám v této oblasti, neboť na jeho půdě vzniklo již více než 20 různých jazykových verzí systémů ASR, z nichž je řada využívána i na komerční bázi.

2.2 Významné existující nástroje pro experimentální vývoj systémů ASR

Ve výzkumu rozpoznávání řeči sehrály významnou roli veřejně přístupné platformy navržené pro experimentální vývoj v této oblasti. Ty umožnily zejména menším akademickým týmům získat přístup k funkčním programům a nástrojům, které již nebylo třeba vytvářet od samého počátku. Naopak se daly snadno modifikovat a využívat pro experimentální ověřování nových nápadů a přístupů. Od 90. let 20. století byl v této oblasti nejznámější produkt univerzity v Cambridge známý pod zkratkou HTK (Hidden Markov Toolkit) [20].

V poslední dekádě převzal vedoucí roli v této oblasti systém Kaldi [21], který jako první umožňoval použití neuronových sítí. Obě programové platformy byly a jsou často využívány pro experimentování s multilingválními systémy.

V oblasti jazykového modelování je nejčastěji používaným nástrojem programový balík SRILM z SRI International [22], který se využívá zejména pro tvorbu n-gramových jazykových modelů.

Na našem pracovišti používáme vlastní rozpoznávací systém zmíněný v kapitole 1. Nicméně pro trénování akustických modelů typu GMM je využíván program HERest z prostředí HTK a pro DNN modely zase nástroje platformy Torch[1]. Jazykové modely jsou vyvíjeny pomocí vlastních programů, které budou zmíněny v kapitole 5.

2.3 Textové korpusy a práce s textovými daty

Textový korpus je základem pro tvorbu slovníku a jazykového modelu. Je potřeba, aby byl tvořen rozmanitými texty v dostatečném množství (řádově stovky MB až jednotky GB), které dostatečně zastupují rozpoznávaný jazyk, případně konkrétní

[1] http://torch.ch/

oblast určenou k rozpoznávání. Existuje mnoho textových korpusů, volně nebo komerčně dostupných. Jedním z nejznámějších distributorů je asociace ELRA[2] (Europen Language Resources Association), která zastupuje okolo 50 světových organizací a zajišťuje prodej a distribuci jejich řečových a textových korpusů. Nicméně ceny jednotlivých nabízených korpusů se mohou vyšplhat i do tisíců eur a jejich užití je vždy spjato s určitými podmínkami.

Další možností jsou národní korpusy jednotlivých zemí, jako je např. Český národní korpus[3] [23]. Ty jsou ale většinou tvořeny s velkým podílem lidské práce, protože obsahují i další lingvistické údaje. I z těchto důvodů bývají pro účely vývoje systémů rozpoznávání řeči příliš malé. Navíc mohou být většinou používány opět pouze s určitými omezeními.

Existují i multilingvální korpusy obsahující stejné texty v různých jazycích, které jsou většinou využívány pro úlohu strojového překladu. Příkladem mohou být korpus dokumentů Evropského parlamentu *Europarl Corpus* a korpus právních dokumentů Evropské unie *Eur-Lex Corpus*[4] [24], které obsahují velké množství různých dokumentů v úředních jazycích Evropské unie. Nicméně texty jsou úředního a právnického charakteru a nemusí tak být příliš vhodné pro modelování běžné řeči, spíše jsou vhodné jako doplněk.

Dalším příkladem multilingválních korpusů je *Parasol*[5] [25]. Korpus je sestaven z originálů a překladů klasické beletrie ve slovanských a několika dalších evropských jazycích. Texty jsou automaticky paralelně zarovnány a navíc mohou obsahovat i další manuálně přidané lingvistické informace. Klasická beletrie nicméně příliš nezastupuje každodenní spontánní řeč a množství také není dostačující pro tvorbu robustních modelů.

Zmínit lze i otevřené multilingvální databáze jako např. *Opus*[6][26] či *Tatoeba*[7]. První zmíněná databáze se zaměřuje na sběr volně dostupných textů na internetu a jejich automatickou syntaktickou a morfologickou anotaci. Druhá databáze je založená na komunitní bázi a jednotliví registrovaní uživatelé mohou přidávat, překládat či kontrolovat jednotlivé věty, kterých je v současné době přes 8 milionů v 369 jazycích.

Kromě korpusů existují i již předpočítané n-gramové modely, např. *Web1T* společnosti Google [27]. Modely jsou spočítány až do úrovně 5-gramů na knihách z databáze Google Books. Může na nich tak být vypočítán jazykový model bez nutnosti tvorby korpusu. Nicméně, jak bylo například ukázáno v [28], systém pro češtinu vyvinutý na TUL využívající jazykový model natrénovaný na vlastním textovém

[2] http://www.elra.info/
[3] https://www.korpus.cz/
[4] https://www.sketchengine.eu/eurlex-corpus/
[5] http://parasolcorpus.org/
[6] http://opus.nlpl.eu/
[7] http://tatoeba.org

korpusu vykázal o 10 % lepší výsledky než systém využívající n-gramové modely Web1T.

Výše zmíněné možnosti se ale netýkají všech jazyků a většinou jsou k dipozici pouze pro velké a větší světové jazyky. Poslední možností tedy je vytvoření vlastního textového korpusu svépomocí z volně dostupných dat na internetu. Nejvhodnějšími zdroji jsou různé internetové noviny, televize a rádia, kde je obsah přidáván každý den a reflektuje tak současné dění. To může být podstatné v úloze online monitoringu televizního a rozhlasového vysílání. Dalšími zdroji jsou např. znalostní databáze jako wikipedia, parlamentní archivy, různé blogy a podobně. Nasbírané texty musí být následně důkladně zpracovány.

Přístupy k vytěžování volně dostupných online zdrojů jsou popsány například v [29] nebo [30], kde byl vytvořen univerzální přístup a nástroje pro automatické stahování textů z webových stránek a jejich zpracování. Naproti tomu postup popsaný v této práci v 5. kapitole není až tak obecný, zaměřuje se na přesnost a využívá multingválních závislostí mezi jazyky pro zefektivnění celého procesu.

2.4 Jazykové modely

Jazykový model poskytuje odhad pravděpodobnosti sekvence rozpoznaných slov. Nejčastější přístup je pomocí statistického modelování za využití slovních n-gramů (bigramy, trigramy a další). Pravděpodobnosti v n-gramových jazykových modelech jsou běžně určovány pomocí maximálního odhadu věrohodnosti. To činí rozložení pravděpodobnosti závislé na trénovacích datech, a proto je vyžadováno co největší množství těchto dat. Jazykový model je pak vypočítán z textového korpusu pro všechna slova ve slovníku.

Pro některé tvaroslovně bohaté jazyky existují přístupy morfologické dekompozice slov na sublexikální jednotky (kořeny slov, přípony, předpony), které jsou pak použity jako slova ve slovníku a v jazykovém modelu. Tento přístup značně zmenšuje velikost slovníku a zvyšuje pokrytí slov daného jazyka. Na druhou stranu zde mohou vznikat problémy s dekódováním řeči jako např. fonetickou nejednoznačnost jednotlivých morfémů, tvoření celých slov z rozpoznaných částic a také nutnost většího n-gramového kontextu (5-gramy až 10-gramy) pro zachycení gramatických vztahů. Tento přístup může být využit dvěma způsoby. Prvním z nich je morfologicko-gramatický přístup, kdy je předem jasný způsob dekompozice slov, což ale vyžaduje hlubší znalosti daného jazyka a připravená data. Druhou možností je statistický přístup, který stojí jen na analýze textu a nevyžaduje tak hlubší lingvistické znalosti. Může zde ale dojít k situaci, kdy jsou slova rozdělena spíše na pseudo-morfémy. Na druhou stranu může být tato metoda využita pro jakýkoliv jazyk. Statistický přístup byl aplikovám například pro slovinštinu za využití statistických modelů [31], pro češtinu za využití rozhodovacích stromů [32] nebo pro turečtinu za využití konečných automatů [33].

V případě některých jazyků, které mají relativně volný slovosled (např. slovanské jazyky), je pro jazykové modelování využíváno syntaktické analýzy. Standardní n-gramové modely vyššího řádu mohou mít vysokou perplexitu a nižší úspěšnost, a proto vyžadují enormní množství trénovacích dat. Některé práce tedy využívají syntaktické informace společně se statistickým jazykovým modelováním jako např. [34] v případě ruštiny, [35] využívající strukturované jazykové modely pro angličtinu, nebo [36] s morfosyntaktickým zpracováním modelů pro francouzštinu. Tyto přístupy vždy zvyšují rozpoznávací přesnost systému, avšak opět vyžadují lingvistické informace pro daný jazyk.

V rámci této práce jsem byl odkázán na použitý ASR systém, který pracuje pouze se statistickými bigramovými modely. Ty sice mají horší výsledky oproti trigramovým modelům, nicméně při práci se slovanskými či jinými flektivními jazyky, které mají relativně volný slovosled a velký slovník, není rozdíl tak markantní, jak ukázala interní studie. Naopak využití bigramů snižuje výpočetní náročnost a umožňuje tak systému pracovat v reálném čase. Pro modelování delších mezislovních kontextů mohou být do slovníku přidávány kolokace (častá slovní spojení), čímž se do jisté míry supluje vliv vyššího n-gramového modelu.

2.5 Fonetická stránka jazyka

Pro každý jazyk existují fonologické studie rozlišující jednotlivé fonémy daného jazyka a případně i jeho různých dialektů. Tyto studie jsou ale většinou velmi detailní v definování i podobně znějících fonémů, a proto je potřeba brát je jen jako pomocné vodítko. Naštěstí pro úlohu rozpoznávání řeči není až tak důležité přesné rozlišování hlásek a jejich variant (alofonů) jako například u syntézy řeči.

V počátcích vývoje ASR systémů, které měly k dispozici malé množství akustických dat pro trénování, se vyplatilo pracovat s většími počty fonetických jednotek definovanými i na základě okolního kontextu. To vedlo k využívání velkých fonetických sad a komplikacím na různých úrovních vývoje, jako například velká nevyváženost trénovacích dat pro jednotlivé modely fonémů či problémy při fonetické anotaci. Příkladem může být jedna z prvních řečových databází pro angličtinu *TIMIT* [37], která ve svých anotacích rozlišuje 58 fonémů, 2 přízvuky a 3 typy neřečových událostí (včetně speciální pauzy před explozivou).

S příchodem nových technologií, metod a množství trénovacích dat bylo možno přejít od kontextově nezávislých jednotek (tzv. monofónů) na trifónové modely (případně vyšší), které už v sobě zahrnují i vliv levého a pravého okolí jednotlivých hlásek. To vedlo ke zmenšení fonetické sady a zefektivnění celého systému [38].

Námi použitý přístup lze označit jako technicky pragmatický. V 6. kapitole jsou popsány použité postupy tvorby fonetických sad a výběru fonémů pro jednotlivé

jazyky například pomocí experimentálního ověřování různých hypotéz. Podobným přístupem se řídíme také při výběru vhodného zdrojového jazyka či kombinace více jazyků a mapování jednotlivých sad mezi sebou při adaptaci systému na cílový jazyk, a rovněž i při míchání sad pro účely identifikace jazyka.

2.6 Tvorba výslovnostních slovníků

Výslovnostní slovník je nezbytná součást rozpoznávacího systému a někdy je v literatuře považován i za součást jazykového modelu. Nicméně z hlediska vývoje ASR systémů je vhodné tyto dvě části oddělit. Slovník obsahuje množinu všech slov, které mohou být systémem rozpoznány a zároveň tvoří spojení mezi lexikální a akustickou částí systému. Slova do slovníku jsou vybírána z textového korpusu, na kterém je následně trénován jazykový model. Hlavním kritériem výběru je jejich četnost v korpusu, případně mohou být další důležitá slova dodána ručně. Velikost slovníku závisí na cílovém jazyku. Například u analytických jazyků, jako je angličtina nebo španělština, obvykle stačí slovník o velikosti do sto tisíc slov pro pokrytí většiny slovní zásoby. U slovanských jazyků, které jsou vysoce flektivní, se ukazuje jako nezbytné množství v řádu stovek tisíc slov.

Výslovnost jednotlivých slov byla v počátcích tvořena ručně odborníky na daný jazyk. To je při současném množství slov ve slovníku nemožné a úlohu je potřeba provádět automaticky. K tomu jsou využity různé metody tzv. G2P (Grapheme-to-Phoneme) konverze, které mohou využívat přesně daná produkční pravidla, mohou být reprezentovány stavovými automaty, nebo využívat metod strojového učení, například na základě neuronových sítí.

V některých případech nemusí být takový převod jednoduchý, nebo naopak je až tak přímý, že není zapotřebí fonetické transkripce. Vždy záleží na ortografické hloubce daného jazyka, tedy vztahu mezi psanou a mluvenou podobou. Některé práce tak pro akustické modelování využívají grafémy místo fonémů a každý znak slova je tedy využit jako základní jednotka řeči. Tento způsob byl s přijatelnými výsledky použit např. pro ruštinu [39], vietnamštinu [40] nebo pro více jazyků zároveň (angličtinu, němčinu a španělštinu) [41]. Nicméně tento přístup nikdy nedosahuje kvality výsledků systémů využívajících fonémy, pouze odstraňuje řešení problémů s fonetickou transkripcí.

Jedním ze způsobů tvorby výslovností je využití již existujících výslovnostních databází či volně dostupných dat, jako například využití otevřeného slovníku Wiktionary[8] aplikovaného v [42]. Autoři zde využívají skutečnosti, že Wiktionary obsahuje fonetické transkripce slov v mezinárodní fonetické abecedě (IPA). Tímto způsobem lze získat i výslovnosti pro jména, názvy měst a další slova, jejichž výslovnost není pro daný jazyk standardní a často neodpovídá G2P pravidlům. Tento způsob je ale

[8] https://www.wiktionary.org/

uplatnitelný jen pro ty jazyky, které mají v databázi Wiktionary dostatek dat.

Způsob využívající předem daná produkční pravidla vyžaduje určitou znalost fonetiky daného jazyka a ruční přípravu pravidel. Nicméně u ortograficky mělkých jazyků, kam patří i slovanské, může být tvorba pravidel poměrně snadná. Navíc zde lze využít určitých obecných principů fonetiky, které jsou často sdíleny napříč jazyky. Systém vytvořený pro češtinu byl popsán v [10] a postup popsaný v kapitole 6 z něho vychází. Podobné systémy byly vytvořeny také pro evropskou portugalštinu [43] s údajnou úspěšností 98,8 % či pro turečtinu [44].

Další práce využívají pro fonetickou transkripci metod strojového učení, kdy se systém na určité množině slov s jejich fonetickými přepisy sám naučí vztahy mezi ortografickou a fonetickou podobou. Takový systém byl nasazen například pro francouzštinu [45] za využití metod statistických strojových překladů (SMT), kde systém Moses dosáhl lepších výsledků než standardní G2P převod využívající předem daná pravidla. Dále byl například v [46] popsán přístup využívající LSTM rekurentních neuronových sítí, který dosáhl úspěšnosti 78,7 % pro angličtinu, či systém využívající též LSTM sítí v kombinaci s konvolučními vrstvami popsaný v [47] s dosaženými výsledky přes 90 % pro angličtinu, češtinu a ruštinu. Tento přístup nicméně vyžaduje již nějaké množství výchozích dat pro trénování, a nemůže tak být využit v počáteční fázi, kdy žádná data k dispozici nejsou.

Postup popsáný v 6. kapitole využívá kombinaci přístupu s předem stanovenými pravidly a strojového učení. Zároveň je zde popsáno využití podobností mezi jazyky.

2.7 Zdroje dat pro akustický model

Zatímco obstarat textová data je v dnešní době internetu jednoduché, s vhodnými akustickými daty je situace složitější. Pro vytvoření akustického modelu je potřeba alespoň několika hodin nahrávek řeči společně s jejich co nejpřesnějšími fonetickými přepisy. Řada subjektů dnes nabízí hotové řečové korpusy, ale za nemalou cenu. Příkladem je již zmíněný tým na Karlsruhe Universität a jejich databáze *GlobalPhone* [48], který během více než 10 let shromáždil řečová data pro 22 jazyků, přičemž pro každý je k dispozici okolo sta různých vět namluvených přibližně sto mluvčími.

Mnoho dalších řečových korpusů nabízí již zmíněná asociace ELRA. Cena takových korpusů se vždy odvíjí od kvality zpracování, rozmanitosti mluvčích, prostředí a dalších faktorů. Tyto korpusy jsou ale většinou tvořeny pro velké a středně velké světové jazyky, a tak může být problém nalézt data v dostatečném množství pro jazyky s řádově několika miliony mluvčími.

Existují i crowd-sourcingové projekty jako *Amazon Mechanical Turk* [49], kde jsou řečové nahrávky připraveny na serveru a lidé mohou vytvářet přepisy těchto nahrávek za určitý finanční obnos, či výzkumné projekty jako [50] a [51], kde byla

v obou případech vytvořena a využita aplikace pro chytré telefony. Ta dobrovolníkům zobrazuje text, který mají přečíst a nahrávka je následně odeslána na server společně s dalšími informacemi, které dobrovolníci vyplní. Dále existují projekty jako například VoxForge[9], kde dobrovolníci nahrávají krátké věty v různých jazycích. Zmínit je potřeba i projekt Librivox[10], kde lidé předčítají knihy v mnoha různých jazycích.

Vždy je ovšem možnost vytvořit si takový řečový korpus svépomocí, což vyžaduje hodně úsilí a zdrojů. Na vytvoření dostatečně robustního akustického modelu je zapotřebí, aby korpus obsahoval co nejvíce nahrávek od co nejvíce různých mluvčích obou pohlaví, v různých věkových kategoriích a nejlépe v různých akustických podmínkách. Sehnat dostatečné množství lidí pro nahrávání může být problematické a nakonec stejně nákladné jako koupě již hotového korpusu. Navíc při tvorbě systému pro cizí jazyky z jiných zemí může být velmi obtížné sehnat rodilé mluvčí.

Kromě ogranizovaného nahrávání je dalším způsobem automatická tvorba vlastních trénovacích dat za využití zdrojů obsahujících nahrávky řeči a přidružený text, jako jsou např. audioknihy, pořady s titulky, případně přepisy z jednání parlamentu a podobně. V tomto případě je využito lehce supervizované učení spočívající v aplikaci existujícího ASR systému k zarovnání textu k nahrávce, včetně detekce případných neshod, které jsou následně odstraněny. Takto získaná data jsou využita k trénování nového systému. Tyto postupy jsou iterativní, kdy v každé iteraci jsou vytvářena nová trénovací data, z nichž je natrénován nový vylepšený model, který je následně použit v dalším kroku.

Takový způsob byl uplatněn například v [52] pro získání trénovacích dat z anglických audioknih. Texty byly časově zarovnány k audio nahrávce, rozděleny na věty, které byly následně automaticky přepsány a přepis porovnán s textem. Pokud došlo ke shodě, byl úsek vyříznut a použit pro trénování. Podobný způsob byl aplikován i pro čínský ASR systém těžící z televizního vysílání s titulky [53] nebo pro jihoafrickou angličtinu za využití pořadů rádiového vysílání s přesnými přepisy [54]. Tyto práce využívaly již existující systém pro zpracovávané jazyky či jim velmi blízké. Také další práce se zabývají tímto způsobem tvorby dat za využití existujícího systému pro jiný jazyk. Například v [55] byla použita angličtina jako tzv. startovací (boostrapping) jazyk pro vývoj telefonního dialogového systému pro tamilštinu.

Některé práce využívají k tvorbě trénovacích dat pouze audio nahrávky a k nim za pomoci existujícího systému a iterativního procesu vytvářejí fonetické přepisy (za využití nesupervizovaného učení). Takový způsob byl například aplikován při vývoji polského systému pro zpracování záznamů Evropského parlamentu za využití španělštiny jako startovacího jazyka [56]. Jednotlivé nahrávky jsou zde rozpoznány a k rozhodnutí, zda budou použity jako trénovací data, slouží důvěryhodnost (confidence measure) vypočtená dekodérem a stanovený práh, který tato hodnota

[9] http://voxforge.org
[10] htttps://librivox.org

musí překročit. Podobný způsob je využit i v [57], kde jsou k vytvoření systémů pro vietnamštinu použity systémy pro 6 evropských jazyků. Přepis je následně přijat na základě shody těchto systémů a hodnoty důvěryhodnosti přepisu.

Tato práce se zaměřuj na získávání trénovacích dat ze zdrojů, které obsahují audio nahrávky a nějaký doprovodný text. Co je v nahrávce obsaženo, není úplně předem známo a o textu se také neví, jestli obsahuje úplný nebo částečný přepis promluvy v nahrávce, či nějakým způsobem promluvu v nahrávce parafrázuje, nebo obsahuje jen popis toho, co se v nahrávce děje. Příkladem jsou zpravodajské weby, kde jednotlivé zprávy či články obsahují audio či video a k němu přidružený text, který popisuje situaci a případně cituje části promluvy. Cílem je získat co největší množství takovýchto volně dostupných dat, pokusit se najít alespoň nějaké shodující se části a vytěžit z nich trénovací data. Podrobně je tento postup popsán v kapitole 6.

2.8 Akustické modely vytvořené z dat více jazyků

Ve chvíli, kdy jsou k dispozici data pro více jazyků, je možno je využít k tvorbě multilinguálních systémů. Ty pak mohou být využity k různým účelům, a to k identifikaci jazyka, k získávání dat pro nový jazyk, ke zlepšení rozpoznávání a podobně.

Například v [58] a [59] byl vytvořen společný akustický model smícháním dat z jednotlivých jazyků databáze Globalphone. Na základě rozhodovacích stromů byly sloučeny jednotlivé fonémy z různých jazyků a vytvořena tak jedna univerzální sada. Podobně bylo učiněno v [60], kde na základě speciálního rozhodovacího algoritmu byla efektivně redukována fonetická sada společného modelu pro hindštinu a tamilštinu v kombinaci s americkou angličtinou, čímž bylo dosaženo zlepšení rozpoznávání. Podobný přístup byl popsán v [61], kde bylo využito sdílení parametrů GMM v podprostoru mezi stavy jednotlivých modelů fonémů, čímž bylo dosaženo jistého zlepšení přesnosti rozpoznávání pro angličtinu, němčinu a španělštinu.

V [62] je uvedeno několik různých přístupů a schémat, jak využít již existující systémy pro rozpoznávání nového cizího jazyka. Popsáno je zde, jak nasadit existující systém na nový jazyk, pro který jsou k dispozici trénovací data, a dále jak využít trénovací data z jiného jazyka v případě, že pro cílový jazyk není dostatek trénovacích dat. Je zmíněn i simultánní multilingvální systém sloužící pro rozpoznávání nahrávky, kde jazyk promluvy není předem znám.

Multilingvální systémy jsou často využívány i pro identifikaci jazyků. Například v [63] je uveden postup sjednocení identifikace jazyka a rozpoznávání řeči do jednoho procesu, kde jsou využity akusticko-fonetické a lexikální informace z přípustných jazyků pro účinnou identifikaci a následné rozpoznávání. Další práce jako [64] či [65] pro identifikaci jazyka a následné rozpoznávání využívají různé architektury neuronových sítí, kde parametry modelu jsou sdíleny napříč všemi jazyky. Stále více se v této oblasti začíná uplatňovat architektura takzvaných end-to-end systémů, které

jsou popsané v následující části.

Přístupy řešené v této práci v 6. kapitole využívají multilingválních modelů především pro nastartování vývoje systému pro nový jazyk a automatické získávání dat. Kdy jsou v první fázi vývoje využita akustická data pouze pro jeden či více již zpracovaných jazyků, ta jsou dále míchána s nově vytěženými daty pro cílový jazyk, až je nakonec systém plně převeden pouze na cílový jazyk.

2.9 End-to-end ASR systémy

V posledních letech (v souvislosti s rozvojem hlubokých neuronových sítí) dochází k realizaci systémů, které již nejsou tvořeny klasickou sestavou dílčích modulů (uvedenou v sekci 1.4), a naopak celý proces zpracování a dekódování akustických dat se přenechává síti s vhodnou architekturou. Ta na svém vstupu přijímá zvuk a na výstupu generuje rozpoznaný text. Jak je ukázáno například v [66], takový systém může dosáhnout výsledků srovnatelných s klasickým přístupem, a to i pro nahrávky v hlučném prostředí. Nevýhodou je potřeba enormního množství trénovacích dat (nejméně tisíce, spíše však desetitisíce hodin), ze kterých je síť schopna naučit se vztahy mezi mluvenou a psanou formou. Proto jsou takové systémy použitelné zatím spíše pro velké světové jazyky, pro něž není problém shromáždit dostatek dat. Pro menší jazyky, a především pro automatickou tvorbu systémů pro nové jazyky s málo zdroji, kterou se zabývá tato práce, je tento přístup nepoužitelný, a proto zde není této problematice věnována další pozornost.

3 Cíle práce

Hlavním tématem předkládané práce je výzkum a vývoj zaměřený na multilingvální systémy automatického rozpoznávání řeči (ML-ASR), a to zejména na jejich jazykově závislou část zahrnující lingvisticko-lexikální moduly (slovníky a jazykové modely) a akusticko-fonetické moduly (fonetické inventáře, výslovnosti a akustické modely). Práce úzce souvisí s projekty řešenými na školícím pracovišti, což ovlivnilo i stanovení cílů a priorit. Ty lze definovat takto:

- Navrhnout co nejefektivnější přístup k vývoji výše uvedených jazykově závislých modulů, který bude po svém nasazení vyžadovat minimum lidské práce, a to jak expertní (zejména v oblasti lingvistické), tak i manuální (například ve formě přepisů, sluchových kontrol či anotací), a který si vystačí s daty veřejně přístupnými prostřednictvím internetu.

- Navrhnout a implementovat kompletní sadu nástrojů, které pomohou zautomatizovat většinu nezbytných prací a úkonů, počínaje sběrem textových a akustických dat, přes tvorbu výslovnostních slovníků, jazykových a akustických modelů, až po finální podobu automaticky přepsaných textů.

- Prozkoumat a navrhnout možnosti nasazení metod strojového učení zejména v nejkritičtější části vývoje, kterou je získávání a anotace dat pro trénování akustických modelů pro jednotlivé jazyky. Zde se zaměřit především na využití tzv. lehce supervizovaného přístupu, v němž hraje úlohu supervizora vlastní vyvíjený ML-ASR systém.

- Navržené postupy aplikovat na všechny slovanské jazyky, a to s využitím jejich podobných rysů a metod založených na mezijazykovém transferu a adaptaci.

- Při vývoji a získávání dat pro učení se zaměřit na cílovou doménu budoucích aplikací, kterou bude (v souladu se zmíněnými projekty) zejména automatický přepis a monitoring televizních a rozhlasových stanic vysílajících v 13 národních slovanských jazycích. Pro tuto oblast také vytvořit sadu reálných testovacích dat použitelných pro objektivní vyhodnocení přesnosti přepisu a porovnání různých přístupů.

- Ověřit použitelnost navržených metod a postupů na několika dalších vybraných evropských neslovanských jazycích.

4 Slovanské jazyky z pohledu rozpoznávání řeči

V této kapitole jsou popsány slovanské jazyky do konkrétních detailů, které jsou důležité pro vývoj ASR systémů. I přestože může být takový systém vytvořen pouze s minimálními znalostmi daného jazyka, je vždy výhodné seznámit se alespoň s jeho základními rysy a charakteristikami. Při práci s multilingválními systémy je též vhodné pochopit vztahy a podobnosti mezi jednotlivými jazyky a pokusit se tyto znalosti co nejlépe využít.

Slovanské jazyky jsou největší skupinou evropských jazyků. Mluví jimi přibližně 290 milionů rodilých mluvčích (z čehož 150 milionů připadá na ruštinu) a dalších přibližně 130 milionů lidí je používá jako svůj druhý jazyk. Dělí se na tři větve - západoslovanské, východoslovanské a jihoslovanské (přičemž ty mohou být také dále děleny na východní a západní skupinu). Dle [67] se jedná o 17-20 jazyků, z nichž 13 je oficiálními jazyky evropských států a jsou spolu s počty mluvčích vypsány v tabulce 4.1. Dalšími minoritními jazyky jsou například hornolužická a dolnolužická srbština, kašubština, rusínština, slezština a některé další dialekty. Ty jsou většinou používány v menších regionech slovanských i neslovanských států. Počty uživatelů těchto menších jazyků jsou v řádu desítek maximálně stovek tisíc mluvčích. Tato práce se zabývá pouze 13 oficiálními jazyky, ostatní minoritní jazyky a dialekty jsou zde jen zmíněny pro doplnění.

Vzájemná podobnost hraje významnou roli mezi jazyky ve stejné větvi. V každé z nich jsou mluvčí víceméně schopni si navzájem porozumět. Vzájemná srozumitelnost mezi jazyky různých větví už je problematičtější, ale stále do jisté míry možná. Důležité jsou i dialekty, které často tvoří most mezi jednotlivými jazyky a větvemi.

Největším pozorovatelným rozdílem mezi slovanskými jazyky je jejich ortografie. Západoslovanské jazyky a západní skupina jihoslovanských jazyků používají k zápisu latinku, což je historicky dáno někdejším vlivem římskokatolické církve. Zatímco východoslovanské a východní skupina jihoslovanských využívají cyrilici, a to díky vlivu pravoslavné církve.

Následující podkapitoly rozebírají jednotlivé skupiny a jazyky s ohledem na vývoj systémů rozpoznávání řeči, tedy se zaměřením na ortografii, fonetiku a do určité míry i morfologii a gramatiku.

Tabulka 4.1: Přehled slovanských jazyků, zdroj: omniglot.com

Skupina	Jazyk	Kód	Mateřský jazyk pro [mil. lidí]	Komun. jazyk pro [mil. lidí]	Písmo
Západoslovanské	Čeština	CZ	10	11	latinka
	Slovenština	SK	5	6	latinka
	Polština	PL	40	55	latinka
Jihoslovanské	Slovinština	SL	2	2	latinka
	Chorvatšina	HR	4	5	latinka
	Srbština	SR	7	8	lat./cyr.
	Bosenština	BS	2	2	lat./cyr.
	Černohorština	CNR	1	1	lat./cyr.
	Makedonština	MK	2	2	cyrilice
	Bulharština	BG	9	9	cyrilice
Východoslovanské	Ruština	RU	150	260	cyrilice
	Ukrajinština	UK	45	50	cyrilice
	Běloruština	BE	3	3	cyrilice

4.1 Písemná podoba jazyků

Slovanské jazyky, jak už bylo zmíněno, využívají dvě různé abecedy - latinku a cyrilici, navíc každý jazyk s různou vlastní obměnou znaků, díky čemuž mohou být i snadněji identifikovány. Západoslovanské jazyky používají latinku, východoslovanské cyrilici a jihoslovanské jsou rozděleny. Slovinština a chorvatština na západě využívá latinku, bulharština a makedonština na východě používá cyrilici a bosenština, srbština a černohorština uprostřed používá zároveň jak latinku, tak cyrilici.

Latinka i cyrilice jsou hláskové abecedy (tj. obsahují znaky jak pro samohlásky, tak pro souhlásky). Latinka byla původně navržena pro starověkou latinu a později doplňována tak, aby vyhovovala výslovnosti spíše románských jazyků. Její zavedení pro slovanské jazyky, které mají mnohem bohatší fonetický inventář, vedlo k potřebě nutných úprav a doplnění. Chybějící hlásky se zpočátku zapisovaly pomocí spřežek (spojením více písmen), které jsou stále hojně využívány především v polštině. S postupem času se přešlo spíše k používání diakritických znamének, tj. háčků, čárek, střížek, vlnovek, ocásků a dalších, které vyznačují alternativní výslovnost.

Oproti tomu cyrilice byla přímo navržena pro výslovnost slovanských jazyků a obsahuje tak více znaků a trochu odlišný systém zápisu a čtení. Historicky se cyrilice vyvinula na dva typy. První používaný ve východoslovanských zemích a v Bulharsku, který obsahuje tvrdé a měkké samohlásky a další znaky změkčující výslovnost předcházející souhlásky, a druhý typ používaný srbo-chorvatskými zeměmi a v Ma-

kedonii, který využívá spíše diakritiky podobně jako u latinky.

Specifickým znakem jazyků používajících cyrilici je způsob psaní názvů a značek západních společností. Využívají se dvě možnosti, buďto je název přepsán foneticky do cyrilice (např. Microsoft je přepsán na МАЙКРОСОФТ - doslova *Majkrosoft*), nebo je naopak název ponechán v textu v původní formě psané latinkou.

Dále se především u východoslovanských jazyků objevuje časté spojování slov spojovníkem. Většinou jsou spojována frekventovaná slovní spojení jako je například *jiho-západ*, ale zároveň se může používat i varianta bez spojovníku - *jihozápad*.

Ortografie jihoslovanských jazyků je velmi fonemická, tedy snaží se zapisovat slova přesně tak, jak se vyslovují. Zohledněna je tak i ve většině případů spodobu znělosti.

Příloha A obsahuje výpis znaků abeced zpracovaných slovanských jazyků. Abecedy jsou vypsány tak, jak jsou oficiálně uváděny a abecedně řazeny, tedy i se spřežkami, pokud jsou dle nich řazeny.

4.2 Fonetická podoba jazyků

Podle [68] a [69] se slovanské jazyky z hlediska výslovnosti značně liší prozodickými rysy jako je délka hlásek, přízvuk a tón. Čeština, slovenština, srbština, chorvatština a částečně slovinština rozlišují fonologickou délku hlásek, zatímco ostatní jazyky délku nerozlišují. Západoslovanské jazyky mají pevně daný přízvuk, zatímco ostatní jazyky mají přízvuk volný a u východoslovanských a u bulharštiny jeho pozice může i rozlišovat význam slova.

Všechny slovanské jazyky mají podobný výčet souhláskových fonémů. Typickým rysem je palatalizace (dělení souhlásek na měkké a tvrdé), která je velmi výrazná především u východoslovanských jazyků, kde většina souhlásek má i svou palatalizovanou verzi. V češtině, slovenštině a ukrajinštině se původní slovanské /g/ změnilo na /h/. Ve většině jazyků znělé párové souhlásky ztrácejí znělost na konci slov. V případě skupiny souhlásek určuje znělost celé skupiny poslední souhláska.

Všechny slovanské jazyky mají standardně 5 základních samohlásek (a, e, i, o, u). Ve východoslovanských jazycích a v polštině se rozlišuje přední a zadnější /i/ (tzv. tvrdé). Čeština a slovenština zachovávají v ortografii historické psaní i/y, ale rozdíl ve výslovnosti zanikl. Jihoslovanské jazyky i/y nerozlišují ani v písmu. Polština má nosové samohlásky (ę, ą). Slovinština rozlišuje mezi polootevřenými a polozavřenými /o, e/. Bulharština oproti ostatním jazykům rozlišuje i samohlásku /ə/, takzvané šva.

V ruštině a běloruštině je výslovnost samohlásek výrazně ovlivněna přízvukem.

Přízvučné samohlásky jsou vysloveny plně a zdůrazněny prodloužením, zatímco v nepřízvučných slabikách se jejich výslovnost redukuje.

Existuje více různých fonologických studií pro každý jazyk a všechny se většinou liší v detailech jednotlivých fonémů. Pro vývoj ASR systému však není nutné pevně dodržovat fonologické dělení fonémů a většinou je jejich počet efektivně redukován na základě testování různých hypotéz. Příloha B obsahuje tabulky s fonémy jednotlivých jazyků, dle kterých se tato práce řídila.

4.3 Morfologie, gramatika a dopad na vývoj ASR

Jak je popsáno v [70] a [71], slovanské jazyky se vyznačují vysokou ohebností a mají tak bohatě rozvinuté skloňování a časování slov, přičemž slovosled je relativně volný. Mají 6-7 pádů pro skloňování, kde dále ještě rozlišují číslo a rod. Výjimkou je bulharština a makedonština, kde bylo skloňování nahrazeno používáním předložek a zůstal zde pouze nominativ a vokativ. Na druhou stranu oba zmíněné jazyky používají u podstatných a přídavných jmen člen, který je připojován na konec slova, což navyšuje počet slovních tvarů ve slovníku.

Časování sloves vykazuje ve všech slovanských jazycích mnoho shodných flektivních rysů. Slovesa vyjadřují kategorie osoby, čísla, času a způsobu. Východoslovanské a západoslovanské jazyky mají standardně tři slovesné časy - přítomný, budoucí a minulý, zatímco jihoslovanské, s výjimkou slovinštiny, rozlišují přítomný, předbudoucí, budoucí a 4 časy minulé.

Všechny tyto rysy vedou k velmi vysoké slovní zásobě a relativně volnému slovosledu. Při vývoji rozpoznávacího systému je tak třeba počítat s mnohem větší slovní zásobou na rozdíl od analytických jazyků, jako je angličtina nebo španělština. Jedná se o stovky tisíc slov oproti desetitisícům u angličtiny. Volný slovosled dále značně snižuje efektivitu n-gramových jazykových modelů a především z technických důvodů je proto často nezbytné vystačit si s nižšími n-gramy. Vyšší n-gramy by při tak velké slovní zásobě vyžadovaly enormní množství trénovacích dat, mnohem rozsáhlejší paměťový prostor a vyšší výpočetní náročnost, zatímco přínos by nebyl adekvátní.

Tabulka 4.2: Ukázka textové podobnosti a odlišnosti u slovanských jazyků prezentovaná na části Všeobecné deklarace lidských práv

CZ	Všichni lidé se rodí svobodní a sobě rovní co do důstojnosti a práv. Jsou nadáni rozumem a svědomím a mají spolu jednat v duchu bratrství.
SK	Všetci ľudia sa rodia slobodní a rovní v dôstojnosti aj právach. Sú nadaní rozumom a svedomím a majú sa k sebe správať v duchu bratstva.
PL	Wszyscy ludzie rodzą się wolni i równi pod względem swej godności i swych praw. Są oni obdarzeni rozumem i sumieniem i powinni postępować wobec innych w duchu braterstwa.
SL	Vsi ljudje se rodijo svobodni in imajo enako dostojanstvo in enake pravice. Obdarjeni so z razumom in vestjo in bi morali ravnati v duhu bratstva.
HR	Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima. Ona su obdarena razumom i sviješću i trebaju jedno prema drugome postupati u duhu bratstva.
SR	Svi ljudi se rađaju slobodni i jednaki u dostojanstvu i pravima. Oni su obdareni razumom i svešću i treba da se odnose jedna prema drugima u duhu bratstva.
BS	Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima. Ona su obdarena razumom i sviješću i treba da jedno prema drugome postupaju u duhu bratstva.
CRN	Sva ljudska bića rađaju se slobodna i jednaka u dostojanstvu i pravima. Ona su obdarena razumom i savješću i jedni prema drugima treba da postupaju u duhu bratstva.
MK	Сите човечки суштества се раѓаат слободни и еднакви по достоинство и права. Тие се обдарени со разум и совест и треба да се однесуваат еден кон друг во духот на општо човечката припадност.
BG	Всички човешки същества се раждат свободни и равни по достойнство и права. Те са надарени с разум и съвест и следва да се отнасят помежду си в дух на братство.
RU	Все люди рождаются свободными и равными в своём достоинстве и правах. Они наделены разумом и совестью и должны поступать в отношении друг друга в духе братства.
UK	Всі люди народжуються вільними і рівними у своїй гідності та правах. Вони наділені розумом і совістю і повинні діяти у відношенні один до одного в дусі братерства.
BE	Усе людзі нараджаюцца вольнымі і роўнымі ў сваёй годнасьці і правох. Яны надзелены розумам і сумленьнем і павінны ставіцца адзін да аднаго ў духу братэрства.

5 Lingvisticko-lexikální část multilingválního systému rozpoznávání řeči

Tato a následující kapitoly popisují výzkum a vývoj metod, které byly použity při tvorbě systémů rozpoznávání řeči pro slovanské a následně i další jazyky. Jsou zde prezentovány dílčí kroky vývoje jednotlivých částí systému. Popis je pro přehlednost rozdělen do dvou kapitol. Tato se zabývá tvorbou textového korpusu, slovníku, jazykového modelu a částečně i postprocessingem. Součástí slovníku jsou ale také výslovnosti jednotlivých slov, což je řešeno v následující kapitole zabývající se akusticko-fonetickou částí ASR systémů.

Funkčnost jednotlivých kroků bude demonstrována zejména na východoslovanských jazycích, na kterých jsem nejintenzivněji pracoval, ale v určitých specifických případech i na ostatních slovanských jazycích.

5.1 Tvorba korpusu pro daný jazyk

Textový korpus sestává z velkého množství textů daného jazyka a následně je z něho vytvářen slovník a jazykový model. Kromě využití již hotových korpusů popsaných v kapitole 2.3 může být takový korpus vytvořen shromážděním a zpracováním volně dostupných dat z internetu.

Pro tvorbu ideálního textového korpusu je zapotřebí nasbírat dostatečné množství textů z různorodých zdrojů, aby tak pokrýval co největší část zpracovávaného jazyka. Řádově se jedná minimálně o stovky MB, nejlépe jednotky GB. Po získání dostatečného množství textových dat je potřeba texty zpracovat do použitelné podoby, tj. normalizovat, vyfiltrovat nežádoucí elementy a naformátovat pro další použití.

5.1.1 Zdroje textových dat

V první řadě je potřeba najít vhodné a dostatečně objemné zdroje dat. Nejvýznamnějším zdrojem textových dat jsou většinou webové stránky novinových zpravodajských portálů či rozhlasu a televize. Ty by v první řadě měly procházet jazykovou korekturou a obsahovat tak spisovnou podobu jazyka. Dále mají velký rozsah témat a pokryjí tak velkou část slovní zásoby, ale hlavně obsahují velké množství dat

v podobě článků přidávaných každý den. Textový korpus tak může být automaticky doplňován i každý den a reflektovat současné dění.

V případě práce s cizími jazyky, kdy se bez znalosti prostředí těžko hledají vhodné weby, jsou dobrým zdrojem například stránky ministerstva zahraničí[1] nebo speciální stránky shromažďující odkazy na zpravodajské weby jako ABYZ News Links[2]. Tabulka 5.1 zobrazuje hlavní zdroje použité pro tvorbu korpusů východoslovanských jazyků.

Tabulka 5.1: Hlavní textové zdroje východoslovanských jazyků

Země	Zdroje	Staženo dat
Rusko	1 TV, NTV, RT, RTR, TV Tsentr, Radio Mayak, Radio Radonezh, Radio Rossii, Gazeta, Integrum, Lenta, Mos News, Pravda, ...	2,38 GB
Ukrajina	1 TV, 112, 5 Kanal, Kanal Ukraina, Novyi Kanal, Radio Svoboda, STV, TSN, Ukrayinske Radio, Unian, MIG News, UKR, ...	4,39 GB
Bělorusko	Radio Svaboda, TVR, Belorusskie Novosti, Charter 97, Novy Chas, Belapan, Belta, Belarus Today, Narodnaia Volia, Zviazda, ...	2,56 GB

Další možnou (spíše doplňující) variantou jsou například knihy v elektronické podobě či úřední a parlamentní zápisy a dokumenty. Ty ale nemusí svým stylem příliš odpovídat normální mluvě, kterou by měl ASR systém rozpoznávat.

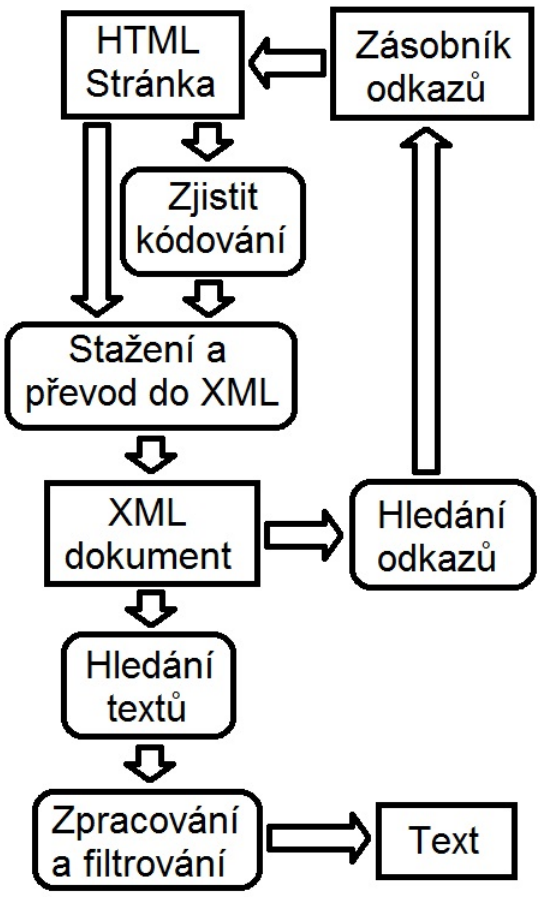
Nevhodná jsou diskuzní fóra, diskuze pod články a jakékoliv další zdroje, které tvoří samotní uživatelé a neprocházejí jazykovou korekturou. Většinou totiž obsahují nespisovné výrazy, pravopisné chyby a často postrádají diakritiku. Je proto potřeba tento obsah identifikovat a filtrovat.

5.1.2 Hromadné stahování dat

Nástrojů na stahování textů ze zadaných webových stránek je velké množství a proces je vždy podobný. Nicméně pro potřeby práce byl vytvořen vlastní nástroj, který umožňuje větší kontrolu nad všemi parametry stahování.

Proces těžení je znázorněn na obrázku 5.1. Zásobník odkazů shromažďuje odkazy a zároveň uchovává informaci, které odkazy již byly zpracovány. Jako první odkaz je sem manuálně vložen odkaz na hlavní stránku vybraného webu. Následně jsou postupně zpracovávány odkazy v zásobníku. Nejdříve je potřeba identifikovat kódování textu na daném webu. Jelikož se tato práce zabývá i jazyky, které používají různé abecedy, weby často obsahují různá kódování, která je potřeba rozpoznat.

[1] http://www.mzv.cz/jnp/cz/udalosti_a_media/media_ve_svete/
[2] http://www.abyznewslinks.com/



Obrázek 5.1: Schéma pro těžení textů z webových stránek

HTML stránka je následně stažena ve správném kódování a převedena do XML dokumentu. S XML dokumentem lze pak jednoduše pracovat za pomoci objektově orientované reprezentace DOM.

Jako první se v dokumentu vyhledají všechny odkazy, a pokud směřují na stejný web (základní doménová adresa se shoduje s první ručně zadanou adresou) a již nejsou v zásobníku odkazů, jsou do něho přidány.

V druhém kroku jsou na stránce vyhledány všechny textové elementy včetně vnořených elementů pro změnu řezu písma (např. kurzíva, tučné písmo atd.). Text z vnořených elementů je vyjmut a zasazen do nadřazeného textu, aby byla zachována správná posloupnost textu. Z každé stránky jsou vybrány všechny texty, které jsou následně podrobeny dalšímu zpracování.

5.1.3 Zpracování a filtrování textů

Stažené texty je potřeba nejprve zpracovat a normalizovat do jednotné podoby. Každý web může používat různé druhy kódování znaků s diakritikou, interpunkce či dalších speciálních znaků týkajících se daného jazyka. Proto je potřeba nejdříve správně dekódovat texty podle zjištěného kódování.

Webové stránky také často obsahují HTML entity, které slouží k reprezentaci znaku nezávisle na kódování. Entity mohou být zapsány názvem (např. é pro znak *é*) nebo číselně pomocí Unicode hodnot (např. đ pro znak *đ*). Tyto entity je nutné nahradit, k čemuž byla využita převodní tabulka[3] všech existujících entit na jejich příslušné Unicode znaky.

Občas se může stát, že webová stránka není plně validní (např. párová značka nebyla ukončena), tudíž plně neprojde převodem do XML dokumentu a některé HTML značky zůstanou v podobě textu. Proto je potřeba nasbírané texty analyzovat, pokusit se tyto elementy detekovat a odstranit je. Jelikož se ale mohly dostat do textu z důvodu nějaké chyby v zápisu, nemusí být stoprocentně vyhledatelné jako HTML značka. Nejefektivnější přístup je hledání podle klíčových slov názvů HTML tagů a jejich atributů a pak nejlépe vyřadit celý řetězec textu.

Dalším důležitým krokem je odstranění textů, které nesplňují určitá kritéria. Těmi jsou např. délka řetězce, počet slov, velký znak na začátku či interpunkční znaménko na konci. Tato kritéria z velké části zaručí, že jsou nakonec použity jen reálné věty a všechny názvy, popisky a další nežádoucí texty jsou odfiltrovány. Jako optimální délka textu byla ve většině případů vybrána hranice 10 znaků a alespoň tří až pěti slov (tokenů oddělených mezerou).

Na některých webech se stává, že se objevuje text bez diakritiky. Jelikož všechny slovanské jazyky psané latinkou obsahují kromě standardních 26 znaků latinské abecedy ještě vlastní znaky s diakritikou, mohou pak být jako podezřelé věty označené i ty, které nemají ani jeden znak s diakritikou. Případně je možné stanovit prahové množství, jelikož může být bez diakritiky jen část textu. To ale nemusí být úplně bezpečné, vždy záleží na konkrétním jazyce a frekvenci užívání znaků s diakritikou, která se dá jednoduše spočítat na již ověřených textech.

Protože se při procházení webu stahují všechny texty, je velice pravděpodobné, že se vícekrát stáhne nějaký delší popisek či prohlášení, které se opakuje na každé stránce. Tento text prošel všemi předchozími filtry, a dá se tedy považovat za plnohodnotnou větu. Nicméně není vhodné, když se v korpusu opakuje stejný text vícekrát, protože by poté zkresloval statistiku pro výběr slov do slovníku a následně i jazykový model. Proto se jako další krok porovnávají všechny řetězce mezi sebou a odstraní se veškerá redundance. Po zkoumání stažených textů bylo objeveno, že se také mohou vyskytovat řetězce, které se liší pouze číselnými hodnotami, např.

[3] https://dev.w3.org/html5/html-author/charref

datem. Proto je vhodné při porovnávání řetězců nebrat v potaz čísla, případně další speciální znaky.

Dále je vhodné včas identifikovat a odstranit webové a e-mailové adresy pomocí regulárních výrazů. Ty by totiž mohly být dále mylně rozděleny na slova a při dostatečné četnosti by se mohly dostat do slovníku. Tyto řetězce je potřeba nahradit za klíčová slova, která budou při dalším zpracování ignorována.

V neposlední řadě je potřeba provést normalizaci textu a odstranit nepotřebné znaky. Jelikož většina jazyků používá nějaké vlastní varianty znaků s diakritikou, tyto znaky mohou být často ve více variantách z různých Unicode rozsahů. Znamená to, že znaky vypadají stejně, ale mají jiný kód (např. znak latinky *C* s Unicode kódem 0x0043 a znak cyrilice *C* s kódem 0x0421). Bohužel v některých slovanských zemích (většinou země užívající cyrilici) není ustálené používání kódování a ve stažených textech se často objevuje užívání znaků z různých Unicode rozsahů. To může být způsobeno buď špatným užíváním kódování, nebo chybným převodem do Unicode z jiných kódování. To samozřejmě narušuje slovní statistiky a další zpracování, a je potřeba tyto záměny normalizovat. Dalším jevem, který se objevuje, je různé psaní znaků s diakritikou. V kódování Unicode existuje pro každý znak s diakritikou jeho vlastní kód, pak ale existuje i varianta zápisu složením základního znaku a znaku diakritiky, což se binárně projevuje jako dva znaky. To opět narušuje zpracování a je potřeba tento jev normalizovat.

V některých zemích se také objevují zdroje, které obsahují texty psané jinými abecedami, například arabštinou v Bosně, kde velkou část tvoří muslimské obyvatelstvo, ale také se často objevují cyrilicí psané texty v jazycích používajících latinku a naopak. Při zpracování konkrétního jazyka je vhodné tyto texty v jiné abecedě odfiltrovat. Toho se dá jednoduše docílit filtrováním celých Unicode rozsahů (např. rozsah 0x0400 až 0x04FF je vyhrazen pro cyrilici, rozsah 0x0600 až 0x06FF pro arabskou abecedu, atd.)

V Tabulce 5.2 je zobrazena statistika pro textový korpus východoslovanských jazyků. Kolik webových serverů bylo použito, kolik dat bylo staženo a kolik následně zbylo po zpracování výše popsaným způsobem.

Tabulka 5.2: Statistika těžení textových dat pro východoslovanské jazyky

Jazyk	**RU**	**UK**	**BE**
Počet zdrojů	12	28	11
Stažených dat	2,38 GB	4,39	2,56 GB
Dat po zpracování	998 MB	2,37 GB	814 MB

5.1.4 Jazyková filtrace

Důležitým momentem při tvorbě textového korpusu je identifikace a filtrace jiných jazyků, než je ten cílový. Ve stažených textech se velmi často objevují části v dalších jazycích. Ve většině případů je to způsobeno více jazykovými mutacemi zdrojového webu, kdy jsou všechny články překládány i do jiných jazyků a při těžení taktéž staženy. V dalších případech to pak jsou různé pasáže (povětšinou citace) v jazyce původního textu. Ve většině případů se jedná o texty v angličtině, vždy ale záleží na konkrétním jazyce a zemi.

Speciálním případem jsou jazyky, které jsou si navzájem srozumitelné (např. čeština a slovenština nebo chorvatština a srbština) a texty tak mohou být ponechány v původním jazyce. Nebo případy, kdy se v dané zemi užívá více jazyků. Například na Ukrajině a v Bělorusku je ruština standardně používána, a tak i zpravodajství a další zdroje jsou často dostupné v obou jazycích a v textech se mohou objevit oba jazyky zároveň.

Identifikaci jazyka textu je možné provést několika způsoby. První nejjednodušší přístup je identifikace na základě použité abecedy. Ten je vhodný při rozpoznávání velmi blízkých jazyků a vyžaduje pouze znalost užívaných abeced. Tento přístup byl využit u východoslovanských jazyků, které jsou si velmi blízké, ale zároveň díky abecedě snadno rozpoznatelné. Tabulka 5.3 ukazuje rozdíly v abecedách jednotlivých východoslovanských jazyků, které byly využity při identifikaci. Tabulka 5.4 ukazuje rozdíly v abecedách bulharštiny a makedonštiny, použité k identifikaci.

Tabulka 5.3: Rozdíly v abecedách východoslovanských jazyků

	RU	**UK**	**BE**
RU	-	э, ё, ы, ъ	и, ъ, щ
UK	ґ, є, і, ї, '	-	ґ, є, и, ї, щ
BE	і, ў, '	э, ё, ы, ў	-

Tabulka 5.4: Rozdíly v abecedách bulharštiny a makedonštiny

	BG	**MK**
BG	-	й, щ, ъ, ь, ю, я
MK	ѓ, ѕ, ј, ќ, љ, њ, џ	-

Pro filtrování vzdálenějších jazyků, jako je pro slovanské jazyky např. angličtina či francouzština, je možné využít metodu, která využívá k identifikaci seznam několika tisíc nejčetnějších slov v daném jazyce. Takové seznamy lze pro větší světové jazyky snadno najít na internetu. Identifikace následně funguje buď porovnáním mezi dvěma jazyky, přičemž vyhrává ten s větším počtem slov, nebo pouze spočtením identifikovaných slov v textu pro jeden jazyk a při přesažení stanovené hranice počtu slov je identifikován jako daný jazyk a odfiltrován.

Dalším možným přístupem je využití znakových n-gramových modelů, kde jsou na trénovacích datech spočítány statistiky užívání n-gramů pro každý jazyk, který má být identifikován. Tento přístup, jak byl popsán v [72], dosáhl velmi dobrých výsledků. Využití tohoto přístupu již vyžaduje dostatečné množství textů pro trénování, a tak nelze jednoduše využít v případě, že první data pro daný jazyk teprve získáváme. Variantou pak jsou již hotové modely nebo celé nástroje jako například *fastText*[4][73], který v současné době obsahuje modely pro 170 jazyků.

Pro východoslovanské jazyky, které jsou na webových stránkách často užívány zároveň, byly využity první dvě popsané metody. První pro identifikaci mezi východoslovanskými jazyky, druhá pro odfiltrování ostatních jazyků. Evaluace úspěšnosti byla provedena na 500 větách rovnoměrně rozdělených mezi všemi jazyky a označených rodilými mluvčími. Výsledná úspěšnost této metody byla 96.8 %. V Tabulce 5.5 je zobrazeno, kolik zbylo dat po aplikaci těchto metod na ukrajinské a běloruské texty, kde byla velmi hojně zastoupena ruština.

Tabulka 5.5: Statistika textových dat ukrajinštiny a běloruštiny po jazykovém filtrování

Jazyk	**UK**	**BE**
Zpracovaná data	2,37 GB	814 MB
Po jaz. filtrování	758 MB	280 MB

5.1.5 Vnitřní kódování jazyků

Jelikož některé jazyky používají jinou abecedu, je pro usnadnění práce vhodné vytvořit si interní abecedu a převodník. Je pak možné používat standardní klávesnici a není potřeba řešit instalaci jiných abeced a fontů do nástrojů používaných k vývoji. Zároveň při potřebě ručních úprav není potřeba hledat speciální znaky, ale je možno vše zapisovat snadno dostupnými znaky na klávesnici.

Existují oficiální způsoby transliterace například pro ruskou azbuku do latinky (tzv. romanizace). Nicméně tyto převody jsou většinou jednosměrné a přesný převod zpět je problematický. Většinou je problém v tom, že některé znaky jsou převedeny na dva či více znaků a při zpětném převodu není přesně jasné, jestli tyto dva znaky převést zpět na jeden či dva (např. znak cyrilice щ bývá převeden na *shch*, při zpětném převodu nemusí být jasné, zda znak převézt zpět na щ, nebo na шч). Proto je potřeba navrhnout a využít převod jedna ku jedné, aby následný výstup z rozpoznávače mohl být jednoznačně transformován zpět do původní abecedy.

Vlastní interní abeceda pro cyrilici byla navržena dle zásad, aby se převod co nejvíce přibližoval české výslovnosti, byl psatelný na české klávesnici, aby usnadňoval práci všem členům týmu a zároveň aby umožňoval přesný převod jedna ku jedné

[4] https://fasttext.cc/

do interní abecedy a zpět do cyrilice. V tabulce 5.6 je ukázka převodu pro jazyky používající cyrilici, kromě jihoslovanských jazyků standardně používajících cyrilici i latinku, kde byla zvolena práce s latinkou a užití oficiálních převodníků mezi těmito abecedami. V příloze C je zobrazena navržená abeceda a mapování všech znaků abeced slovanských jazyků používajících cyrilici.

Tabulka 5.6: Ukázka převodu různých národních variant cyrilice do interní abecedy

RU	Член Общественной палаты России Ирина Волынец предложила изменить правила трудоустройства россиян по совместительству
	Člěn Obşěstvěnnoj palaty Rossii Irina Volyněc prědložila izměnit‸ pravila trudoustrojstva rossiân po sovměstitěl‸stvu
UK	Колишній небожитель політичного Олімпу Давид Жванія, публікує одне скандальне відео за іншим
	Kolyšnij nebožytel‸ polityčnoho Olimpu Davyd Žvaniâ, publikuě odne skandal‸ne video za inšym
BE	Мы паспяхова процідзейнічаем незаконнай міграцыі, наркатрафіку, кантрабандзе і гэтак далей
	My paspâxova procidzějničaěm nězakonnaj mihracyi, narkatrafiku, kantrabandzě i hetak daléj
BG	По-късно днес правителството ще се събере и официално ще приеме постановление за удължаване на извънредното положение до средата
	Po-kăsno dnes pravitelstvoto şe se săbere i oficialno şe prieme postanovlenie za udălžavane na izvănrednoto položenie do sredata
MK	Некаде има многу голем број на ученици, па ќе мора да се воведе учење во смени, а во други пак, има помали паралелки и нема да има потреба
	Nekade ima mnogu golem broj na učenici, pa ťe mora da se vovede učeňe vo smeni, a vo drugi pak, ima pomali paralelki i nema da ima potreba

5.1.6 Textový preprocessing

Na závěr je ještě vhodné pro zvýšení efektivity rozpoznávání provést preprocessing textového korpusu. Základem je zpracování číslic a dat, ale i častých zkratek, jednotek a případně i interpunkčních znamének a dalších symbolů, pokud je systém tvořen například pro účely diktování.

U flektivních jazyků vyvstává problém detekce správného pádu číslovky. Efektivním řešením je nahrazení všech číslovek v textu zástupnými tokeny, které jsou pak přidány do slovníku s různou výslovností pro všechny pády. U některých jazyků se dají identifikovat i řadové číslovky podle tečky za číslicí, jiné jazyky tento zápis neužívají. Pro tento úkon je potřebná znalost ortografie a výslovnosti všech číslovek. A zároveň je dobré znát i různé varianty výslovnosti, jelikož číslovky většinou tvoří

dlouhé řetězce slov a často jsou vyslovovány rychle v nejsnadněji vyslovitelné formě, která nemusí odpovídat fonetickým pravidlům (např. české výslovnosti /*sedm*/ - /*sedum*/, /*čtrnáct*/ - /*štrnác*/).

U velmi častých zkratek či jednotek je také vhodné je nahradit zástupným tokenem a přidat všechny možné pády a výslovnosti. Nejčastěji se to týká zkratek státních úřadů, velkých firem, psaných zkratek typu *tj.* či *atd.*, peněz nebo fyzikálních jednotek.

Tabulka 5.7: Ukázka preprocessingu chorvatského textu pomocí zástupných tokenů

Původní text	Odredbom članka 29. § 1. 6. Zakona o posebnom porezu na duhanske proizvode
Po preprocessingu	odredbom članka 20# 9# točka stavak 1# točka 6# točka zakona o posebnom porezu na duhanske proizvode

5.2 Tvorba slovníku

Slovník je jednou ze stěžejních částí rozpoznávače. Obsahuje všechna slova, která mohou být rozpoznávačem rozpoznána. Ke každému slovu ve slovníku jsou přiřazeny přípustné výslovnosti (což bude probráno v následující kapitole zabývající se akusticko-fonetickou stránkou ASR).

Slovník je vytvořen výběrem slov (řetězců oddělených mezerami) dle jejich četnosti v textovém korpusu. To znamená, že jsou z korpusu vybrána všechna unikátní slova a napočítána jejich četnost, podle níž jsou seřazena. Počet slov vybraných do slovníku záleží na typu jazyka. Například u analytických jazyků, jako je angličtina nebo španělština, vystačí slovník o velikosti do sto tisíc slov pro pokrytí většiny slovní zásoby. U slovanských jazyků, které mají velkou míru skloňování a časování slov, se ukazuje jako nezbytné množství alespoň tři sta tisíc slov. Dostačující množství slov pro slovník je určeno pomocí míry množství slov mimo slovník OOV (tzv. Out-of-vocabulary). Ta se spočítá jako poměr počtu unikátních slov v korpusu, která nebyla vybrána do slovníku, k celkovému počtu slov. Přijatelná míra OOV je kolem 1-3 %.

Ne všechna vybraná slova jsou reálná slova z daného jazyka, především ta krátká vznikající ze zkratek či rozdělených slov. Je proto vhodné zkontrolovat všechna jedno, dvou a případně i třípísmenná slova, zdali opravdu v daném jazyce existují. Dále je vhodné detekovat a zkontrolovat další podezřelá slova, např. příliš dlouhá, obsahující znaky nepatřící do abecedy zpracovávaného jazyka, dlouhá slova bez samohlásek, atd.

V některých jazycích je časté užívání spojovníků. Může se jednat jak o gramatické vyjádření (např. v češtině vyjádření podmiňovacího způsobu přidáním -li),

tak o spojování častých slovních spojení (vyskytuje se hodně ve východoevropských jazycích, např. jiho-západ). Tento jev může v některých případech vést až ke zbytečné redundanci slov ve slovníku a tím i k možnému zhoršování jazykového modelu. Vhodným řešením pro odstranění těchto redundancí je výběr jen jedné varianty dle četnosti. Tedy pro všechna slova se spojovníkem se zjistí, zdali existují i jednotlivá slova po rozpojení či spojená varianta bez spojovníku, a poté se spočítají četnosti jednotlivých variant, podle kterých je zvolena varianta, která bude ponechána. Zároveň je potřeba podle toho upravit i korpus.

Dalším krokem, který může zlepšit rozpoznávání, je přidání kolokací, tedy velmi častých slovních spojení, kdy dvě či více slov jsou v korpusu spojena do jednoho tokenu a následně přidána do slovníku. Může se jednat o unikátní slovní spojení, kdy se samostatná slova téměř nevyskytují (např. Addis Abeba, Los Angeles, ad absudrum, de iure, ...). Další jsou častá spojení slov především s předložkami či spojkami, a to zejména jednofonémovými. Tím může být značně zlepšeno rozpoznávání, jelikož krátká slova jako předložky a spojky jsou často špatně rozpoznávána. Dále to mohou být častá jména osob, institucí a firem, zejména skládají-li se z krátkých slov. Přidáním kolokací je možné zlepšit rozpoznávací skóre i v řádu procent, jak bylo ukázáno v [74]. Zároveň se přidáním kolokací supluje vyšší n-gramový jazykový model. Kolokace jsou opět vybírány podle statistiky četnosti v korpusu, případně na základně tzv. vzájemné informace. Počet přidaných kolokací záleží na daném jazyce, ale rámcově se pohybuje v řádu tisíců až několika desetitisíců. Hodnotu lze nalézt testováním na validačních datech.

Tabulka 5.8 zobrazuje, jak velké slovníky byly vybrány z korpusů pro východoslovanské jazyky a procento zbývajících slov v korpusu, která nebyla přidána do slovníku.

Tabulka 5.8: Statistika vytvořených slovníků pro východoslovanské jazyky

Jazyk	Velikost korpusu	Počet slov	OOV
RU	998 MB	326 tis.	2,02 %
UK	758 MB	324 tis.	1,94 %
BE	280 MB	293 tis.	1,30 %

5.3 N-gramový jazykový model

Na závěr je pro slova ze slovníku a korpusu vypočítán n-gramový jazykový model. V této práci bylo využito bigramových modelů z důvodu použitého ASR systému. Nicméně vzhledem k tomu, že slovanské jazyky mají relativně volný slovosled a velkou slovní zásobu, byl by vyžadován mnohem větší korpus pro natrénování n-gramového modelu vyššího řádu. Zároveň by se razantně zvýšila výpočetní náročnost rozpoznávání a systémy by tak nemuselo být možné nasadit pro online rozpoznávání v reálném čase.

Pro vytvoření bigramového jazykového modelu jsou spočítány četnosti slov a dvojic slov v korpusu. Podmíněná pravděpodobnost pro každý bigram je poté spočítána podle vztahu 5.1, kde C je četnost výskytu slova či sekvence slov v korpusu.

$$P(w_i|w_{i-1}) = \frac{C(w_i, w_{i-1})}{C(w_i)} \tag{5.1}$$

Pro neviděné sekvence slov, které by měly četnost 0, je použito vyhlazování pomocí Witten-Bellova algoritmu [75].

Jak již bylo zmíněno v předchozí části, pro doplnění většího kontextu jsou přidány kolokace těch nejčastějších slovních spojení, což částečně supluje n-gramový model vyššího řádu. Z tabulky 5.9, která zobrazuje statistiky ruského korpusu a jazykového modelu, lze vidět, že doplněné kolokace jsou obsaženy přibližně v 10 % všech bigramů.

Tabulka 5.9: Statistika ruského korpusu a jazykového modelu

Velikost korpusu	998 MB
Celkový počet slov	149 352 988
Počet unikátních slov	1 594 109
Slov ve slovníku	326 324
Míra OOV	2,02 %
Celkový počet bigramů	144 233 687
Počet bigramů pro slova ve slovníku	127 231 285
Počet unikátních bigramů	30 864 417
Počet unikátních bigramů pro slova ve slovníku	28 589 865
Počet kolokací	1328
Počet bigramů obsahujících kolokaci	13 589 189
Počet unikátních bigramů s kolokací	3 566 667

5.4 Textový postprocessing

Posledním jazykově závislým textovým modulem je textový postprocessing, který dodatečně upravuje výstup dekodéru. Základem je převod na číselné a formátované údaje jako jsou data, zkratky atd. Dále převod z interního na originální kódování, pokud bylo použito. Nakonec to může být přidání interpunkce či dalšího formátování textu. Zde vždy záleží na cílové aplikaci a tato práce se tímto tématem podrobněji nezabývá.

Tabulka 5.10: Ukázka postprocessingu chorvatského textu

Výstup z rozpoznávače	pronađeni su ostaci devetog se jednog tisuću devetsto osamdeset šesta godine
Po postprocessingu	Pronađeni su ostaci 9. 7. 1986 godine.

6 Akusticko-fonetická část multilingválního systému rozpoznávání řeči

V této části jsou řešeny moduly a nástroje, které souvisí s převodem akustického signálu řeči na její textový přepis, tj. sestavení foneticko-akustického inventáře základních zvuků (fonémů a neřečových událostí), vytvoření výslovnostní části slovníku a s tím související vývoj grafémově-fonémového převodníku a zejména vývoj akustického modelu spojený s automatizovaným vytvářením trénovací databáze.

6.1 Foneticko-akustický inventář

Cílem je sestavit optimalizovaný inventář fonémů pro konkrétní jazyk (a doplnit ho o jazykově nezávislý inventář nejčastěji se vyskytujících neřečových zvuků). Pro každý jazyk existují fonetické studie rozlišující jednotlivé fonémy daného jazyka, případně i jeho různých dialektů. Tyto studie jsou ale většinou velmi důsledné v detailním rozlišování jednotlivých fonémů, případně jejich kontextově závislých variant. Pro úlohu rozpoznávání řeči není však důležité detailní dělení fonémů, a tak jsou tyto studie brány jako počáteční přehled, podle kterého se následně sestavuje fonetický inventář.

Při sestavování inventáře se řídíme následujícími kritérii:

- Jako základ pro tvorbu fonetického inventáře slouží fonetické studie a tabulky jak je popsáno v kapitole 4.2.
- Snaha o sdílení zvukově blízkých fonémů mezi jazyky, čehož může být následně využito při tvorbě multilinguálních modelů, které mohou být následně využity v počátečních fázích vývoje akustického modelu pro nový jazyk.
- Snaha o usnadnění automatického převodu mezi ortografickou a fonetickou podobou slova tak, aby bylo možné co nejjednodušeji pracovat s výslovnostmi, zejména při manuálních úpravách, a to i při minimální znalosti daného jazyka.
- Snaha o minimalizaci počtu fonémů, zejména sdružováním foneticky blízkých fonémů a alofonů, a to s ohledem na schopnosti akustického modelu „naučit“ se a v rámci pravděpodobnostních parametrů pokrýt různé fonémové alternativy. Může být navrženo několik hypotéz, které jsou průběžně testovány a nakonec vybrána ta nejvhodnější.

6.1.1 Vlastní fonetická abeceda

Fonetické studie a tabulky obecně používají pro zápis fonémů mezinárodní fonetickou abecedu (IPA[1]). Ta nicméně není příliš vhodná pro práci na standardní klávesnici ani pro strojové zpracování. Často využívá velice speciální znaky a pro zápis některých fonémů, jejich obměn výslovnosti či prozodie využívá více znaků a s různou diakritikou. Pro zjednodušení práce je vhodnější varianta, kdy každý foném má svůj vlastní znak, a navíc tyto znaky jsou zvoleny tak, aby je bylo možné psát na standardní klávesnici (v případě této práce na české).

Pro počítačové zpracování byla navržena například abeceda SAMPA, respektive její rozšířená varianta X-SAMPA[2], která mapuje abecedu IPA pomocí základních znaků anglické abecedy, které navíc rozlišuje pomocí kapitalizace. Ale opět pro některé fonémy či prozodii využívá doplňování písmen o další různé znaky (např. lomítko, dvojtečka, apostrof), čímž se značně znepřehledňuje zápis a komplikuje čtení.

Pro co největší zefektivnění práce byl využit přístup navržený v [76] pro českou fonetickou abecedu. Ten se řídí následujícími zásadami:

- Fonémy jsou označovány pouze jedním znakem pro snadnou čitelnost a zamezení nejednoznačností.
- Rozlišuje se mezi velkými a malými znaky.
- K označení fonému je využit znak, který se s daným fonémem nejčastěji pojí (při práci v českém týmu je tedy zohledněna česká výslovnost).
- Znak by měl být zapsatelný na české klávesnici, pro usnadnění rychlého zápisu a oprav.

Pro některé interní operace, kde nelze použít symboly s diakritikou (např. při trénování AM v prostředí HTK), je využita ještě další interní abeceda, která může být víceznaková a odpovídá konkrétnímu nástroji.

Obecně je v první fázi vývoje výhodnější vybrat fonetickou abecedu spíše větší, aby pokryla co nejvíce fonémů. Po prvních experimentech lze analyzovat výstupy rozpoznávače řeči a na jejich základě rozhodnout o redukci abecedy. Analýza probíhá na základně kontroly výstupu testovacích dat, kdy lze porovnat správnou variantu s výstupem rozpoznávače a odhalit tak, kde a z jakého důvodu vznikají chyby.

Posledním důležitým kritériem pro výběr fonetické abecedy je výběr tzv. zdrojového jazyka pro situaci, kdy počáteční vývoj systému pro cílový jazyk bude založen na využití akustických dat z jiného, zdrojového jazyka (nebo více jazyků). V tu chvíli je potřeba naplánovat způsob mapování fonetických abeced mezi těmito jazyky.

[1] https://www.internationalphoneticalphabet.org/
[2] https://commons.wikimedia.org/wiki/X-SAMPA

Tabulka 6.1 zobrazuje ukázku fonetické transkripce ruského textu s mezikrokem převodu do interní abecedy a pro srovnání převod do X-SAMPA, který je viditelně mnohem hůře čitelný než vlastní navržená abeceda.

Tabulka 6.1: Ukázka fonetické transkripce ruštiny

Původní text	В связи с реорганизацией совхозов в тысяча девятьсот девяносто третьем году возглавил крестьянско-фермерское хозяйство в Новосергиевском районе
Převod do interní abecedy	V svâzi s rěorganizaciěj sovhozov v tysâča děvât‿sot děvânosto trět‿ěm godu vozglavil krěst‿ânsko-fěrměrskoě hozâjstvo v Novosěrgiěvskom rajoně
Vlastní fonetická transkripce	f sVázi s Reorgaňizácijej sofxózof f týSača ďeVáťsod ďeVánosto tRéťjem gódu vozglávil kResťjánskoFérMerskoje xoZájstvo v novoSérgijefskom rajóňe
Transkripce do X-SAMPA	f sV'"azi s r'eorgan'iz"atsijej sofx"ozof f t"1s'atS'a d'ev'"acót'sot d'ev'"anosto tr'"et'jem g"odu vozgl"áv'il kr'est'j"anskof'"erm'erskoje xoz'"ajstvo v novos'"erg'ijefskom raj"on'e

6.1.2 Neřečové zvuky a jejich symboly

Fonetickou sadu je potřeba doplnit i o neřečové zvuky. Ty mohou být sdíleny mezi jazyky, což je užitečné především v prvotní fázi vývoje akustického modelu, kdy není k dispozici dostatek dat. Tabulka 6.2 zobrazuje používané neřečové zvuky.

Tabulka 6.2: Seznam používaných neřečových zvuků s jejich kódováním

Ticho	-
Fonetický ráz	0
Klik	1
Slabý kratší hluk	2
Nádech	3
Silnější delší hluk	4
Hesitace, apod.	5

6.2 Vytváření výslovnostní části slovníku

Pro vytvoření výslovnosti, tedy převod z ortografické podoby slova na fonetickou, využíváme tzv. G2P převodníku (Grapheme-to-Phoneme). Ten může být různých typů od systému využívající předem vytvořená produkční pravidla až po systémy využívající metod strojového učení. Produkčních pravidel je využíváno zejména v počátečních fázích vývoje, kdy nejsou k dispozici žádná data, na kterých by bylo možné trénovat.

Slovanské jazyky jsou tzv. ortograficky mělké, tedy rozdíl mezi výslovností a psanou podobou je relativně malý a pro fonetickou transkripci tedy stačí menší množství pravidel. Mohou tak být využita produkční pravidla využívající okolní kontext fonému, kde fonetická transkripce probíhá tak, že je procházen ortografický tvar slova znak po znaku, je kontrolováno jeho okolí a podle toho je použito vhodné produkční pravidlo. Pravidla jsou definována ve tvaru:

$$A \rightarrow B / C\_D \tag{6.1}$$

To znamená, že pokud řetězci A předchází řetězec C a je následován řetězcem D, je přepsán na řetězec B. Pro usnadnění zápisu mohou řetězce C a D obsahovat i zástupné skupiny (např. pro samohlásky, znělé souhlásky, atd.), pro které jsou vygenerovány všechny možné variace při porovnávání pravidla. Použitý systém byl vyvinut na základě systému popsaného v [10], kde lze nalézt podrobný popis generování fonetické transkripce pro češtinu.

Příklad několika pravidel pro běloruštinu vypadá následovně:

```
šsâ => Sa / _
t^sâ => ťSa / _
t^sě => ťSe / _
t^sô => ťSo / _
t^si => ťSi / _
t^sű => ťSu / _
t^ => ď / _<W,WW,QW>Q
t^ => ť / _<Q,QQ,WQ>W
t^ => ď / _Q
t^ => ť / _
```

V pravidlech je využito zástupných symbolů jako W pro neznělé souhlásky a Q pro znělé souhlásky. Tyto skupiny jsou vždy vyjmenovány dopředu.

Stejně tak jsou vyjmenovány i znělé-neznělé páry souhlásek, které pak jsou využity v obecnějších pravidlech pro řešení spodoby znělosti. Tato pravidla pak mohou být aplikována pro více jazyků, pokud se jejich výslovnost nějak specificky neliší. Znak A pak říká, že má být změněna znělost, znak O naopak říká, že má být znělost

ponechána. Příklad několika použitých obecných pravidel vypadá následovně:

```
W => O / _<v,Wv -v,-Wv,W-v>
W => O / _<Q,W>-<Q,W>W
W => A / _<Q,W>-<Q,W>Q
W => O / _<Q,W>-W
W => A / _<Q,W>-Q
W => O / _-<Q,W>W
Q => O / _<Q,W>Q-<Q,W>Q
Q => O / _<Q,W>-<Q,W>Q
Q => O / _<Q,W>Q-Q
Q => O / _<Q,W>-<Q,W>Q
Q => A / _<Q,W>Q-
```

V každém jazyce existují slova, která mohou mít více různých přípustných výslovností. Nejčastěji se to týká spodoby znělosti na konci slov. Proto při generování výslovnosti je potřeba kontrolovat každé slovo, zda končí znělou či neznělou souhláskou (případně celou skupinou) a vygenerovat výslovnost s opačnou znělostí této poslední souhlásky (respektive skupiny souhlásek). Zbylé případy jsou zavedené výslovnosti odlišné od základních pravidel, většinou z historických důvodů. Pro takové případy, pokud se dají popsat novou sadou pravidel, se vygeneruje další výslovnost používající tato upravená pravidla. Pokud se jedná o nesystémové výjimky, je vhodné najít způsob, jak tyto výjimky detekovat ve slovníku a následně pro ně vygenerovat výslovnost. Na závěr může být přistoupeno k ruční tvorbě výslovností pro konkrétní specifická slova.

Pravidla pro fonetickou transkripci mohou být nalezena ve fonetických studiích pro daný jazyk, případně dovozena z dalších zdrojů. Těmi mohou být například internetové lekce jazyka, případně je potřeba výslovnost "naposlouchat" na nahrávkách s textem a pravidla vydedukovat. Tato část je místem, kde je potřeba co nejvíce porozumět zpracovávanému jazyku a oproti ostatním částem vývoje vyžaduje alespoň základní znalosti fonetiky.

Důležitým krokem, který vyžaduje větší pozornost, je výslovnost zkratek a číslovek. Každý jazyk má svůj způsob výslovnosti zkratek. Zkratky mohou být hláskovány různými způsoby, anebo čteny jako normální slova. Každý jazyk má jednu nebo více hláskovacích abeced. Pro vytvoření výslovností je tedy potřeba detekovat zkratky a vygenerovat pro ně přípustné výslovnosti. Tabulka 6.3 ukazuje příklad hláskování části ruské abecedy.

Tabulka 6.3: Ukázka hláskování části ruské abecedy

а	б	в	г	д	е	ё	ж	з	и
a	be	ve	ge	de	e	jo	že	ze	i

Číslovky se ve slovanských jazycích často vyslovují velice zrychleně a zkráceně. Je proto vhodné zjistit všechny možné výslovnosti a přidat je do slovníku.

Jelikož jsou si některé jazyky podobné z hlediska výslovnosti, je možné při práci na novém jazyce využít těchto podobností a použít již vytvořená pravidla doplněná jen o odlišnosti nového jazyka. V největší míře se to týká pravidel pro spodobu znělosti či pro palatalizaci.

V pozdějších fázích vývoje, kdy už je k dispozici dostatek dat pro trénování, je pro tvorbu výslovností nasazen G2P systém využívající neuronové sítě. Tím se zabývají další členové týmu v pozdějších fázích, kdy je ASR systém již nasazován do praxe, a není to tak součástí této práce.

6.3 Vytváření databáze trénovacích nahrávek

Z trénovacích nahrávek je trénován akustický model. Tvorba akustického modelu je tou nejobtížnější částí při vývoji celého systému rozpoznávání řeči. Pro vytvoření akustického modelu použitelného pro rozpoznávání spojité řeči je zapotřebí alespoň několik hodin nahrávek řeči společně s jejich přesnými fonetickými přepisy. Nahrávky musí pocházet od více mluvčích a měly by být dostatečně fonémově rozmanité. Fonetické přepisy by zároveň měly obsahovat i anotaci neřečových zvuků. Hlavním cílem této kapitoly je popsat způsob automatického vytěžování trénovacích dat z volně dostupných zdrojů, které jsou k dispozici na internetu.

6.3.1 Dedikované trénovací databáze

Jak už bylo uvedeno v kapitole 2.7, pro mnoho jazyků existují již hotové řečové databáze. Nicméně jejich dostupnost a kvalita většinou koresponduje s počtem mluvčích daného jazyk. Nemusí tak být k dispozici dostatek dat a v dostatečné kvalitě pro tvorbu obstojně fungujícího systému.

Kvalitní databází, již popsanou dříve, je databáze Globalphone. Ta obsahuje dostatečné množství nahrávek pro spoustu jazyků a byla tak využita i v rámci zmíněných souvisejících projektů při tvorbě systémů pro ruštinu a srbochorvatské jazyky. Nicméně obsahuje jen nahrávky a textové anotace, je tedy nutné si vytvořit vlastní fonetické přepisy pro trénování akustického modelu. Navíc se při tvorbě fonetických přepisů podařilo najít menší nepřesnosti v anotacích.

Existují další volně dostupné databáze, které byly v rámci vývoje využity, jako například zmiňovaný VoxForge či polský Clarin [77]. VoxForge obsahuje velice přesné textové přepisy, ale má bohužel velmi špatnou kvalitu nahrávek a nemá fonetické přepisy. Clarin má naopak dobrou kvalitu nahrávek a obsahuje fonetické přepisy, nicméně v přepisech bylo nalezeno mnoho, i velmi závažných, nepřesností.

Při tvorbě ASR systémů pro slovanské jazyky tak bylo využito těchto dedikovaných databází pouze jako doplňku, případně pro validaci či testování, a hlavní data pro trénování byla vytvořena pomocí postupů popsaných dále.

6.3.2 Vlastní systém vytváření trénovacích dat

Základní myšlenkou navrhovaného přístupu tvorby vlastních trénovacích dat je nalézt na internetu co největší množství audio nebo video záznamů, ke kterým je připojen nějaký text, jenž by měl obsahovat promluvy v nahrávkách. Nahrávky jsou přepsány existujícím systémem a tyto přepisy jsou následně porovnány s připojeným textem.

Během porovnávání jsou hledány úseky, kde se shoduje automatický přepis s připojeným textem. Tyto segmenty jsou vyříznuty a dále použity. Pokud se výstup rozpoznávače plně shoduje s textem, je segment přidán do trénovací množiny. Pokud se shoduje jen částečně (např. více než 80 %), může být zkontrolován a opraven manuálně za využití speciálního nástroje, nebo být znovu rozpoznán s novým, lepším modelem, kdy má šanci být správně rozpoznán.

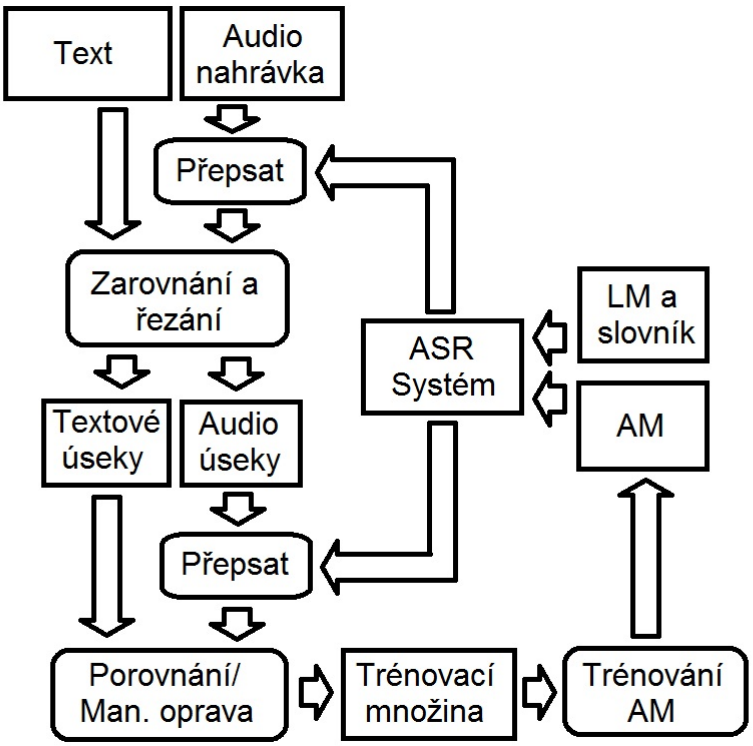
Po získání určitého množství nových trénovacích dat může být natrénován nový akustický model a celý proces se opakuje se zbylými nahrávkami, nebo pouze s těmi krátkými úseky, které neměly 100% shodu a nebyly tak přidány do trénovací množiny. Celý proces zpracování dat a tvorby akustického modelu je zachycen na obrázku 6.1. Jednotlivé kroky jsou detailněji popsány dále.

Hledání zdrojů

Jako první krok je potřeba nalézt vhodné zdroje. Nejcennějším zdrojem dat jsou webové stránky zpravodajských či televizních a rádiových stanic, které mají volně dostupné zpravodajské i jiné pořady společně s titulky či doslovnými přepisy (v tom nejlepším případě), anebo alespoň doplněné nějakým textem, který popisuje či cituje, o čem se mluví v pořadu. Druhým velmi dobrým zdrojem jsou parlamentní archivy se záznamy ze zasedání. Pokud jsou audio/video záznamy přístupné i společně se zápisy z jednání, je to velmi dobrý zdroj s velkým množstvím mluvčích. Nahrávky jsou také pořízeny v rušnějším prostředí, což může zvýšit robustnost akustického modelu. Dále je možno nalézt a použít volně dostupné audioknihy, internetové lekce daného jazyka a různé další zdroje.

Hledání vhodných zpravodajských zdrojů v cizím jazyce je obtížné bez znalosti tohoto jazyka a bez znalosti prostředí. Pokud nejsou k dispozici rodilí mluvčí, kteří by mohli odkázat na konkrétní zdroje, jedinou možností je hledání zdrojů vlastnoručně, nejlépe za využít některých serverů, sdružujících informace o zpravodajských serverech, jako již zmíněný server ABYZ News Links[3] nebo stránky Ministerstva

[3] http://www.abyznewslinks.com/



Obrázek 6.1: Schéma iterativního těžení akustických dat

zahraničních věcí[4], kde jsou k dispozici odkazy na hlavní zpravodajské servery v zahraničí.

Většina zpravodajských serverů ale neobsahuje videa, pouze články. Je proto potřeba prohledat všechny odkazy a zjišťovat, zda obsahují audio/video záznamy s nějakým textem, který by odpovídal tomu, co je řečeno v nahrávkách. Zjistit automaticky, jestli text odpovídá promluvě, je pro zcela cizí jazyk velmi obtížné, takže ideálním řešením je stáhnout a zpracovat vše, a nechat na systému, aby si s tím poradil.

Získávání a zpracování dat

Po nalezení vhodných zdrojů přichází na řadu stažení všech dostupných audio/video souborů a k nim přidružených textů (dále nazývaných referenční). Na to bohužel neexistuje univerzální řešení, protože každý web má odlišnou strukturu a využívá různé přehrávače a způsoby uchování audia/videa. Je proto v tomto případě nutné upravit stahování pro každý web tak, aby nejdříve procházel stránky, kde se nacházejí audia/videa a následně je stáhl. Ne vždy je tento automatický přístup možný, jelikož některé servery využívají ochranné mechanismy proti stahování.

[4]https://www.mzv.cz/jnp/cz/udalosti_a_media/media_ve_svete/index.html

Stažená data často bývají v různých formátech (nejčastěji však mp3/mp4). Je proto nutné data převézt do formátu stravitelného pro rozpoznávač. Stažený text je také potřeba upravit do tvaru použitelného pro porovnání s výstupem z rozpoznávače. To znamená odstranit všechnu interpunkci a další netextové znaky, kromě spojovníků, apostrofů a dalších znaků, které mohou být součástí slov. V případě jiných abeced je potřeba text převést do interní abecedy. Dále je vhodné rozepsat zkratky a číslice, alespoň do základního tvaru. To sice nemusí odpovídat správnému skloňování, ale i tak je zvýšena pravděpodobnost správného rozpoznání, protože v opačném případě, pokud budou zkratky a číslice ponechány v původním tvaru, budou ve všech případech vyhodnoceny jako neshoda.

Když jsou data připravena, přichází na řadu samotný proces těžení dat znázorněný na obrázku 6.1. Pro zjednodušení nejdříve uvažujme, že již je k dispozici fungující systém, tedy existuje výslovnostní slovník, jazykový model a akustický model pro zpracovávaný jazyk. (Následující podkapitola se zabývá případem, kdy se začíná s novým jazykem od úplného začátku a nejsou tudíž k dispozici žádná akustická data). Za využití tohoto systému jsou postupně rozpoznávány jednotlivé nahrávky jedna po druhé a výstup rozpoznávače je vždy porovnán s referenčním textem. K porovnání je využit následující postup z [78], kde je využito metody hledání nejmenší vzdálenosti (Minimum edit distance) pro zarovnání obou textů.

K zarovnání dochází na lokální úrovni nejvíce se shodujících částí obou textů. Metoda hledá optimální zarovnání slov referenčního textu r_j o počtu slov J a slov ve výstupu rozpoznávače w_i o počtu slov I. Počty slov I a J se mohou výrazně lišit, podle toho, jak moc odpovídá referenční text promluvě v nahrávce. Tato úloha je řešena pomocí dynamického programování za využití přiřazovací matice A, ve které je hledáno optimální řešení. Proces začíná inicializací prvního řádku a sloupce matice A:

$$A(i,0) = P_D \cdot (i-1), 1 \leq i \leq I; A(0,j) = P_I \cdot (i-1), 1 \leq i \leq J \tag{6.2}$$

Následně jsou rekurzivně dopočítány zbývající hodnoty matice dle vztahu:

$$A(i,j) = min[A(i-1,j-1)+d(r_i,w_i)-b_{i-1,j-1}; A(i,j-1)+P_I; A(i-1,j)+P_D] \tag{6.3}$$

kde

$$d(r_i,w_j) = \begin{cases} 0, & \text{pokud } r_i = w_j \\ P_S, & \text{pokud } r_i \neq w_j \end{cases} \tag{6.4}$$

a

$$b_{i-1,j-1} = \begin{cases} 0, & \text{pokud } r_i \neq w_j \\ P_S, & \text{pokud } r_i = w_j \end{cases} \tag{6.5}$$

Hodnoty P_D, P_I, a P_S jsou hodnoty pro penalizaci v případě delecí, inzercí a substitucí slov. Obvyklou hodnotou penalizace je 1. Hodnota $b_{i,j}$ napomáhá k vyhledávání nepřerušených sekvencí přidáním bonusu v (6.3). Po dopočítání matice A

je nalezeno nejlepší zarovnání zpětným průchodem matice z posledního bodu (I,J) do počátku (1,1) po nejmenších hodnotách. Každé slovo je při zpětném průchodu označeno jako shoda, delece, inzerce nebo substituce.

Když jsou texty zarovnány, je dále použit algoritmus hledající jakékoliv souvislé segmenty, které se do určité míry shodují a jsou ohraničeny tichem nebo některým z hluků. Shoda nemusí být 100%, rozpoznaný text nemusí být zcela správně, ale referenční text ano. Při vybrání i těchto segmentů je šance, že budou správně rozpoznány v následujících iteracích s lepším akustickým modelem. Tyto vybrané segmenty jsou následně vyříznuty z původní audio nahrávky i z textu. K tomu je využit výstup rozpoznávače, který zobrazuje časové značky začátku a konce jednotlivých slov a neřečových elementů. Algoritmus přijímá několik vstupních parametrů, a to:

- minimální a maximální počet slov, které mohou být ve vyřezávaném úseku,
- seznam hluků, které mohou tvořit hranice,
- procento shody mezi těmito úseky.

Když jsou z původní nahrávky vyřezány shodující se segmenty ve formě audio nahrávek a textu, nahrávky jsou znovu zpracovány rozpoznávačem a opět porovnány s jejich příslušnými texty. K vyhodnocení je použita míra WER.

Toto porovnání je provedeno pro všechny vyřezané segmenty. Ty, které mají hodnotu WER nulovou, jsou přidány do trénovací množiny pro trénování nového akustického modelu. Tento přístup zaručuje dostatečnou úroveň kontroly a je víceméně zaručeno, že fonetické transkripce vytvořené tímto procesem skutečně odpovídají tomu, co bylo řečeno v nahrávce.

Tabulka 6.4 ukazuje část výstupu rozpoznávače, kde je zobrazena ortografická i fonetická podoba rozpoznaných slov společně s čísly jejich počátečních a koncových framů vstupního signálu, podle kterých lze dopočítat přesná místa pro střih.

Tabulka 6.4: Ukázka výstupu rozpoznávače řeči aplikovaného na ukrajinštinu

Ort:	porâdok	kraîny	na	vykonannâ	social^nyx	iniciatyv	prezydenta
Phon:	poRadok	krajiny	na	vykonaňa	sociaLnyX	iňiciatyf	prezydenta
Start:	101	137	148	196	258	318	376
Stop:	137	148	196	258	318	376	426

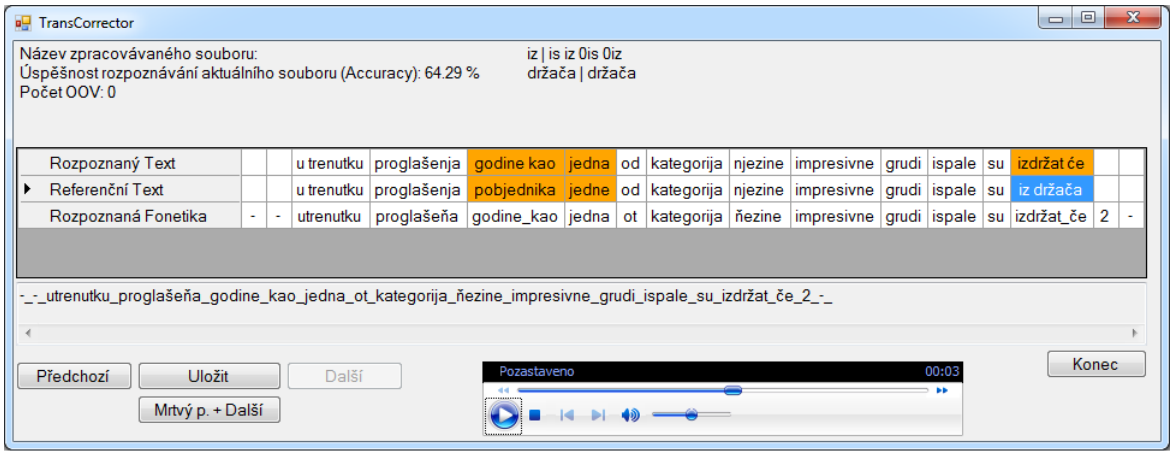
Manuální kontrola a editace

Pomocí speciálního nástroje (na obrázku 6.2) je možno zobrazit jednotlivé segmenty a jejich referenční a rozpoznaný text v orografické a fonetické podobě.

Oranžově jsou zobrazeny části, kde se referenční text s výstupem rozpoznávače neshoduje. K tomu si lze poslechnout nahrávku a podle ní odhadnout správnou variantu a manuálně opravit neshodující se části. Chyba může být jak ve výstupu rozpoznávače, tak ale i v referenčním textu, který neodpovídá přesně promluvě v nahrávce. Lze zvolit, jaké úseky se budou zobrazovat, podle procent shody referenčního textu a výstupu rozpoznávače. Je tedy tak možné kontrolovat a opravovat jen ty úseky, které mají chybu jen v jednom či dvou slovech.

Tento krok manuální editace není nezbytný, ale slouží zaprvé k urychlení práce v počáteční fázi vývoje, kdy akustický model není ještě příliš kvalitní a výtěžnost dat je nízká. Zadruhé lze pomocí tohoto kroku snadno odhalit chyby ve výslovnosti nebo ortograficky chybná slova, která se dostala do slovníku.

Práce s tímto nástrojem je velice efektivní neboť anotátor se může zaměřit především na barevně vyznačené úseky. Rozhodnutí, zde je správné rozpoznané nebo referenční slovo lze učinit jediným kliknutím myši. Pro každou nahrávku tak často stačí pouze dva až tři jednoduché úkony.



Obrázek 6.2: TransCorrector - nástroj na kontrolu a opravu fonetických přepisů

Trénování nového modelu

Posledním krokem celého algoritmu je natrénování nového akustického modelu z vytěžených segmentů. Nový model může být použit zaprvé pro zpracování zbývajících segmentů. Tím, že byl model trénován na segmentech z jedné nahrávky, se adaptoval na akustické podmínky v této nahrávce a šance úspěšného rozpoznání a vytěžení dalších segmentů se tím zvyšuje. Zadruhé je nový model použit pro zpracování dosud nepoužitých nahrávek. Formální zápis procesu pro zpracování dat a trénování akustického modelu je zapsán následujícím pseudokódem:

`Iterativní přetrénování:`

1. `Pro každý dokument`
    - `Proveď segmentaci shodných úseků`
2. `Pro každý segment`
    - `Přepiš segment`
    - `Porovnej a přidej do trénovací sady/(Oprav manuálně)`
3. `Přetrénuj akustický model`
4. `Opakuj krok 1. nebo 2.`

6.3.3 Využití multilingválního akustického modelu pro tvorbu trénovacích dat

V případě, že se začíná s vývojem systému pro nový jazyk a nejsou zatím k dispozici žádná akustická data pro tento jazyk, je zde možnost využít v těžícím schématu akustický model pro jiný jazyk (tzv. jazyk zdrojový), případně i více jazyků. V anglicky psané literatuře se tento přístup většinou označuje jako bootstrapping.

Z dostupných jazyků se snažíme vybrat jazyk foneticky nejbližší. Čím jsou si vybrané jazyky foneticky bližší, tím efektivnější je celý proces. Aby systém mohl využívat akustický model pro jiný jazyk, je nutné namapovat fonetickou abecedu cílového jazyka na abecedu jazyka zdrojového a změnit podle toho výslovnosti ve slovníku. Jazykový model zůstává původní pro cílový jazyk.

Akustický model pro čistě zdrojový jazyk je použit pouze v první iteraci těžícího algoritmu. Ve chvíli, kdy je získáno dostatečné množství trénovacích dat (řádově to mohou být desítky minut až jednotky hodin), je natrénován nový akustický model smícháním trénovacích dat pro zdrojový jazyk a nově získaných dat pro cílový jazyk. Tímto procesem se akustický model postupně adaptuje na cílový jazyk a výtěžnost nových dat se zvyšuje.

Ve chvíli, kdy je získáno dostatečné množství trénovacích dat pro cílový jazyk (většinou alespoň 5 hodin), může být zdrojový jazyk odstraněn z trénování, všechna získaná data převedena zpět do fonetické abecedy cílového jazyka a natrénován akustický model pouze pro cílový jazyk, který je dále využíván pro získávání dalších dat. Tento proces využití multilingválního systému je zapsán následovně:

`Multilingvální trénování:`

1. `Namapuj fonémy na zdrojový jazyk`
2. `Přidej zdrojový jazyk do trénovací množiny`
3. `Iterativní přetrénování`
4. `Odstraň zdrojový jazyk z trénovací množiny`
5. `Přemapuj fonémy zpět`
6. `Přetrénuj`

Výběr vhodného zdrojového jazyka

Při volbě zdrojového jazyka jsme v první řadě omezeni na výběr z jazyků, které jsou k dispozici. Ne vždy to musejí být ty nejbližší jazyky, a proto je potřeba vybrat ten nejvhodnější, či nějakou kombinaci jazyků. Dalším kritériem výběru jsou také možnosti mapovaní fonetických abeced.

Ne vždy může být toto mapování jednoduché, jelikož jednotlivé jazyky většinou mají unikátní fonémy, které v ostatních jazycích nejsou (jako např. české *ř* nebo ukrajinské měkké *c*). Často se tedy mapuje foném jedné abecedy na nějaký nejbližší foném druhé abecedy či na více fonémů.

Pro výběr nejvhodnějšího jazyka (jazyků) je nutné udělat úvodní test. K tomu jsou potřeba alespoň nějaká testovací sada v cílovém jazyce a připravený slovník a jazykový model. Následně se slovník namapuje na fonetické abecedy testovaných jazyků či jejich kombinací a provede se testování. Následuje ukázka takového postupu pro výběr vhodného zdrojového jazyka pro ukrajinštinu a bulharštinu.

Příklad výběru zdrojového jazyka pro vývoj ukrajinštiny

Pro výběr zdrojového jazyka pro ukrajinštinu bylo uvažováno mezi českým, polským a ruským akustickým modelem, které byly v tu chvíli k dispozici.

Přestože se může ukrajinština jevit jako velmi blízká ruštině, v některých ohledech se liší (např. nemá redukci nepřízvučných samohlásek nebo tak rozsáhlou palatalizaci souhlásek jako ruština) a podle některých zdrojů má foneticky blíž k polštině (zejména na západě Ukrajiny). Tabulka 6.5 ukazuje příklad, jak byla ukrajinská fonetická sada mapována na ruskou.

Tabulka 6.5: Mapování ukrajinské fonetické sady na ruskou

UK	a	e	i	o	u	X	ć	Ć
RU	á	é	í	ó	ú	x	cj	Cj

Pro testování byly zvoleny tři různé sady. První sada byla vytvořena nahráváním několika rodilých mluvčích, jako druhá byla využita databáze VoxForge pro ukrajinštinu po manuálním odstranění velmi nekvalitních nahrávek a jako třetí sada bylo využito několik krátkých zpravodajských pořadů ze stanice 5UA, které byly manuálně přepsány. Další informace jsou v Tabulce 6.6.

Tabulka 6.6: Testovací sady pro ukrajinštinu

Testovací sada	**Velikost**	**Mluvčích**
Studiové nahrávky	57 min	5
VoxForge	40 min	9
Zpravodajství 5UA	53 min	-

Na těchto datech byly otestovány všechny tři systémy a vyhodnoceny pomocí vzorce WER. Akustický model pro každý jazyk byl natrénován na stejném množství dat za využití databáze GlobalPhone, která byla k dispozici. Díky tomu byly všechny modely vytvořeny na stejném typu dat a byly tak zajištěny rovnocenné podmínky pro porovnání. V Tabulce 6.7 jsou zobrazeny výsledky testu. Nejlepšího skóre dosáhl ruský model, který byl tedy následně vybrán jako zdrojový jazyk pro vývoj ukrajinštiny.

Tabulka 6.7: Výsledky multilingválního testu pro výběr vhodného jazyka pro vývoj ukrajinštiny

AM	**Velikost**	**WER [%]**		
		Studiové nahrávky	**VoxForge**	**Zprav. 5UA**
RU	11 hod.	40,3	78,3	59,5
CZ	11 hod.	65,8	92,1	74,8
PL	11 hod.	63,0	91,4	79,7

Příklad výběru zdrojového jazyka pro vývoj bulharštiny

Při výběru jazyka pro vývoj bulharštiny bylo rozhodováno mezi češtinou, slovenštinou, polštinou a chorvatštinou. Dále byla vyzkoušena i kombinace všech těchto jazyků. K testování byla v tomto případě vytvořena jedna sada z nahrávek čtyř rodilých mluvčích v délce 33 minut.

Jelikož má bulharština relativně malou fonetickou sadu, všechny testované jazyky již obsahovaly její fonémy a nemuselo tak dojít k mapování.

Pro trénování akustických modelů bylo použito podobné množství dat z různých zdrojů, multilingvální model byl vytvořen smícháním všech těchto dat. Výsledky testu jsou zobrazeny v tabulce 6.8. Nejlepšího skóre dosáhla chorvatština, která byla jen o několik desetin procenta lepší než multilingvální model.

Tabulka 6.8: Výsledky multilingválního testu pro výběr vhodného jazyka pro vývoj bulharštiny

AM	**Velikost**	**WER** [%]
CZ	10 hod.	28,7
SK	10 hod.	27,1
PL	10 hod.	31,6
HR	10 hod.	25,7
CZ+SK+PL+HR	10 hod.	26,1

6.3.4 Nesupervizovaný přístup tvorby trénovacích dat

V případě, že nejsou k dispozici žádné nahrávky s textem, které by mohly být zpracovány, přichází na řadu možnost tvorby trénovacích dat pouze z nahrávek bez textu pomocí nesupervizovaného přístupu. K tomu je zapotřebí mít již existující akustický model pro daný jazyk s dostatečným množství trénovacích dat.

Trénovací data jsou rozdělena do různých skupin a z nich jsou natrénovány různé akustické modely. Rozdělení probíhá nejlépe podle zdrojů, ze kterých byly vytvořeny, aby každý model byl natrénován na rozdílných typech dat a dospělo se k určité objektivnosti srovnávání. Případně mohou být pro každý model nastaveny i jiné parametry systému.

Zpracovávané nahrávky jsou následně rozděleny na krátké segmenty v místech, kde je ticho či nějaký hluk, a každý segment je následně rozpoznán pomocí všech vytvořených akustických modelů. Pokud se všechny přepisy shodují, je segment považován za správně rozpoznaný a i se svým automaticky vytvořeným fonetickým přepisem přidán do trénovací sady. K tomuto účelu mohou být použity i akustické modely z jiných jazyků, čímž se ale může snížit efektivita.

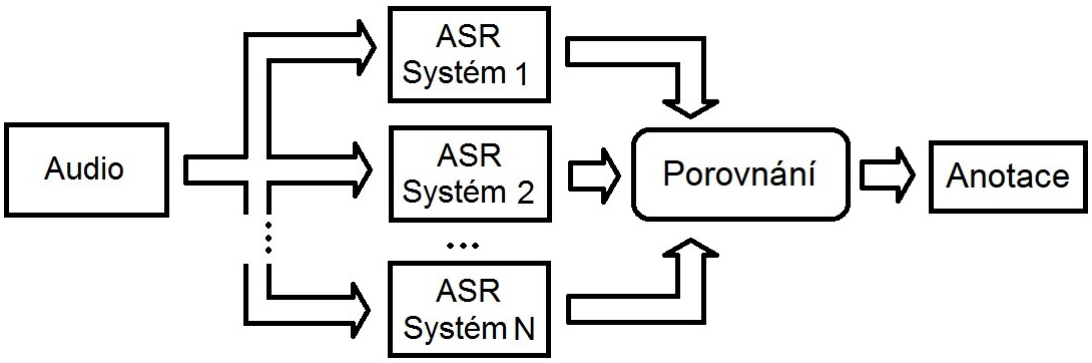
Formálně je proces zapsán následovně:

`Nesupervizované trénování:`

1. `Pro každý dokument`
   - `Proveď segmentaci`
2. `Pro každý segment`
   (a) `Pro každý rozpoznávač`
      - `Přepiš segment`

   (b) `Pokud se všechny výstupy shodují`
      - `Přidej do trénovací sady`
3. `Přetrénuj model`

Tento přístup byl aplikován na zpracování 120 polských televizních pořadů o délce 30 minut. Pro zpracování byly využity 4 rozpoznávače natrénované na různých datech. Celkem bylo tímto přístupem vytěženo 16,4 hodin. Pro dodatečné vyhodnocení byla manuálně zkontrolována náhodná podmnožina dat a bylo zjištěno, že 9 z 10 segmentů bylo správně přepsáno a zbývající segmenty obsahovaly pouze marginální chyby ve výslovnosti. Nesupervizované těžící schéma je zobrazeno na obrázku 6.3.



Obrázek 6.3: Nesupervizované těžení akustických dat

6.4 Identifikace a filtrování cizích jazyků v nahrávce

Při automatickém stahování nahrávek může nastat případ, kdy jsou buď celé nahrávky, nebo jen některé jejich úseky v jiném jazyce. Například ve zpravodajství se může jednat o záznamy promluv v cizím jazyce doplněných titulky. V takovýchto případech, kdy se v nahrávce vyskytne jazyk zcela odlišný od zpracovávaného jazyka, je naprosto minimální pravděpodobnost, že by se přepis této promluvy shodoval

s porovnávaným textem, a tato část nahrávky by tak prošla zpracováním a dostala se do trénovacích dat. Tyto případy neprojdou použitým algoritmem, a proto není nutné se jimi při těžení dat nijak zabývat.

Vážnější situace nastává při vývoji jazyků z multilingválních zemí, jako je například Ukrajina či Bělorusko, kde z různých důvodů velké množství obyvatel mluví rusky, a tudíž je i ruština často používána ve zpravodajství, televizních a rádiových pořadech a dalších zdrojích využívaných pro tvorbu trénovacích dat. V Bělorusku je dokonce situace taková, že více než 70 % obyvatel považuje ruštinu za svůj mateřský jazyk a ruština je tak prakticky hlavním jazykem této země. Jelikož jsou si tyto jazyky velmi blízké, je zde určitá šance, že některé věty budou správně rozpoznány a do trénovacích dat by se tak dostal ruský jazyk. Proto bylo při zpracování běloruštiny tedy zapotřebí zavést identifikaci a filtrování cizích jazyků, aby bylo docíleno čistě běloruských trénovacích dat.

Existuje několik metod identifikace jazyka, které ale většinou vyžadují trénovací data, na nichž jsou natrénovány modely pro jednotlivé jazyky. To je ale v případě počáteční fáze vývoje systému pro nový jazyk nemožné, když nejsou zatím k dispozici žádná akustická data. Proto byl navržen postup, který byl odvozen z postupu použitého pro češtinu v [79], kdy je vytvořen multilingvální systém z jazyků, které mají být identifikovány.

V tomto případě došlo ke smíchání ruštiny, ukrajinštiny a běloruštiny. K tomu účelu byl vytvořen speciální slovník smícháním sto tisíc nejčetnějších slov z každého jazyka a každé slovo bylo označeno svým jazykem. Jazykový model byl natrénován smícháním textových korpusů ze všech jazyků a každé slovo bylo označeno podle jeho jazyka. Fonetická abeceda byla použita ruská, jelikož ukrajinská a běloruská z ní vycházejí a liší se pouze dvěma fonémy, které byly přemapovány. Akustický model byl nakonec natrénován smícháním ruských a ukrajinských dat, jelikož žádná běloruská data zatím nebyla k dispozici. Následná identifikace jazyka proběhla podle největšího množství slov z daného jazyka, která byla rozpoznána v nahrávce pomocí takto vytvořeného systému, při překročení určité prahové hodnoty počtu slov.

Pro vyhodnocení této metody a nalezení prahové hodnoty byla vytvořena testovací sada, z jejíchž výsledků byla vypočtena matice záměn (Confusion matrix) pro různé prahové hodnoty. Testovací sada obsahovala 300 nahrávek, jejichž jazyk byl označen rodilými mluvčími. Následně byly vypočteny hodnoty pro míru shody (Accuracy), přesnost (Precision) a senzitivitu (Recall), dle vztahů 6.6, 6.7 a 6.8, kde TP (True positive) je počet nahrávek správně vyhodnocených jako posuzovaný jazyk, FP (False positive) je počet nahrávek špatně vyhodnocených jako posuzovaný jazyk, FN (False negative) je počet nahrávek špatně vyhodnocených jako jiný jazyk a N je počet nahrávek, které prošly prahem filtru. K těmto hodnotám byly dále dopočteny hodnoty pro chybovost (Errors), pouze jakožto doplněk přesnosti dle vztahu 6.9, a výtěžnost (Yield) dle vztahu 6.10 udávající, jaké množství nahrá-

vek prošlo filtrem z celkového počtu nahrávek pro daný jazyk (zde tedy závisí na prahové hodnotě).

$$ACC = \frac{TP}{N} \tag{6.6}$$

$$PREC = \frac{TP}{TP + FP} \tag{6.7}$$

$$REC = \frac{TP}{TP + FN} \tag{6.8}$$

$$ERR = 1 - PREC \tag{6.9}$$

$$YLD = \frac{TP}{N_L} \tag{6.10}$$

Výsledná tabulka 6.9 zobrazuje hodnoty pro běloruštinu při využití tří různých prahových hodnot. Z výsledků lze vyčíst, že i přes vysokou přesnos, kdy naprostá většina nahrávek identifikovaných jako běloruština je opravdu v běloruštině, spousta dalších běloruských nahrávek je mylně identifikována jako jiný jazyk a celková výtěžnost je tak nízká. Na základě těchto hodnot byl vybrán práh 40 %, kde je sice 4% možnost zanesení jiného jazyka do trénovacích dat, ale výtěžnost je při této hodnotě prahu o 14 % vyšší než při s prahem 50 %.

Tabulka 6.9: Výsledky experimentu identifikace jazyka pro běloruštinu

Práh	**ACC** [%]	**PREC** [%]	**REC** [%]	**ERR** [%]	**YLD** [%]
30 %	50,63	90,74	52,13	9,26	52,12
40 %	53,62	95,56	55,13	4,44	45,74
50 %	60,82	100,00	60,00	0,00	31,91

6.5 Trénování akustického modelu

Jak již bylo zmíněno v kapitole 1.4, pro účely této práce byl využit systém rozpoznávání řeči vyvinutý na Ústavu informačních technologií a elektroniky Fakulty mechatroniky TUL [80]. Tento systém byl vyvinut pro češtinu, ale díky jeho modulární struktuře je možné ho jednoduše adaptovat na nové jazyky. V dnešní době je schopen pracovat se slovníkem velikosti půl milionu slov v reálném čase. Dále umožňuje práci jak s GMM, tak s DNN modely.

GMM jsou trénovány jako 32-složkové gaussovské směsi a jako příznakové vektory vstupních segmentů jsou využity 39-dimenzionální MFCC (Mel-Frequency Cepstral

Coefficients). Vstupní signál je segmentován po 25 ms s překryvem 10 ms. Pro trénování je použit nástroj HTK Speech Recognition Toolkit.[5] GMM jsou využívány pro zarovnání dat pro trénování DNN.

V případě DNN je využita standardní architektura s pěti skrytými vrstvami v konfiguraci 1024-768-768-512-512 a aktivační funkcí ReLU. Vstupní signál je parametrizován pomocí 39-dimenzionálních Log filter banks. Segmentace vstupu je stejná jako v případě GMM. Pro dodání kontextu je každý frame doplněn pěti předchozími a pěti následujícími framy. Pro trénování je využita knihovna Torch.[6]

6.6 Výsledky aplikace popsaných metod a postupů na východoslovanské jazyky

Na závěr jsou zde pro ukázku aplikace popsaných metod popsány některé vybrané kroky z vývoje akustických modelů pro všechny tři východoslovanské jazyky. Údaje ohledně korpusu, slovníku a jazykového modelu jsou popsány v kapitole 5. zabývající se textovou částí.

Jako první byl vývoj zahájen na ruštině za pomoci českého akustického modelu. Ruština byla následně použita pro vývoj ukrajinštiny a ta následně pro běloruštinu. Hlavním cílem bylo vytvořit systém určený především pro monitoring médií, zpravodajské pořady byly proto nejvhodnějšími daty. Krom toho byly zpracovány a použity i záznamy z ruského parlamentu a ruský GlobalPhone.

6.6.1 Ruština

Jako první byla zpracována ruština za využití českého akustického modelu. Pro začátek byl k dispozici ruský řečový korpus GlobalPhone. Ten ale neobsahuje fonetické transkripce, a proto bylo nutné je vytvořit s pomocí českého akustického modelu.

Tabulka 6.10: Vývoj AM pro ruštinu na databázi GlobaPhone

AM	**Velikost**	**WER** [%]
CZ	10 hod.	58,3
RU	5 hod.	30,7
RU	10 hod.	21,6
RU	23 hod.	18,2

Ruská fonetická abeceda byla namapována na českou a podle toho byly upraveny výslovnosti ve slovníku. Následně byl aplikován popsaný algoritmus automatického těžení dat dohromady s manuální kontrolou a korekcí přepisů pro kontrolu a urychlení procesu. Po vytěžení prvních pěti hodin byl systém převeden na čistě ruský

[5] http://htk.eng.cam.ac.uk/
[6] http://torch.ch

a těžení pokračovalo. Postupný proces vývoje je zobrazen v tabulce 6.10.

Pro vyhodnocování tohoto procesu bylo z GlobalPhone odebráno 10 mluvčích, jejichž přepisy byly manuálně zkontrolovány a následně využity pro testování. Údaje o testovací a výsledné trénovací sadě jsou v tabulce 6.11.

Tabulka 6.11: Statistika ruské trénovací a testovací sady využívající databáze Globalphone

	Trénovací sada	**Testovací sada**
Počet mluvčích	105	10
Počet nahrávek	10644	1202
Délka	23 hod.	2,4 hod.

Následně bylo pokračováno ve zpracování dalších zdrojů z ruských zpravodajských webů a z ruského parlamentu. Celkové množství získaných a zpracovaných dat shrnuje tabulka 6.13.

6.6.2 Ukrajinština

Ruský model byl následně využit pro vývoj ukrajinštiny. Zde nebyla k dispozici žádná data pro začátek, pouze testovací sady využité pro výběr jazyka. Ty byly využity společně s ruskými daty k vytěžení prvních pěti hodin. Následně byly odebrány z trénovací sady a dále využívány pouze pro testování průběhu vývoje.

Ukrajinská fonetická sada byla navržena jako podmnožina ruské fonetické sady s rozdílem ukrajinského "měkkého" /*c*/, které bylo namapováno na standardní ruské "tvrdé" /*c*/.

Zpracovávány zde byly opět zpravodajské pořady a celkové množství použitých a zpracovaných dat je uvedeno v tabulce 6.13. Postupný průběh vývoje ukrajinského modelu je zobrazen v tabulce 6.12.

Tabulka 6.12: Vývoj AM pro ukrajinštinu

AM	**Velikost**	**WER** [%]		
		Studiové nahrávky	**VoxForge**	**Zprav. 5UA**
RU	25 hod.	40,3	78,3	59,5
UK+RU	30 hod.	28,3	69,6	30,0
UK+RU	42 hod.	26,3	67,9	25,5
UK	8 hod.	29,7	68,7	28,2
UK	22 hod.	26,1	58,5	22,3
UK	40 hod.	26,7	57,3	19,7

6.6.3 Běloruština

Pro vývoj běloruštiny byla popsaným způsobem vybrána ukrajinština jakožto nejvhodnější zdrojový jazyk a vývoj probíhal obdobným způsobem, jako u předchozích jazyků. Nicméně jak už bylo zmíněno, u běloruštiny bylo zapotřebí odfiltrovat nahrávky v ruštině a ukrajinštině, které se objevovaly ve velkém množství. Tím se objem celkově vytěžených dat snížil na 16 hodin, jak ukazuje tabulka 6.13 společně s dalšími údaji.

Tabulka 6.13: Statistika zpracování akustických dat pro východoslovanské jazyky

Jazyk	**RU**	**UK**	**BE**
Zdrojů	7	6	3
Stažených dat	4627 hod.	4015 hod.	955 hod.
Celkem vyřezaných úseků	192 hod.	161 hod.	48 hod.
Celkem vytěžených dat	58,3 hod.	48 hod.	16,4 hod.

6.6.4 Vyhodnocení

Z tabulky 6.13 zobrazující statistiku akustických dat lze vyčíst, že bylo staženo obrovské množství dat, ale jen zlomek byl nakonec vytěžen. Důvodů je k tomu několik.

Hlavní příčinou byla použitá data. Většina dat byly automaticky stažené pořady z webových stránek spolu s textem, který se u nich vyskytoval. Ten ale ve většině případů vůbec neodpovídal promluvě v nahrávkách a byl většinou jen slovní popis obsahu nahrávky, nebo jakýkoliv další možný text k danému tématu.

Bylo tak potřeba zpracovat ohromné množství dat, aby byly nalezeny alespoň některé úseky, kde se text shoduje s promluvou v nahrávce. Druhým důvodem, proč byla výtěžnost tak malá, je také fakt, že při tvorbě přepisů a jejich porovnávání s textem není systém 100% úspěšný. Tedy ne vše přepíše správně, a tudíž nemusí být nalezeny úplně všechny shodující se části.

Další důvod je také ten, že stažená data obsahují různé znělky, hudbu, nebo to jsou jen videa bez promluvy a tak i přes to, že byla zpracována, nemohlo z nich být vytěženo nic.

Na druhou stranu tento přístup zaručuje vytvoření trénovacích dat s velmi přesnými fonetickými přepisy. Jak je důležitá přesnost trénovacích dat, bylo ověřeno v experimentech popsaných v následující části.

6.7 Analýza vlivů některých aspektů automatizovaného vývoje na přesnost rozpoznávání

Souběžně s vývojem jednotlivých systémů pro slovanské jazyky bylo provedeno několik různých experimentů, které měly prověřit efektivitu navržených postupů. Jednalo se na jedné straně o simulované experimenty, zkoumající vliv různých typů chyb ve fonetických přepisech na výsledné skóre rozpoznávače, a na straně druhé bylo provedeno několik porovnávacích experimentů s reálnými daty. Zde je uvedeno jen několik klíčových ze všech provedených testů. Všechny experimenty nicméně potvrdily, že přesnost anotace trénovacích dat je zásadní pro dosažení co nejvyšší úspěšnosti rozpoznávání.

6.7.1 Simulované experimenty zkoumající přesnost fonetických přepisů

Tato série experimentů zjišťovala, jaký vliv mají chyby ve fonetickém přepisu v trénovacích datech na přesnost rozpoznávání. Byly provedeny dva typy experimentů upravující fonetické přepisy. Jeden náhodně zaměňující fonémy ve slovech a druhý, kde se náhodně přidávala nebo odebírala slova. Experimenty byly prováděny na obou GMM i DNN modelech.

V trénovacích datech bylo vždy uměle vytvořeno určité množství chyb, následně byl standardním způsobem natrénován GMM model a pomocí něho byly zarovnány nahrávky s přepisem a natrénován DNN model.

Experimenty byly provedeny na 4 slovanských jazycích (čeština, polština, chorvatština, ruština), pro které byly k dispozici databáze Globalphone, aby výsledky byly porovnatelné. Z nich bylo vždy vybráno prvních deset mluvčích pro testování a zbytek pro simulaci chyb a trénování. Výsledky nicméně byly velmi podobné pro všechny jazyky a tak jsou následně zobrazeny jen hodnoty pro polštinu.

Tyto experimenty měly za úkol prověřit, do jaké míry mohou chyby v přepisech ovlivnit akustický model a následně přesnost rozpoznávače. Výsledky experimentů nám říkají, jak zásadně je třeba řešit přesnost přepisů v průběhu vývoje systému.

Záměny fonémů ve fonetických přepisech

Jako první byly simulovány chyby fonetické anotace v trénovacích datech. Náhodně byly měněny fonémy za jiné z použité fonetické sady v různých množstvích. Vyzkoušeny byly změny od 1 % až do 50 % všech fonémů.

Následně bylo testováno i rozložení chyb, kde byly chyby provedeny u dvou a tří po sobě následujících fonémů. Tím byly simulovány chyby typu špatné produkční pravidlo, které při transkripci ovlivní celý shluk fonémů (např. kvůli znělosti).

V tabulce 6.14 jsou zobrazeny výsledky experimentu. K vyhodnocení byla použita míra WER. Z výsledků je patrné, že už i malé množství chyb mírně ovlivňuje přesnost rozpoznávání, z čehož vyplývá, že pro co největší přesnost systému jsou zapotřebí co nejpřesnější přepisy.

Zároveň je ale také vidět, že i při velkém množství chyb není propad až tak zásadní, jak by se dalo očekávat. To poukazuje na robustnost použitých modelů a trénovacích algoritmů. Také je vidět, že DNN modely jsou více náchylné k chybám než GMM. Při 15 % záměn mají podobné skóre jako GMM a s více záměnami přesnost klesá mnohem razantněji.

Tabulka 6.14: Výsledky experimentu záměny fonémů ve fonetických přepisech

	WER [%]					
	1 foném		**2 fonémy**		**3 fonémy**	
Změněno fonémů	**GMM**	**DNN**	**GMM**	**DNN**	**GMM**	**DNN**
0 %	13,80	12,55	13,80	12,55	13,80	12,55
1 %	14,32	13,24	14,40	13,50	15,09	13,19
2 %	14,62	13,92	14,54	13,43	14,43	13,18
5 %	16,64	14,19	14,70	14,54	15,39	14,31
10 %	15,43	15,38	16,17	15,38	16,21	15,15
15 %	16,23	16,84	16,43	16,47	16,74	16,00
20 %	17,07	17,63	16,60	17,06	16,80	16,74
30 %	18,27	20,61	17,21	19,11	17,95	18,81
40 %	20,23	28,55	18,59	22,10	19,44	20,66
50 %	24,00	32,91	20,67	26,90	21,31	24,08

Přidávání a odebírání slov ve fonetických přepisech

Další typ experimentu se zaměřil na chyby typu chybějících nebo přebývajících slov v přepisech. Tento typ chyb se může objevit z několika důvodů. Při ruční anotaci je nějaké slovo přeslechnuto (zejména krátká slova), nebo naopak anotátor přidá slovo, které nebylo řečeno, ať už z důvodu, že si myslí, že ho slyšel, anebo se snaží doplnit správné chybějící slovo, které mluvčí omylem vynechal, aby věta dávala smysl.

Při automatické tvorbě trénovacích dat se často stává, že referenční text obsahuje gramaticky správnou větu, kterou ale mluvčí neřekl, nebo naopak řekl něco navíc a referenční text slovo neobsahuje. Jelikož proces tvorby dat popsaný v této práci primárně cílí na používání 100% přesných trénovacích dat, bylo vhodné otestovat i situace, kdy by byly přidány přepisy s chybějícími a přebývajícími slovy.

Pro simulaci toho problému byla náhodně mazána nebo opakována slova v přepisech opět do určitého množství a vše testováno. Tabulka 6.15 obsahuje výsledky těchto testů vyhodnocené mírou WER. Z výsledků je opět vidět, že přesnost systému klesá s množstvím chyb, ale opět ne tak dramaticky, jak by se dalo očekávat.

Tabulka 6.15: Výsledky experimentu přidávání a odebírání slov ve fonetických přepisech

	WER [%]			
	Odebírání		**Přidávání**	
Množství změn	**GMM**	**DNN**	**GMM**	**DNN**
0 %	13,80	12,55	13,80	12,55
1 %	15,05	13,29	14,90	12,94
2 %	14,97	13,45	14,97	12,92
5 %	15,57	14,47	15,32	13,62
10 %	15,86	15,23	15,72	14,36
15 %	16,48	15,91	16,49	14,29
20 %	16,72	16,95	16,76	15,40
30 %	17,45	19,20	17,72	16,05
40 %	19,41	21,11	18,96	17,31

6.7.2 Experimenty s reálnými daty

Další testy vlivu nepřesné fonetické anotace akustických trénovacích dat byly provedeny na reálných datech obsahujících různé typy chyb. Tyto experimenty měly prověřit efektivitu navržených postupů pro tvorbu akustických trénovacích dat.

Experiment s polštinou

První experiment byl proveden s polskou databází Clarin, kde bylo při automatickém testování objeveno mnoho chyb v přepisech, a to jak chybějících nebo přebývajících slov, tak například i chyby ve špatném skloňování. Celá databáze tedy byla zpracována popsaným procesem automatické tvorby dat, kde byly pro trénování přijaty jen 100% přesně souhlasící nahrávky.. Na těchto datech byl natrénován akustický model a porovnám s modelem natrénovaným na původních fonetických anotacích.

Modely byly otestovány na polské testovací sadě popsané výše, která byla vytvořena z databáze Globalphone. Automatickým procesem bylo vytěženo 31 z původních 56 hodin. Tabulka 6.16 ukazuje, že modely natrénované na těchto vytěžených datech mají o několik procent lepší přesnost rozpoznávání díky velké přesnosti přepisů, a to s polovičním množstvím dat než původní distribuovaná databáze s chybami.

Tabulka 6.16: Experiment s polskou databází Clarin

Clarin	Původní anotace		Automatické anotace	
Velikost	56 h		31 h	
Typ modelu	GMM	DNN	GMM	DNN
WER [%]	24,53	17,72	19,26	14,54

Experiment s ruštinou

Další experiment byl proveden při vývoji ruštiny s ruskou databází GlobalPhone, která byla porovnána s automaticky získanými daty. Tento experiment měl ukázat, jak kvalitní je akustický model natrénovaný na automaticky získaných datech z internetu oproti dedikované řečové databázi.

Pro testování bylo použito prvních deset mluvčích z databáze Globalphone a zpravodajská testovací sada pro ruštinu, která je detailně popsána v tabulce 7.1 v následující kapitole popisující souhrnné výsledky. První akustický model byl natrénován na zbylých 105 mluvčích z GlobalPhonu a druhý na všech datech automaticky vytěžených z dat z internetu.

Oba modely byly testovány na obou testovacích sadách a z tabulky 6.17 lze vyčíst, že model natrénovaný na automaticky vytěžených datech předčil databázi Globalphone v obou testech. Samozřejmě mnohem výrazněji na zpravodajské testovací sadě, jelikož byl trénován na podobném typu dat.

Tabulka 6.17: Experiment s ruskou databází Globalphone a automaticky získanými daty

	Trénovací sada / WER [%]	
Testovací sada	Globalphone (22 h)	Automaticky získaná data (58 h)
GlobalPhone (150 min)	18,21	14,30
Zpravodajství (93 min)	50,74	23,02

Experiment s běloruštinou

Další experiment byl proveden na běloruštině. Jelikož se v Bělorusku běžně používá ruština více než běloruština, byl také problém získat dostatečné množství dat pro trénování. Nakonec automatickým zpracováním prošlo pouze 16 hodin použitelných dat. Vznikla proto otázka, jak by byl model ovlivněn, kdyby se do trénovací sady přidaly i úseky, které neměly 100% shodu.

Byly tak vybrány tři různé hodnoty shody mezi přepisem z rozpoznávače a porovnávaným textem. Všechny úseky, které měly shodu stejnou nebo vyšší, byly přidány do trénovací sady. Experiment byl testován na běloruské zpravodajské testovací sadě popsané v tabulce 7.1 v další kapitole.

Tabulka 6.18 zobrazuje hodnoty WER pro jednotlivé trénovací sady. Je z ní jasně vidět, že i při vyšším množství dat se celková přesnost rozpoznávání postupně zhoršuje kvůli množství chyb v trénovacích datech.

Tabulka 6.18: Výsledky experimentu s trénováním AM pro běloruštinu na nepřesných datech

Přesnost dat	**Velikost**	**WER [%]**
100 %	16,6 hod.	35,9
95+ %	19,2 hod.	38,2
90+ %	20,7 hod.	39,3
80+ %	27,2 hod.	41,8

6.7.3 Závěry z vyhodnocování

Výše uvedené výsledky zřetelně ukazují, že navržený způsob automatické tvorby trénovacích dat pro vývoj akustických modelů je funkční, robustní a vede k výsledkům, které jsou srovnatelné s výsledky získatelnými ručně anotovanými textovými a fonetickými přepisy. Popsaný přístup je plně automatizovatelný a navíc dovede odhalit případné chyby anotací. Je schopen učit se na veřejně přístupných datech, která jsou navíc bližší cílovým aplikacím než standardní databáze čtené řeči.

7 Souhrnné výsledky dokumentující vývoj ASR systémů pro slovanské jazyky

Pro závěrečné otestování vyvinutých systémů byly vytvořeny standardizované testovací sady. Cílem bylo systémy otestovat na reálných datech z televizního a rozhlasového vysílání, aby bylo možné posoudit jejich použitelnost v reálném provozu při monitorování médií. Systémy byly vytvořeny pro všech 13 slovanských jazyků. Nicméně pro bosenštinu a černohorštinu byl kvůli nedostatku akustických dat použit srbochorvatský akustický model. Rovněž se u těchto jazyků nepodařilo sestavit standardizovaná testovací data, jelikož nebyl nalezen rodilý mluvčí pro jejich validaci. Proto bylo výsledné testování provedeno pouze na 11 slovanských jazycích.

Je nutné dodat, že systém pro češtinu byl na pracovišti vyvíjen od 90. let minulého století a již byl úspěšně nasazen v mnoha výzkumných i komerčních aplikacích. Na jeho základě byly pak vytvořeny systémy pro slovenštinu a polštinu a následně započal vývoj systémů pro východoslovanské a jihoslovanské jazyky, kterého jsem se už v rámci týmu zúčastnil.

V této kapitole je popsán výběr a tvorba testovacích dat, následuje popis a statistiky vytvořených systémů a jejich výsledky dosažené na testovacích datech.

Tabulka 7.1: Statistika testovacích sad pro slovanské jazyky

Jazyk	**Kód**	**Délka [min]**	**Počet slov**
Čeština	CZ	95	13494
Slovenština	SK	92	12365
Polština	PL	105	14742
Slovinština	SL	109	14943
Chorvatština	HR	104	15319
Srbština	SR	89	12791
Makedonština	MK	94	12916
Bulharština	BG	100	15197
Ruština	RU	93	12277
Ukrajinština	UK	75	9440
Běloruština	BE	82	11716

7.1 Standardizovaná testovací sada

Pro testování byly vybrány zpravodajské pořady z hlavních televizních a rozhlasových stanic v jednotlivých zemích. Pro každý jazyk byly zvoleny 3 pořady z alespoň dvou různých stanic v celkové délce okolo 90 minut. Testovací data byla vytvořena z dat odvysílaných několik měsíců (v několika případech roků) později po natrénování akustických a jazykových modelů, aby byla zajištěna skutečná nezávislost experimentů.

Jedná se o kompletní pořady (od otevírací až po závěrečnou znělku) obsahující všechny typy zvuků a promluv běžně se vyskytujících v těchto zprávách - tj. čistá řeč ve studiu, řeč s hudbou či hlukem na pozadí, spontánní řeč lidí na ulici, dabovaná řeč s původní promluvou na pozadí a podobně. Referenční texty byly vytvořeny a zkontrolovány rodilými mluvčími pro všech 11 testovaných jazyků.

Zároveň byly v referenčních textech označeny úseky v jiném než cílovém jazyce. U některých pořadů se může jednat o pasáže v cizím jazyce s přidanými titulky. U dalších, především u východoslovanských zemí, které se vyznačují silně bilingválním prostředím, se často objevuje používání více jazyků v rámci jednoho pořadu. Konkrétně je to především užívání ruštiny v ukrajinských a běloruských pořadech, ale v jednom běloruském pořadu se objevil i rozhovor v polštině.

Tabulka 7.1 zobrazuje statistiky testovacích sad pro jednotlivé jazyky společně s použitým jazykovým kódem dle ISO 639-1. Testovací sady byly zároveň zpřístupněny a jsou veřejně dostupné[1].

Tabulka 7.2: Charakteristiky vytvořených modulů pro slovanské jazyky

Jazyk	**Autorův podíl**	**Velikost korpusu [GB]**	**Velikost slovníku [tis.]**	**Počet fonémů**	**Velikost trénovací sady [h]**	**Počáteční jazyk**
CZ	0	6,2	388	41	1050	x
SK	0	2,9	302	41	118	CZ
PL	1	3,00	303	36	58	CZ
SL	1	0,91	300	32	42	HR
HR	1	1,10	304	32	45	CZ
SR	1	1,23	307	32	40	CZ
MK	1	0,83	265	33	40	BG
BG	1	0,98	283	33	41	HR
RU	1	0,98	326	53	58	CZ
UK	2	0,75	324	39	48	RU
BE	2	0,28	293	36	16	UK

[1] https://owncloud.cesnet.cz/index.php/s/qLTs9K5LAeqIZAV

7.2 Charakteristiky vytvořených modulů

V tabulce 7.2 jsou vypsány výsledné charakteristiky všech modulů systému pro jednotlivé jazyky. Jako první je uvedeno v jaké míře jsem se na kterém jazyce osobně podílel v průběhu jeho vývoje. Použito je kódování 0 - nepodílel, 1 - ve spolupráci se školitelem, 2 - samostatně.

Dále je uvedena velikost textového korpusu, ze kterého byl trénován jazykový model, velikost použitého slovníku, rozsah fonetické sady a množství akustických dat použitých pro trénování akustického modelu.

Nakonec je uveden počáteční jazyk, který byl využit v počátečních fázích vývoje akustického modelu daného jazyka (bootstrapping). Jelikož byly systémy pro jednotlivé jazyky vyvíjeny průběžně, vždy bylo vybíráno pouze z dostupných a již dobře fungujících systémů, a tak tedy ve většině případů byla použita čeština.

Trénovací data pro západoslovanské jazyky, a to především pro češtinu a slovenštinu, výrazně převyšují množství dat pro ostatní jazyky, jelikož byly vyvíjeny mnohem delší dobu a jsou již komerčně nasazeny v mnoha aplikacích. Je však třeba říci, že v testech byly použity pouze vývojové verze systémů, nikoliv aktuální produkční verze, které se liší zejména novými typy neuronových sítí využitých v akustických modelech a rovněž pravidelně aktualizovanými slovníky a jazykovými modely. Tyto modifikace jsou již řešeny jinými členy týmu.

Tabulka 7.3: Výsledky rozpoznávání na vytvořených testovacích sadách

Jazyk	**OOV** [%]	**OOL** [**min**]	**WER GMM** [%]	**WER DNN** [%]
CZ	0,87	2,6	24,01	14,72
SK	1,37	1,2	28,05	18,42
PL	0,92	2,3	25,91	20,80
SL	0,68	4,0	23,84	16,16
HR	0,99	0,7	27,11	20,07
SR	0,41	0,3	26,25	18,90
MK	0,52	1,6	26,43	14,54
BG	0,61	0,1	27,66	20,86
RU	2,18	3,7	33,76	22,08
UK	2,75	8,9	36,32	30,15
BE	3,12	15,7	41,83	35,95

7.3 Výsledky rozpoznávání na vytvořených testovacích sadách

Na testovacích sadách byly následně ověřeny všechny systémy. Tabulka 7.3 zobrazuje kromě dosažených výsledků i míru OOV referenčních textů a délky úseků v jiném než cílovém jazyce. Úseky označené jako OOL byly vyřazeny z vyhodnocení pro zachování objektivity.

Výsledné hodnoty WER jsou uvedeny jak pro GMM tak pro DNN modely z důvodu, že v průběhu vývoje byly využívány oba typy modelů a rovněž z důvodů popsaných v předchozích kapitolách. Nicméně DNN modely vždy dosahují vyšší úspěšnosti rozpoznávání a jsou využívány ve finálních verzích systémů.

Z výsledků je patrné, že nejlepších výsledků dosahuje samozřejmě čeština, která se může opřít o mnohem více trénovacích dat a rovněž precizněji připravený slovník. Většina dalších jazyků se však vešla pod hranici 20 % WER, což již znamená poměrně dobře použitelné přepisy pro účely analýz a monitoringu, a také jako základ pro případnou efektivní ruční editaci. Velice dobré výsledky jsou vidět u makedonštiny a slovinštiny, což do jisté míry souvisí i s akustickou kvalitou dat, výslovností mluvčích, obsahem, apod. Tam, kde byl větší podíl promluv profesionálních řečníků snímaných ve studiu, mohl systém dosáhnout mnohem lepších výsledků než tam, kde byly ve větší míře využity spontánní rozhovory z ulice či jinak rušného prostředí.

Dále je vidět, že hodnoty WER u východoslovanských jazyků byly obecně vyšší, což je dáno jak větší složitostí těchto jazyků (oproti prvním dvěma skupinám), tak i vyšší výslovnostní variabilitou způsobenou regionálními odlišnostmi a bilingvismem v těchto zemích. Vyšší je i míra OOV, a to i při větším počtu slov ve slovníku. Nejhoršího skóre dosáhla běloruština, což bylo způsobeno především nízkým množstvím trénovacích dat, které se podařilo získat procesem automatického těžení (důvody jsou podrobněji vysvětleny v předchozí kapitole).

8 Příklady aplikace metod na další jazyky

V této kapitole je krátce zdokumentována aplikace popsaných metod na jiné než slovanské jazyky. Systémy pro tyto jazyky byly vyvíjeny ať už v rámci různé spolupráce nebo jen pro otestování navržených metod a postupů na jiném typu jazyka.

Tabulka 8.1: Základní přehled dalších zpracovaných jazyků

Jazyk	Španělština	Lotyština	Albánština
Rodina	Románské	Baltické	Albánské
Typ	Analytické	Flektivní	Flektivní
Mluvčích	558 mil.	1,75 mil.	7,5 mil.
Písmo	Latinka	Latinka	Latinka
Kód	ES	LV	SQ (ALB)

Vybrány byly tři jazyky, na kterých jsem pracoval samostatně s částečnou pomocí rodilých mluvčích při tvorbě dat a konzultacích různých detailů jazyka. První byla vybrána španělština, jakožto jeden z největších světových jazyků, dále lotyština, která má určitým způsobem blízko ke slovanským jazykům, a na závěr albánština, osamocený evropský jazyk s omezenými zdroji. Tabulka 8.1 shrnuje základní údaje o těchto jazycích. V tabulce 8.2 je pro ukázku uvedena část všeobecné deklarace lidských práv v těchto jazycích.

Tabulka 8.2: Všeobecná deklarace lidských práv v dalších zpracovaných jazycích

ES	Todos los seres humanos nacen libres e iguales en dignidad y derechos y, dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.
LV	Visi cilvēki piedzimst brīvi un vienlīdzīgi savā pašcieņā un tiesībās. Viņi ir apveltīti ar saprātu un sirdsapziņu, un viņiem jāizturas citam pret citu brālības garā.
SQ	Të gjithë njerëzit lindin të lirë dhe të barabartë në dinjitet dhe në të drejta. Ata kanë arsye dhe ndërgjegje dhe duhet të sillen ndaj njëri tjetrit me frymë vëllazërimi.

8.1 Španělština

Španělština je klasickým zástupcem jednoho z nejpoužívanějších jazyků na světě, díky čemuž je k dispozici velké množství zdrojů. Zároveň je ideálním zástupcem analytického jazyka, kam patří například angličtina, čínština či další románské jazyky. Tím je vhodným kandidátem k otestování popsaných metod, které byly primárně navrženy a aplikovány na slovanské jazyky.

Analytické jazyky se vyznačují tím, že gramatický význam věty je dán především pořadím slov a užitím předložek, na rozdíl od flektivních jazyků, kde je dán i tvarem slov. Tím odpadá potřeba velkého množství slovních tvarů a výslovnostní slovník tak může obsahovat mnohem méně slov než v případě slovanských jazyků. Nicméně španělština využívá do velké míry flexe při časování sloves a rozlišuje mezi 14 slovesnými časy (7 základních a 7 doplňkových), které jsou dále ovlivněny osobou, číslem, způsobem, rodem a videm. Tím slovní zásoba částečně narůstá, ale stále je mnohem menší než u slovanských jazyků.

Španělská fonetika je mnohem jednodušší než u slovanských jazyků, obsahuje menší množství fonémů a vztah mezi psanou a mluvenou formou je velmi přímý, takže jej lze popsat několika jednoduchými produkčními pravidly. V mluvě se rozlišuje přízvuk, který je většinou na předposlední slabice. Nicméně se může objevit i jinde, čímž může měnit význam slova. Při tom je vyznačen čárkou nad samohláskou.

Španělština díky velkému množství mluvčích a dalším historickým okolnostem zahrnuje velké množství dialektů a s tím spojené i různé varianty výslovností, které bylo potřeba zohlednit ve výslovnostním slovníku.

Díky rozšířenosti španělštiny bylo k dispozici velké množství textových i akustických zdrojů. Cílem bylo vytvořit systém primárně určený pro dialogové systémy a hlasové ovládání. Systém byl proto vyvíjen především za využití volně dostupných řečových databází a audioknih a jen doplněn zpravodajskými pořady. Vývoj byl zahájen za využití češtiny, která obsahuje dostatečné množství fonémů, které bylo možné namapovat na španělské. Pro testování bylo využito části řečové databáze Albaycin [81]. Finální výsledky společně s dalšími charakteristikami vytvořených modulů shrnuje tabulka 8.4.

8.2 Lotyština

Lotyština je baltický jazyk užívaný v Lotyšsku a má poměrně malý počet mluvčích. Baltické jazyky jsou někdy slučovány se slovanskými do Balto-slovanské rodiny na základě pravděpodobného společného historického vývoje. Nicméně lotyština je historicky ovlivněna jak slovanskými jazyky, tak i němčinou. Jedná se o flektivní jazyk s gramatikou velmi podobnou slovanským jazykům a tak byl očekáván podobný průběh vývoje.

Díky komplexní morfologii obsahuje lotyština velké množství slov a slovních variant, které bylo potřeba přidat do slovníku. Zároveň tak má i relativně volný slovosled. Foneticky má velmi blízko slovanským jazykům, obsahuje prakticky stejné fonémy jako západoslovanské jazyky a také rozlišuje mezi krátkými a dlouhými samohláskami. Čeština proto byla nejvhodnější kandidát jako počáteční jazyk.

I přesto, že je Lotyšsko malá země s malým počtem obyvatel, podařilo se najít velké množství dat vhodných ke zpracování. Primárně byly využity zpravodajské pořady a parlamentní archivy. Průběh vývoje probíhal obdobně jako u slovanských jazyků. Pro testování bylo využito pořadů TEDx s celkem přesnými titulky o délce 2 hodiny a k tomu bylo vytvořeno 30 minut studiových nahrávek rodilých mluvčích. Výsledné charakteristiky a dosažené skóre je v tabulce 8.4.

8.3 Albánština

Albánština je jazyk používaný především v Albánii a Kosovu, který nemá žádné příbuzné jazyky a tak tvoří vlastní jazykovou skupinu. Jedná se o flektivní jazyk s víceméně podobnou morfologií, jakou mají slovanské jazyky. To tedy opět vede k velkému množství slovních variant a volnému slovosledu. Díky tomu bylo opět potřeba přidat velké množství slov do slovníku a natrénovat jazykový model na velkém množství textů.

Foneticky se albánština velmi liší od slovanských jazyků a obsahuje spoustu fonémů, které se objevují napříč různými jazyky. Vztah mezi psanou a mluvenou podobou jazyka je poměrně přímý a tak bylo relativně snadné vytvořit produkční pravidla pro fonetickou transkripci.

Tabulka 8.3: Výsledky multilingválních testů pro zahájení vývoje albánštiny

Model	**WER** [%]
CZ+BG	54,15
CZ+RO	54,15
CZ+HR	55,55
RO+BG	64,80
HR+RO	47,45
CZ+HR+RO	52,10
CZ+BG+RO	50,50
HR+RO+BG	65,45
CZ+HR+RO+BG	49,40
CZ+HR+RO+BG+ES	48,10

Jelikož byla albánština vyvíjena jako jeden z posledních jazyků, bylo již v době vývoje k dispozici vícero systémů pro další jazyky a tak se nabízelo pro zahájení otestovat různé multilingvální akustické modely. Pro testování bylo využito sedmi různých jazyků, kde kromě slovanských byly použity i španělština, rumunština a maďarština. Jazyky byly nejdříve testovány samostatně a následně slučovány dohromady, kdy byla sloučena fonetická sada a akustický model byl natrénován na mixu trénovacích dat z použitých jazyků.

Pro testovaní byla vytvořena testovací sada nahráním dvou rodilých mluvčích v celkové délce 31 minut. Tabulka 8.3 zobrazuje výsledky všech multilingválních testů. Vyzkoušeno bylo mnoho různých kombinací a nejlepšího skóre dosáhl model vytvořený smícháním chorvatštiny s rumunštinou a model vytvořený smícháním pěti jazyků - češtiny, chorvatštiny, rumunštiny, bulharštiny a španělštiny. Pro zahájení vývoje byl nakonec vybrán druhý model, jelikož obsahoval větší fonetický inventář, který usnadňoval následné mapování na albánskou fonetickou sadu.

Následný vývoj probíhal standardně s využitím navržených metod. Data byla získávána ze zpravodajských webů a výsledné charakteristiky finálního systému jsou v tabulce 8.4.

Tabulka 8.4: Výsledné parametry a dosažené výsledky u dalších vybraných jazyků

Jazyk	Španělština	Lotyština	Albánština
Textový korpus	2,05 GB	1,73 GB	2,21 GB
Slovník	90 tis.	377 tis.	349 tis.
Výslovností	102 tis.	541 tis.	506 tis.
Fonémů	26	35	37
Zpracováno akust. dat	672 h.	3560 h.	422 h.
Získáno akust. dat	53 h.	118 h.	22 h.
Testovací sada	3 h.	2,5 h.	31 min.
WER	12,6 %	21,55 %	19,95 %

8.4 Zhodnocení aplikace metod na další jazyky

Výsledky dosažené pro uvedené 3 jazyky jsou shrnuty v tabulce 8.4. Je třeba říci, že dosažené hodnoty WER nelze přímo srovnávat s čísly uvedenými v předchozí kapitole, neboť použitá testovací data neměla stejný standardizovaný charakter jako u slovanských jazyků. Nicméně z nich vyplývá, že metody vyvinuté původně pro vzájemně příbuzné slovanské jazyky jsou plně použitelné i pro další evropské jazyky a s velkou pravděpodobností by mohly být aplikovány i pro některé neevropské. Vývoj u těchto tří jazyků probíhal v maximální míře automaticky a pouze v jediné osobě jsem byl schopen v řádu několika měsíců vytvořit systémy dosahující slušné úrovně WER pod 22 %. Navržené metody a postupy tak prokázaly svou efektivitu a určitou univerzálnost.

Závěr

V disertační práci jsem se zaměřil na řešení teoretických a praktických otázek spojených s efektivním vývojem multilingválních systémů automatického rozpoznávání řeči. Cílem bylo navrhnout, implementovat a na reálných datech ověřit postupy umožňující relativně rychle a s co nejmenšími náklady adaptovat existující systém pro nové jazyky.

Původní zadání počítalo se zaměřením na slovanské jazyky, u nichž bylo možné využít řadu společných lingvistických i fonetických rysů, nicméně se ukázalo, že navržený postup je dobře použitelný i pro jazyky z jiných jazykových skupin.

Základem celého přístupu je práce s textovými a akustickými řečovými daty, přičemž se ukazuje, že dobrých a v praxi použitelných výsledků lze dosáhnout s daty, která jsou veřejně přístupná na internetu, a s využitím vhodných metod strojového učení.

Navržený postup lze stručně popsat následujícím schématem:

1. Vytvoření dostatečně rozsáhlého a reprezentativního korpusu textů ze zdrojů přístupných na internetu a jeho následná úprava.
2. Vytvoření slovníku ze slov nejčastěji se vyskytujících v korpusu.
3. Vytvoření jazykového modelu pro daný slovník a korpus.
4. Definice fonetického inventáře pro daný jazyk.
5. Vygenerování výslovností pro všechna slova ve slovníku.
6. Shromáždění co největšího množství řečových nahrávek spolu s doprovodnými textovými daty, která více či méně odrážejí mluvený obsah nahrávek.
7. Iterativní proces výběru těch nahrávek, v nichž text odpovídá mluvenému obsahu. Shodu mluveného a textového obsahu určuje samotný vyvíjený rozpoznávací systém. Ten v počáteční fázi využívá akustický model jiného (již zvládnutého) jazyka a postupně ho adaptuje na nový jazyk na základě fonetických přepisů vytvořených systémem.

8. Při dostatečném množství trénovacích dat v cílovém jazyce pokračuje iterativní proces už s vlastním akustickým modelem, přičemž lze využít i nesupervizovaného přístupu, kdy se trénovací sada postupně rozšiřuje na základě vyhodnocování shody mezi různě nakonfigurovanými rozpoznávacími systémy.

9. Uvedený postup může běžet téměř automaticky. Je však vhodné doplnit ho o řízenou kontrolu těch nahrávek, ve kterých se referenční a rozpoznaný text liší v malém počtu (jednoho až dvou) slov, u nichž lze pomocí vhodně navrženého nástroje snadno (i pro laika) odhalit a opravit zdroj chyby.

10. Experimenty provedené na více než deseti různých jazycích ukazují, že popsaným postupem lze vytvořit funkční rozpoznávací systém pracující s reálnými daty s chybovostí nižší než 30 %, a který lze využít jak pro demonstrační účely (např. pro budoucího klienta), tak jako základ pro vývoj skutečné komerční aplikace.

Výše uvedené schéma se postupně vyvíjelo a optimalizovalo s každým dalším zpracovávaným jazykem. Pro často se opakující činnosti byly vyvíjeny nástroje, kterými jsem se snažil tyto činnosti co nejvíce zautomatizovat a převést je na programy či skripty s volitelnými parametry specifickými pro každý jazyk. Díky tomuto postupu a také díky modelům pro již zpracované jazyky je nyní možné uskutečnit základní vývoj systému pro další jazyk v průběhu cca 3-6 měsíců a dosáhnout přesnosti na úrovni výsledků zmíněných v kapitolách 7 a 8.

Shrnutí přínosů práce k rozvoji vědního oboru

Vědecké přínosy práce jsou shrnuty v následujících bodech:

- Byla navržena a prakticky prověřena metodologie procesu vývoje jazykově závislých modelů systému rozpoznávání řeči za využití co největší míry automatizace procesu především v časově náročných a opakujících se úlohách.

- V rámci metodologie byly vytvořeny jednotné zásady pro zpracování a tvorbu dat týkající se především způsobu značení, kódování či formátování dat.

- Pro jednotlivé dílčí kroky vývoje byly vytvořeny efektivní nástroje dodržující stanovené zásady, které mohou být využity jak pro tvorbu systémů pro další jazyky, tak i pro jiné úlohy v oblasti zpracování řečového signálu.

- Při práci s jazyky používající odlišné abecedy byl navržen způsob efektivního převodu mezi abecedami pro usnadnění práce lidí neznalých dané abecedy.

- V mnoha krocích vývoje bylo využito různých metod strojového učení, především tzv. lehce supervizovaného, které byly adaptovány pro konkrétní účely.

- Navržené metody byly aplikovány na všechny národní slovanské jazyky a tedy i ty, pro které zatím podle dostupné literatury žádné systémy pro rozpoznávání spojité řeči neexistovaly, jako běloruština, bosenština, černohorština či makedonština.

- Dále byly metody úspěšně aplikovány i na jiné neslovanské jazyky, a to i jazyky s nedostatečnými zdroji jako lotyština či albánština, kde podle dostupných zdrojů existuje pouze jeden systém rozpoznávání řeči v případě lotyštiny a pravděpodobně žádný v případě albánštiny.

- Zároveň byly ve spolupráci s rodilými mluvčími vytvořeny unifikované testovací sady z televizních a rozhlasových zpravodajských pořadů pro 11 slovanských jazyků (tedy téměř všech národních jazyků kromě bosenštiny a černohorštiny), na kterých byly testovány finální systémy. Data byla veřejně zpřístupněna a využita nejen námi, ale i zahraničními týmy, například v [82].

Shrnutí přínosů práce pro praxi

Vytvořené systémy byly postupně nasazovány partnerskou firmou Newton technologies, a.s., do praxe v rámci společných projektů "MULTILINMEDIA" a "DeepSpot" pro včasné upozorňování a monitorování médií nebo jako další komerční aplikace například pro automatické titulkování pořadů programem Beey či pro diktovací software Newton Dictate použitelný jak pro obecné diktování, tak i například pro soudy či speciální lékařská oddělení.

Tyto produkty jsou již komerčně nasazeny kromě České republiky také na Slovensku, v Polsku, ve Slovinsku, v Chorvatsku a v Srbsku a jsou podle potřeb neustále aktualizovány novými daty, která jsou zpracovávána popsanou metodologií.

Navazující a budoucí práce

Nástroje, data a metodologie vytvořené během této práce už průběžně jsou nebo budou využity v dalších projektech a výzkumech řešených na pracovišti.

Celý postup již byl úspěšně aplikován i dalšími členy týmu pro tvorbu nových systémů pro další evropské jazyky. V plánu je aplikace i pro další jazyky jako například švédština a norština v rámci nově zahajovaného mezinárodního projektu.

Data a znalosti získané v rámci této práce jsou využívány rovněž v rámci výzkumu a vývoje systémů pro identifikaci jazyka, především pro slovanské jazyky [83].

V neposlední řadě všechna vytvořená řečová data slouží také při vývoji společného multilingválního mnohavrstevného akustického modelu typu DNN, kterým se zabývají další členové týmu.

Literatura

[1] Huang, Xuedong, et al., „Spoken language processing: A guide to theory, algorithm, and system development“, Prentice hall PTR, 2001.

[2] Juang, Biing-Hwang, and Lawrence R. Rabiner. "Automatic speech recognition–a brief history of the technology development." Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara 1 (2005): 67.

[3] O'Shaughnessy, Douglas. "Automatic speech recognition: History, methods and challenges." Pattern Recognition 41.10 (2008): 2965-2979.

[4] "Special Eurobarometer 386: Europeans and their Languages", Eurobarometer Special Surveys, 2014-12-08

[5] Čermák, František. Jazyk a jazykověda. Karolinum Press, 2011.

[6] Sgall, Petr. Jazyk, mluvení, psaní. Karolinum Press, 2011.

[7] Pala, Karel, and Klára Osolsobě. Základy počítačové lingvistiky. Masarykova univerzita, 1992.

[8] Ashby, Michael, and Maidment, John, „Úvod do obecné fonetiky“, Charles University in Prague, Karolinum Press, 2015.

[9] Volín, J., „Fonetika a fonologie“, MSoČ 1, 2010, 43–45.

[10] Nouza, J., Koldovský, Z., Vích, R., „Řeč a počítač: principy hlasové komunikace, úlohy, metody a aplikace: sborník článků“, Liberec: Technická univerzita v Liberci, 2009. ISBN 978-80-7372-548-8.

[11] Alpaydin, Ethem. Introduction to machine learning. MIT press, 2020.

[12] Derouault, A. M., et al. "The IBM speech server series and its applications in Europe." Applications of Speech Technology. 1993.

[13] Lai, Jennifer, and John Vergo. "MedSpeak: Report creation with continuous speech recognition." Proceedings of the ACM SIGCHI Conference on Human factors in computing systems. 1997.

[14] Newman, Dan. Dragon NaturallySpeaking Guide: Speech Recognition Made Fast and Simple. Waveside Publishing, 1999.

[15] Gauvain, J-L., and Lori Lamel. "Large-vocabulary continuous speech recognition: advances and applications." Proceedings of the IEEE 88.8 (2000): 1181-1200.

[16] Adda-Decker, Martine, Lori Lamel, and Natalie D. Snoeren. "Comparing monomultilingual acoustic seed models for a low e-resourced language: a case-study of Luxembourgish." Eleventh Annual Conference of the International Speech Communication Association. 2010.

[17] Do, Cong-Thanh, Lori Lamel, and Jean-Luc Gauvain. "Speech-to-Text Development for Slovak, a Low-Resourced Language." Spoken Language Technologies for Under-Resourced Languages. 2014.

[18] Schultz, Tanja. Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen. Shaker, 2000.

[19] Schultz, Tanja, Martin Westphal, and Alex Waibel. "The globalphone project: Multilingual lvcsr with janus-3." Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop. 1997.

[20] Young, Steve J., and Sj Young. "The HTK hidden Markov model toolkit: Design and philosophy." (1993): 69.

[21] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

[22] Stolcke, Andreas. "SRILM-an extensible language modeling toolkit." Seventh international conference on spoken language processing. 2002.

[23] Kucera, Karel. "The Czech National Corpus: principles, design, and results." Literary and linguistic computing 17.2 (2002): 245-257.

[24] Koehn, Philipp. "Europarl: A parallel corpus for statistical machine translation." MT summit. Vol. 5. 2005.

[25] von Waldenfels, Ruprecht. "ParaSol: introduction to a Slavic parallel corpus." Prace Filologiczne 63 (2012): 293-302.

[26] Tiedemann, Jörg, and Lars Nygaard. "The OPUS Corpus-Parallel and Free: http://logos. uio. no/opus." LREC. 2004.

[27] Brants, T., Franz, A.: Web 1T 5-gram, version 1. Linguistic Data Consortium, Philadelphia (2006)

[28] Prochazka, Vaclav, et al. "Performance of Czech speech recognition with language models created from public resources." Radioengineering (2011).

[29] Švec, Jan, et al. "Web text data mining for building large scale language modelling corpus." International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, 2011.

[30] Schultz, Tanja. "SPICE-An Interactive Toolkit for Rapid Portability of Speech Processing Systems to new Languages." Multilingual Speech and Language Processing. 2006.

[31] Rotovnik, T., Maucec, M. S., Kacix, Z., „Large vocabulary continuous speech recognition of an inflected language using stems and endings", Speech Communication. 49(6), 2007. pp. 437–452.

[32] Oparin, I., Glembek, O., Burget, L., Černocký J., „Morphological random forests for language modeling of inflectional languages", In Proc. IEEE Workshop on Spoken Language Technology SLT'08, Goa, India, 2008.

[33] Sak, H., Saraclar, M., Güngör, T., „Morphology-based and sub-word language modeling for Turkish speech recognition". ICASSP 2010, pp. 5402-5405.

[34] Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A., „Large vocabulary Russian speech recognition using syntactico-statistical language modeling", Speech Communication Journal, Special Issue on Processing Under-Resourced Languages, 2013.

[35] Chelba, C., Jelinek, F., „Structured language model", Computer Speech and Language. Vol. 10, 2000, pp. 283–332.

[36] Huet, S., Gravier, G., Sebillot, P., „Morpho-syntactic postprocessing of N-best lists for improved French automatic speech recognition", Computer Speech and Language, 24(4), 2010.

[37] Garofolo, John S., et al. "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1." STIN 93 (1993): 27403.

[38] Sharada, C. S., Vijaya, C., "Speech Recognition Using Monophone and Triphone Based Continuous Density Hidden Markov Models", International journal of research and scientific innovation 2015, pp. 30-35.

[39] Stüker, S., Tanja, S., „A grapheme based speech recognition system for Russian", 9th Conference Speech and Computer. 2004.

[40] Le, V., Besacier, L., „Automatic Speech Recognition for Under-Resourced Languages: Application to Vietnamese Language" IEEE Transactions on Audio, Speech and Language Processing. Volume 17, Issue 8, Nov. 2009, pp. 1471 – 1482.

[41] Killer, M., Stüker, S., Schultz, T., „Grapheme Based Speech Recognition", Interspeech 2003, Geneva, Switzerland, 2003.

[42] Schlippe, T., Ochs, S., Schultz, T., „Wiktionary as a Source for Automatic Pronunciation Extraction", Interspeech 2010, Makuhari, Japan, 2010.

[43] Braga, Daniela, Luís Coelho, and Fernando Gil Vianna Resende. "A rule-based grapheme-to-phone converter for TTS systems in European Portuguese." 2006 International Telecommunications Symposium. IEEE, 2006.

[44] Altinok, Duygu. "Towards Turkish ASR: Anatomy of a rule-based Turkish g2p." arXiv preprint arXiv:1601.03783 (2016).

[45] Laurent, A., Deleglise, P., Meignier, S., „Grapheme to phoneme conversion using an SMT system", Interspeech 2009, pp. 708-711, Brighton, UK, 2009.

[46] Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[47] Juzová, Markéta, Daniel Tihelka, and Jakub Vít. "Unified Language-Independent DNN-Based G2P Converter." INTERSPEECH. 2019.

[48] Schultz, T., „Globalphone: a multilingual speech and text database developed at Karlsruhe university", Interspeech 2002, Denver, USA, 2002.

[49] Parent, G., Eskenazi, M., „Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data", In Proceedings of IEEE Workshop on Spoken Language Technology, pp- 312– 317, Berkeley, California, 2010.

[50] Hughes, T., Nakajima, K., Ha, L., Moreno, P., LeBeau, M., „Building transcribed speech corpora quickly and cheaply for many languages", in Proc. Interspeech 2010, Makuhari, Japan, 2010, pp. 1914-1917.

[51] De Vries, N. J., Badenhorst, J., Davel, M.H., Barnard, E., De Waal, A., „Woefzela-an opensource platform for ASR data collection in the developing world", in Proc. Interspeech, 2011, pp. 3177-3180.

[52] Braunschweiler, N., Gales, M., Buchholz, S., „Lightly supervisedrecognition for automatic alignment of large coherent speech recordings", in Proc. Interspeech, Makuhari, Japan, Sept. 2010, pp. 2222–2225.

[53] Meng, M., Wang, S., Liang, J., Ding, P., Xu, B., „Full utilization of closed-captions in broadcast news recognition", in Proc. IS-CSLP, Kent Ridge, Singapore, 2006.

[54] Davel, M. H., van Heerden, C., Kleynhans, N., Barnard, E., „Efficient harvesting of Internet audio for resource-scarce ASR", in Proc. Interspeech, 2011, pp. 3153-3156.

[55] Cetin, O., „Unsupervised adaptive speech technology for limited resource languages: a case studyfor Tamil", SLTU'08, Hanoi, Vietnam, 2008.

[56] Loof, J., Gollan, C., Ney, H., „Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System“, Interspeech 2009. Brighton, UK. 2009.

[57] Vu, N. T., Kraus, F., Schultz, T., „Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training“, in Proc. Interspeech, 2011.

[58] Schultz, Tanja, and Alex Waibel. "Multilingual and crosslingual speech recognition." Proc. DARPA Workshop on Broadcast News Transcription and Understanding. 1998.

[59] Schultz, Tanja, and Alex Waibel. "Development of Multi-lingual Acoustic Models in the GlobalPhone Project." Proceedings of the 1st Workshop on Text, Speech, and Dialogue (TSD). 1998.

[60] Kumar, C. Santhosh, V. P. Mohandas, and Haizhou Li. "Multilingual speech recognition: A unified approach." Ninth European Conference on Speech Communication and Technology. 2005.

[61] Burget, Lukáš, et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models." 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.

[62] Uebler, Ulla. "Multilingual speech recognition in seven languages." Speech communication 35.1-2 (2001): 53-69.

[63] Ma1, Bin, et al. "Multilingual speech recognition with language identification." Seventh International Conference on Spoken Language Processing. 2002.

[64] Tong, Sibo, Philip N. Garner, and Hervé Bourlard. "An investigation of deep neural networks for multilingual speech recognition training and adaptation." Proc. of INTERSPEECH. No. CONF. 2017.

[65] Watanabe, Shinji, Takaaki Hori, and John R. Hershey. "Language independent end-to-end architecture for joint language identification and speech recognition." 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2017.

[66] Hannun, Awni, et al. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).

[67] Sussex, Roland, and Paul Cubberley, „The Slavic Languages. Cambridge University Press“, 2006.

[68] Bethin, Christina Y., and Christina y Bethin. Slavic prosody: Language change and phonological theory. Cambridge University Press, 1998.

[69] Stankiewicz, Edward. The Slavic languages: unity in diversity. Walter de Gruyter, 1986.

[70] Siewierska, Anna, and Ludmila Uhlirova, „An overview of word order in Slavic languages“, Constituent order in the languages of Europe 20.1 (1998): 105.

[71] Franks, Steven, „Parameters of Slavic morphosyntax“, Oxford University Press, 1995.

[72] Valta, Jan. ”Identifikace jazyka textového dokumentu”. Diplomová práce. Technická Univerzita v Liberci, 2012.

[73] Grave, Edouard, et al. ”Learning word vectors for 157 languages.” arXiv preprint arXiv:1802.06893 (2018).

[74] Kolorenč, Jan. Tvorba a adaptace lingvistické vrstvy pro systém rozpoznávání mluvené češtiny. Diss. Technická Univerzita v Liberci, 2007.

[75] Witten, Ian H., Bell, Timothy C., ”The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression”, IEEE Transactions on Information Theory, 37(4):1085–1094, 1991.

[76] Nouza, J., Psutka, J., Uhlíř, J., „Phonetic Alphabet for Speech Recognition of Czech“, In Radio Engineering, vol. 6, no. 4, 1997, pp. 16-20.

[77] Hinrichs, E. and Krauwer, S., “The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars.” LREC-2014, May 2014, pp. 1525–31.

[78] Nouza, J., Červa, P., Kuchařová, M., „Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages“. Radioengineering, 22(3), 2013.

[79] Nouza, J., Cerva, P. and Silovsky, J., „Dealing with Bilingualism in Automatic Transcription of Historical Archive of Czech Radio“, International Conference on Image Analysis and Processing. Springer Berlin Heidelberg, 2013, pp. 238-246.

[80] Nouza, J., „A Czech Large Vocabulary Recognition System for Real-Time Applications“, In Text, Speech and Dialogue (eds. Sojka, Kopecek, Pala) Springer-Verlag, Heidelberg, 2000, pp. 217-222.

[81] Moreno Bilbao, M. Asunción, et al. ”Albayzin speech database: Design of the phonetic corpus.” EUROSPEECH 1993: 3rd European Conference on Speech Communication and Technology: Berlin, Germany: September 22-25, 1993. . EUROSPEECH, 1993.

[82] Abdullah, Badr, et al. ”Cross-Domain Adaptation of Spoken Language Identification for Related Languages: The Curious Case of Slavic Languages.” arXiv preprint arXiv:2008.00545 (2020).

[83] Matějů, L., Červa, P., Žďánský, J. a Šafařík, R. Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 1. vyd. Indie: ISCA, 2018 S. 1803 – 1807. ISSN: 2308-457X.

Autorovy publikace

1. Šafařík, R., Nouza, J., „Methods for rapid development of Automatic Speech Recognition“, ECMSM 2015, Liberec, 2015.

2. Nouza, J., Červa, P., Šafařík, R., „Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources“, In LTC'15, Poznań, Poland, November 2015.

3. Šafařík, R., Matějů, L., „Impact of Phonetic Annotation Precision on Automatic Speech Recognition Systems“, In Proc. TSP 2016, Vienna, Austria, pp. 311-314, June 2016.

4. Nouza, J., Šafařík, R., Červa, P., „ASR for South Slavic Languages Developed in Almost Automated Way“, In Proc. INTERSPEECH 2016, San Francisco, USA, September 2016.

5. Boháč, M., Matějů, L., Rott, M., Šafařík, R., „Automatic Syllabification and Syllable Timing of Automatically Recognized Speech - for Czech“, In Proc. TSD 2016, Brno, Czech Republic, pp. 540-547, September 2016.

6. Šafařík, R., Matějů, L., „The Impact of Inaccurate Phonetic Annotations on Speech Recognition Performance“, In Proc. TSD 2017, Prague, Czech republic, pp. 402-411, August 2017.

7. Šafařík, R., Nouza, J., „Unified Approach to Development of ASR Systems for East Slavic Languages“, In Proc. SLSP 2017, Le Mans, France, pp. 193-203, October 2017.

8. Nouza, J., Šafařík, R., „Parliament archives used for automatic training of multi-lingual automatic speech recognition systems“, In Proc. TSD 2017, Prague, Czech republic, pp. 174-183, August 2017.

9. Nouza, J., Červa, P., Šafařík, R., „Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources“, In Proc. Lecture Notes in Artificial Intelligence (LNAI), 2017.

10. Šafařík, R., Matějů, L.: „Automatic Development of ASR System for an Under-Resourced Language“, In proc. of 41st International Conference on Telecommunications and Signal Processing, TSP 2018, Athens, Greece, pp. 1-4, July 2018.

11. Šafařík, R., Matějů, L., Weingartová, L.:„The Influence of Errors in Phonetic Annotations on Performance of Speech Recognition System“, In proc. of 21st International Conference on Text, Speech and Dialogue, TSD 2018, Brno, Czech republic, pp. 419-427, September 2018.

12. Mateju, L., Cerva, P., Zdansky J., and Safarik R.: „Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal“, Interspeech 2018, Hyderabad, India, pp. 1803-1807, September 2018.

Příloha A – Ortografické tabulky

Znaky západoslovanských jazyků

Jazyk	Znaky
Čeština - velké	A, Á, B, C, Č, D, Ď, E, É, Ě, F, G, H, CH, I, Í, J, K, L, M, N, Ň, O, Ó, P, Q, R, Ř, S, Š, T, Ť, U, Ú, Ů, V, W, X, Y, Ý, Z, Ž
Čeština - malé	a, á, b, c, č, d, ď, e, é, ě, f, g, h, i, í, j, k, l, m, n, ň, o, ó, p, q, r, ř, s, š, t, ť, u, ú, ů, v, w, x, y, ý, z, ž
Slovenština - velké	A, Á, Ä, B, C, Č, D, Ď, DZ, DŽ, E, É, F, G, H, CH, I, Í, J, K, L, Ĺ, Ľ, M, N, Ň, O, Ó, Ô, P, Q, R, Ŕ, S, Š, T, Ť, U, Ú, V, W, X, Y, Ý, Z, Ž
Slovenština - malé	a, á, ä, b, c, č, d, ď, dz, dž, e, é, f, g, h, ch, i, í, j, k, l, Í, ľ, m, n, ň, o, ó, ô, p, q, r, ŕ, s, š, t, ť, u, ú, v, w, x, y, ý, z, ž
Polština - velké	A, Ą, B, C, Ć, D, E, Ę, F, G, H, I, J, K, L, Ł, M, N, Ń, O, Ó, P, R, S, Ś, T, U, W, Y, Z, Ź, Ż
Polština - malé	a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, r, s, ś, t, u, w, y, z, ź, ż

Znaky jihoslovanských jazyků používajících latinku

Jazyk	Znaky
Slovinština - velké	A, B, C, Č, D, E, F, G, H, I, J, K, L, M, N, O, P, R, S, Š, T, U, V, Z, Ž
Slovinština - malé	a, b, c, č, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, š, t, u, v, z, ž
Chorvatština, srbština, bosenština - velké	A, B, C, Č, Ć, D, DŽ, Đ, E, F, G, H, I, J, K, L, LJ, M, N, NJ, O, P, R, S, Š, T, U, V, Z, Ž
Chorvatština, srbština, bosenština - malé	a, b, c, č, ć, d, dž, đ, e, f, g, h, i, j, k, l, lj, m, n, nj, o, p, r, s, š, t, u, v, z, ž
Černohorština - velké	A, B, C, Č, Ć, D, DŽ, Đ, E, F, G, H, I, J, K, L, LJ, M, N, NJ, O, P, R, S, Š, Ś, T, U, V, Z, Ž, Ź
Černohorština - malé	a, b, c, č, ć, d, dž, đ, e, f, g, h, i, j, k, l, lj, m, n, nj, o, p, r, s, š, ś, t, u, v, z, ž, ź

Znaky jihoslovanských jazyků používajících cyrilici

Jazyk	Znaky
Srbština, bosenština - velké	А, Б, В, Г, Д, Ђ, Е, Ж, З, И, Ј, К, Л, Љ, М, Н, Њ, О, П, Р, С, Т, Ћ, У, Ф, Х, Ц, Ч, Џ, Ш
Srbština, bosenština - malé	а, б, в, г, д, ђ, е, ж, з, и, ј, к, л, љ, м, н, њ, о, п, р, с, т, ћ, у, ф, х, ц, ч, џ, ш
Černohorština - velké	А, Б, В, Г, Д, Ђ, Е, Ж, З, З́, И, Ј, К, Л, Љ, М, Н, Њ, О, П, Р, С, С́, Т, Ћ, У, Ф, Х, Ц, Ч, Џ, Ш
Černohorština - malé	а, б, в, г, д, ђ, е, ж, з, з́, и, ј, к, л, љ, м, н, њ, о, п, р, с, с́, т, ћ, у, ф, х, ц, ч, џ, ш
Makedonština - velké	А, Б, В, Г, Ѓ, Д, Е, Ж, S, З, И, Ј, К, Ќ, Л, Љ, М, Н, Њ, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Џ, Ш
Makedonština - malé	а, б, в, г, ѓ, д, е, ж, s, з, и, ј, к, ќ, л, љ, м, н, њ, о, п, р, с, т, у, ф, х, ц, ч, џ, ш
Bulharština - velké	А, Б, В, Г, Д, Е, Ж, З, И, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ь, Ю, Я
Bulharština - malé	а, б, в, г, д, е, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ь, ю, я

Znaky východoslovanských jazyků

Jazyk	Znaky
Ruština - velké	А, Б, В, Г, Д, Е, Ё, Ж, З, И, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ы, Ь, Э, Ю, Я
Ruština - malé	а, б, в, г, д, е, ё, ж, з, и, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ъ, ы, ь, э, ю, я
Ukrajinština - velké	А, Б, В, Г, Ґ, Д, Е, Є, Ж, З, И, І, Ї, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ь, Ю, Я
Ukrajinština - malé	а, б, в, г, ґ, д, е, є, ж, з, и, і, ї, й, к, л, м, н, о, п, р, с, т, у, ф, х, ц, ч, ш, щ, ь, ю, я
Běloruština - velké	А, Б, В, Г, Д, ДЖ, ДЗ, Е, Ё, Ж, З, І, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ў, Ф, Х, Ц, Ч, Ш, Ы, Ь, Э, Ю, Я
Běloruština - malé	а, б, в, г, д, дж, дз, е, ё, ж, з, і, й, к, л, м, н, о, п, р, с, т, у, ў, ф, х, ц, ч, ш, ы, ь, э, ю, я

Znaky dalších zpracovaných jazyků

Jazyk	Znaky
Španělština - velké	A, B, C, D, E, F, G, H, I, J, K, L, M, N, Ñ, O, P, Q, R, S, T, U, V, W, X, Y, Z
Španělština - malé	a, b, c, d, e, f, g, h, i, j, k, l, m, n, ñ, o, p, q, r, s, t, u, v, w, x, y, z
Lotyština - velké	A, Ā, B, C, Č, D, E, Ē, F, G, Ģ, H, I, Ī, J, K, Ķ, L, Ļ, M, N, Ņ, O, P, R, S, Š, T, U, Ū, V, Z, Ž
Lotyština - malé	a, ā, b, c, č, d, e, ē, f, g, ģ, h, i, ī, j, k, ķ, l, ļ, m, n, ņ, o, p, r, s, š, t, u, ū, v, z, ž
Albánština - velké	A, B, C, Ç, D, Dh, E, Ë, F, G, Gj, H, I, J, K, L, Ll, M, N, Nj, O, P, Q, R, Rr, S, Sh, T, Th, U, V, X, Xh, Y, Z, Zh
Albánština - malé	a, b, c, ç, d, dh, e, ë, f, g, gj, h, i, j, k, l, ll, m, n, nj, o, p, q, r, rr, s, sh, t, th, u, v, x, xh, y, z, zh

Příloha B – Fonetické tabulky

Čeština

Samohlásky	Přední	Střední	Zadní
Zavřené	ɪ iː		u uː
Polozavřené	ɛ ɛː		o oː
Otevřené		a aː	

Souhlásky	Lab.	Labdent.	Alv.	Postalv.	Pal.	Vel.	Glot.
Nazály	m		n		ɲ	ŋ	
Plozivy	p b		t d		c ɟ	k g	
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ			
Frikativy		f v	s z	ʃ ʒ		x	ɦ
Vibranty			r				
Frik. vibranty			ɾ̝				
Aproximanty			l		j		

Slovenština

Samohlásky	Přední	Střední	Zadní
Zavřené	i iː		u uː
Polozavřené	ɛ ɛː		ɔ ɔː
Otevřené	æ		a aː

Souhlásky	Lab.	Labdent.	Alv.	Postalv.	Pal.	Vel.	Glot.
Nazály	m		n		ɲ	ŋ	
Plozivy	p b		t d		c ɟ	k g	
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ			
Frikativy		f v	s z	ʃ ʒ		x	ɦ
Vibranty			rː				
Verberanty			ɾ				
Aproximanty			l lː		j ʎ		

Polština

Samohlásky	Přední	Střední	Zadní
Zavřené	i	ɨ	u
Polozavřené	ɛ ɛ̃		ɔ ɔ̃
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.	Glot.
Nazály	m(ʲ)		n		ɲ	ŋ	
Plozivy	p(ʲ) b(ʲ)		t(ʲ) d(ʲ)			k(ʲ) g(ʲ)	
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ	t͡ɕ d͡ʑ		
Frikativy		f(ʲ) v(ʲ)	s z	ʃ(ʲ) ʒ(ʲ)	ɕ ʑ	x(ʲ) (ɣ)	(ɦ)
Vibranty			r(ʲ)				
Aproximanty	w		l(ʲ)		j		

Slovinština

Samohlásky	Přední	Střední	Zadní
Zavřené	i		u
Polozavřené	e ɛ	ə	o ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m		n			
Plozivy	p b		t d			k g
Afrikáty			t͡s	t͡ʃ d͡ʒ		
Frikativy		f	s z	ʃ ʒ		x
Verberanty			ɾ			
Aproximanty	ʋ		l		j	

Chorvatština, bosenština, srbština

Samohlásky	Přední	Střední	Zadní
Zavřené	i iː		u uː
Polozavřené	e eː		o oː
Otevřené		a aː	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m		n		ɲ	
Plozivy	p b		t d			k g
Afrikáty			t͡s	t͡ʃ d͡ʒ	t͡ɕ d͡ʑ	
Frikativy		f v	s z	ʃ ʒ		x
Vibranty			r			
Aproximanty			l		j ʎ	

Černohorština

Samohlásky	Přední	Střední	Zadní
Zavřené	i iː		u uː
Polozavřené	e eː		o oː
Otevřené		a aː	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m		n		ɲ	
Plozivy	p b		t d			k g
Afrikáty			t͡s	t͡ʃ d͡ʒ	t͡ɕ d͡ʑ	
Frikativy		f v	s z	ʃ ʒ	ɕ ʑ	x
Vibranty			r			
Aproximanty			l		j ʎ	

Makedonština

Samohlásky	Přední	Střední	Zadní
Zavřené	i		u
Polozavřené	ɛ	(ə)	ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m	n			ɲ	
Plozivy	p b	t d			c ɟ	k g
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ		
Frikativy		f v	s z	ʃ ʒ		x
Vibranty			r			
Aproximanty			l		j ʎ	

Bulharština

Samohlásky	Přední	Střední	Zadní
Zavřené	i		u
Polozavřené	ɛ	ɤ	ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m(ʲ)		n		ɲ	ŋ
Plozivy	p(ʲ) b(ʲ)		t(ʲ) d(ʲ)			k(ʲ) g(ʲ)
Afrikáty			t͡s(ʲ) d͡z(ʲ)	t͡ʃ d͡ʒ		
Frikativy		f(ʲ) v(ʲ)	s(ʲ) z(ʲ)	ʃ ʒ		x(ʲ) (ɣ)
Vibranty			r(ʲ)			
Aproximanty			ɫ		j ʎ	

Ruština

Samohlásky	Přední	Střední	Zadní
Zavřené	i	ɨ	u
Polozavřené	e		o
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.
Nazály	m(ʲ)		n		ɲ	
Plozivy	p(ʲ) b(ʲ)		t(ʲ) d(ʲ)			k(ʲ) g(ʲ)
Afrikáty			t͡s(ʲ)		t͡ɕ	
Frikativy		f(ʲ) v(ʲ)	s(ʲ) z(ʲ)	ʂ ʐ	ɕː (ʑː)	x(ʲ) (ɣ)
Vibranty			rʲ	r		
Aproximanty			ɫ lʲ		j	

Ukrajinština

Samohlásky	Přední	Střední	Zadní
Zavřené	i	ɪ	u
Polozavřené	ɛ		ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Postalv	Pal.	Vel.	Glotal
Nazály	m		n		ɲ		
Plozivy	p b		t(ʲ) d(ʲ)			k g	
Afrikáty			t͡s(ʲ) d͡z(ʲ)	t͡ʃ d͡ʒ			
Frikativy		f(ʲ)	s(ʲ) z(ʲ)	ʃ ʒ		x	ɦ
Vibranty			r(ʲ)				
Aproximanty	w		l(ʲ)		j ʎ		

Běloruština

Samohlásky	Přední	Střední	Zadní
Zavřené	i	ɨ	u
Polozavřené	ɛ		ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Alv.	Retrof.	Pal.	Vel.
Nazály	m(ʲ)		n		ɲ	
Plozivy	p(ʲ) b(ʲ)		t(ʲ) d(ʲ)			k(ʲ) g(ʲ)
Afrikáty			t͡s(ʲ) d͡z(ʲ)	t͡ʂ d͡ʐ		
Frikativy		f(ʲ) v(ʲ)	s(ʲ) z(ʲ)	ʂ ʐ		x(ʲ) (ɣ)
Vibranty			r			
Aproximanty	w		ɫ		j ʎ	

Španělština

Samohlásky	Přední	Střední	Zadní
Zavřené	i		u
Polozavřené	e		o
Otevřené		a	

Souhlásky	Lab.	Labdent.	Dent.	Alv.	Pal.	Vel.
Nazály	m			n	ɲ	
Plozivy	p b			t d		k g
Afrikáty					tʃ	
Frikativy		f (v)	θ	s	ʃ	x
Vibranty				r		
Aproximanty				ɾ l	j ʎ	

Lotyština

Samohlásky	Přední	Střední	Zadní
Zavřené	i iː		u eː
Polozavřené	e eː		(ɔ ɔː)
Otevřené	æ æː	a aː	

Souhlásky	Lab.	Labdent.	Alv.	Postalv.	Pal.	Vel.
Nazály	m		n		ɲ	ŋ
Plozivy	p b		t d		c ɟ	k g
Afrikáty			t͡s d͡z	t͡ʃ d͡ʒ		
Frikativy		f v	s z	ʃ ʒ		x
Vibranty			r		(rʲ)	
Aproximanty			l		j ʎ	

Albánština

Samohlásky	Přední	Střední	Zadní
Zavřené	i y		u
Polozavřené	ɛ		ɔ
Otevřené		a	

Souhlásky	Lab.	Labdent.	Dent.	Alv.	Postalv.	Pal.	Vel.	Glot.
Nazály	m			n		ɲ	ŋ	
Plozivy	p b			t d			k g	
Afrikáty				t͡s d͡z	t͡ʃ d͡ʒ	c͡ç ɟ͡ʝ		
Frikativy		f v	θ ð	s z	ʃ ʒ			ɦ
Vibranty				r				
Aproximanty				ɾ l		j	ɫ	

Příloha C – Mapování znaků cyrilice na latinku

CZ	**A**	**B**	**V**	**G**	**D**	**Ď**	**Ě**	**Ô**	**Ž**	**Ż**	**Z**	**I**	**J**	**Î**	**K**	**Ť**	**L**	**Ľ**	**M**	**N**	**Ň**
RU	А	Б	В	Г	Д	-	Е	Ё	Ж	-	З	И	Й	-	К	-	Л	-	М	Н	-
UK	А	Б	В	Ґ	Д	-	Є	-	Ж	-	З	І	Й	Ї	К	-	Л	-	М	Н	-
BE	А	Б	В	-	Д	-	Е	Ё	Ж	-	З	І	Й	-	К	-	Л	-	М	Н	-
BG	А	Б	В	Г	Д	-	-	-	Ж	-	З	И	Й	-	К	-	Л	-	М	Н	-
MK	А	Б	В	Г	Д	Ѓ	-	-	Ж	S	З	И	Ј	-	К	Ќ	Л	Љ	М	Н	Њ

CZ	**O**	**P**	**R**	**S**	**T**	**Ů**	**U**	**F**	**H**	**X**	**C**	**Č**	**Š**	**Ş**	**~**	**Y**	**^**	**E**	**Ű**	**Â**	**Ă**
RU	О	П	Р	С	Т	-	У	Ф	Х	-	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я	-
UK	О	П	Р	С	Т	-	У	Ф	Г	Х	Ц	Ч	Ш	Щ	'	И	Ь	Е	Ю	Я	-
BE	О	П	Р	С	Т	Ў	У	Ф	Г	Х	Ц	Ч	Ш	-	-	И	Ь	Э	Ю	Я	-
BG	О	П	Р	С	Т	-	У	Ф	Х	-	Ц	Ч	Ш	Щ	-	-	Ь	Е	Ю	Я	Ъ
MK	О	П	Р	С	Т	-	У	Ф	Х	-	Ц	Ч	Ш	-	-	-	-	Е	-	-	-