Tvorba a adaptace lingvistické vrstvy pro systém rozpoznávání mluvené češtiny

autoreferát disertační práce

Jan Kolorenč

Tvorba a adaptace lingvistické vrstvy pro systém rozpoznávání mluvené češtiny

Disertační práce

Disertant:	Jan Kolorenč
Studijní program:	2612V Elektronika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Laboratoř počítačového zpracování řeči, Ústav informačních technologií, Fakulta mechatroniky a mezioborových inženýrských studií, Technická univerzita v Liberci
Školitel:	Prof. Ing. Jan Nouza, CSc.

Rozsah práce:

Počet stran:	101
Počet obrázků:	20
Počet tabulek:	31
Počet příloh:	2

Anotace

Tvorba lingvistické vrstvy pro systém rozpoznávání mluvené češtiny je v tomto díle chápána jako komplexní úloha skládající se z logicky navazujících kroků. Jednotlivé kroky využívají různorodé přístupy od využití hrubé výpočetní síly přes metody umělé inteligence, využití rad expertů až po různé heuristiky. Často dochází k fúzi těchto přístupů.

Nejprve jsou diskutovány otázky různých zdrojů textových dat a problémy při jejich využití. Jsou též uvedeny metody čištění textového korpusu a jejich vliv na úspěšnost rozpoznávaní.

V části o slovníku a fonetickém přepisu je diskutován vliv velikosti slovníku. Dále je uvedena metoda pro semiautomatické nalezení nových fonologických pravidel vylepšující automatickou fonetickou transkripci. Přidáním slovních párů do slovníku lze téměř bezpracně zlepšit úspěšnost rozpoznávaní. Tato metoda je uvedena na závěr části týkající se slovníku.

Velký slovník způsobuje problém při implementaci počítání jazykového modelu. Tento problém je vyřešen pro různé konfigurace počítačů v závislosti na preferenci malé spotřeby paměti nebo rychlosti výpočtu. Dosavadní programy pro výpočet jazykového modelu jsou výrazně zrychleny, čímž mohlo být uskutečněno mnoho experimentů.

Pro efektivní zvyšování úspěšnosti rozpoznávání je nutné co nepřesněji identifikovat a kvantifikovat chyby. Je proto zlepšena metoda vyhodnocování výsledků rozpoznávání.

Adaptace jazykového modelu je v literatuře velmi diskutovanou částí automatického rozpoznávání řeči. Úspěšnost adaptace závisí na mnoha faktorech. Proto je uvedena řada experimentů ukazujících vliv adaptace jazykových modelů na úspěšnost rozpoznávání rozpoznávačů vyvinutých v Laboratoři počítačového zpracování řeči Technické univerzity v Liberci. Tyto experimenty bylo možné provést též díky výraznému zvýšení rychlosti rozpoznávačů a programů pro vytváření jazykového modelu.

Na závěr je uvedena metoda automatické interpunkce zvyšující čitelnost výstupu rozpoznávače spojité řeči. Uvedená metoda je schopna odhadnout pozici interpunkce pouze na základě výstupu rozpoznávače oproti jiným metodám vyžadujícím též přítomnost akustického signálu.

Annotation

Development of a language model layer for an automatic speech recognition system is understood as a complex task. This task consists of many logically following steps. Approaches used in these steps range from computational brute force, artificial intelligence, experts' help to several heuristics. Combination of different approaches is often required.

The first task is to collect text data. Several sources and their specific advantages and problems are discussed in this work. Collected text data are called text corpus. This corpus has to be cleaned before it is used. Cleaning methods partly depend on data source. Effectiveness of common cleaning methods are evaluated with respect to recognition accuracy.

Next step is to create vocabulary and assign phonetic transcription to each word in the vocabulary. Semiautomatic approach for creation of new phonological rules is presented. These rules are used in the automatic phonetic transcription. Multi-words in the vocabulary easily increase recognition accuracy. This part also discuss influence of vocabulary size on recognition speed and accuracy.

Language model computation is problematic when large vocabulary is needed. The computation requires large amount of memory. This problem is solved for different requirements. The first approach maximally saves required memory, the second one maximizes computation speed. Current software for language model computation is significantly improved, so many experiments can be performed.

Effectiveness of speech recognition improvement depends on proper experiment evaluation. The better mistakes are identified the more effective recognizer's enhancement can be. This work present improved method of results' evaluation, so mistakes are better identified.

Language mode adaptation is often discussed because of it's dependence on various factors and different results. Several experiments are performed to demonstrate influence of the adaptation on recognizer developed in SpeechLab.

Finally, automatic punctuation approach is presented. Punctuation increases readability of recognizer's output. Presented approach uses only output of the SpeechLab's recognizer, because it's output also includes information of various noises.

Obsah

1 Úvod

Prudký rozvoj hlasových technologií v posledních desetiletích je z velké části zapříčiněn výrazným nárůstem výkonu výpočetní techniky, neboť zpracování přirozené řeči, zejména její rozpoznávání, je výpočetně velice náročné. Nemalý vliv má též velké množství publikovaných prací a jistá stabilizace postupů zpracování řeči.

V současné době se lze již setkat s množstvím aplikací využívajících zpracování řeči. Syntéza řeči je používaná v dialogových a navigačních systémech. Důležitou aplikací syntézy řeči jsou systémy pomáhající nevidomým jako je například čtečka obrazovky.

Automatické rozpoznávání mluvené řeči se též aplikuje v dialogových systémech. Příkladem je systém Infocity zahrnující jak syntézu tak i rozpoznávání. Infocity je dialogový telefonní systém podávající informace o Liberci z oblasti dopravy, kultury, sportu, atd. Tento systém byl vyvinut v Laboratoři počítačového zpracování řeči Technické univerzity v Liberci. Další aplikací rozpoznávání mluvené řeči jsou programy umožňující hlasové ovládání počítače a jednoduché diktování. Nejznámější jsou Dragon NaturallySpeaking od firmy Nuance Communications, ViaVoice od IBM a SpeechMagic od firmy Phillips, která se specializuje na rozpoznávání řeči v lékařské oblasti. Pro rozpoznávání češtiny byl vyvinut systém MyVoice [1], který má pomoci zejména handicapovaným lidem v přístupu k výpočetní technice a informačním technologiím. MyVoice pochází z Laboratoře počítačového zpracování řeči Technické univerzity v Liberci a je prodáván firmou Fugasoft.

Rozsáhlejší systémy, které zahrnují rozpoznávač mluvené řeči jsou používány při přepisu televizních a rozhlasových pořadů [2]. Takový systém byl vyvinut v Laboratoři počítačového zpracování řeči. Díky segmentaci vstupního akustického signálu je rozpoznávání distribuováno na více počítačů, čímž je dosažena přijatelná odezva celého přepisovacího systému. Automaticky přepsané zprávy jsou manuálně opravovány. Tento systém byl vyvinut pro firmu Newton IT.

Přestože je již oblast automatického rozpoznávání mluvené řeči zkoumána dlouhou dobu, není zatím možné používat hlasové technologie tak pohodlně, jak bychom si přáli. Překážky pro masové rozšíření hlasových technologií jsou:

- Vysoká citlivost na prostředí, ve kterém je řeč rozpoznávána.
- Zaškolení uživatelů, aby mluvili plynule a nesnažili se různě intonovat, křičet či hláskovat, když rozpoznávač dělá chybu.
- Lokalizace rozpoznávače, zejména akustického a jazykového modelu a slovníku, neboť je manuálně, časově i finančně náročná a je nutno ji provést pro každý jazyk zvlášť.

Rozpoznávání mluvené řeči se skládá z moha úkonů, které lze rozdělit do tří základních vrstev.

Akustická vrstva se stará o nahrání a zpracování řeči do podoby příznaků vhodných pro rozpoznávání. Cílem této vrstvy je potlačit nežádoucí složky akustického signálu, jako je šum a různorodost řečníků.

Technická vrstva zahrnuje rozpoznávací proces, kdy se k signálu přiřazuje nejpravděpodobnější sekvence hlásek, které tvoří slova.

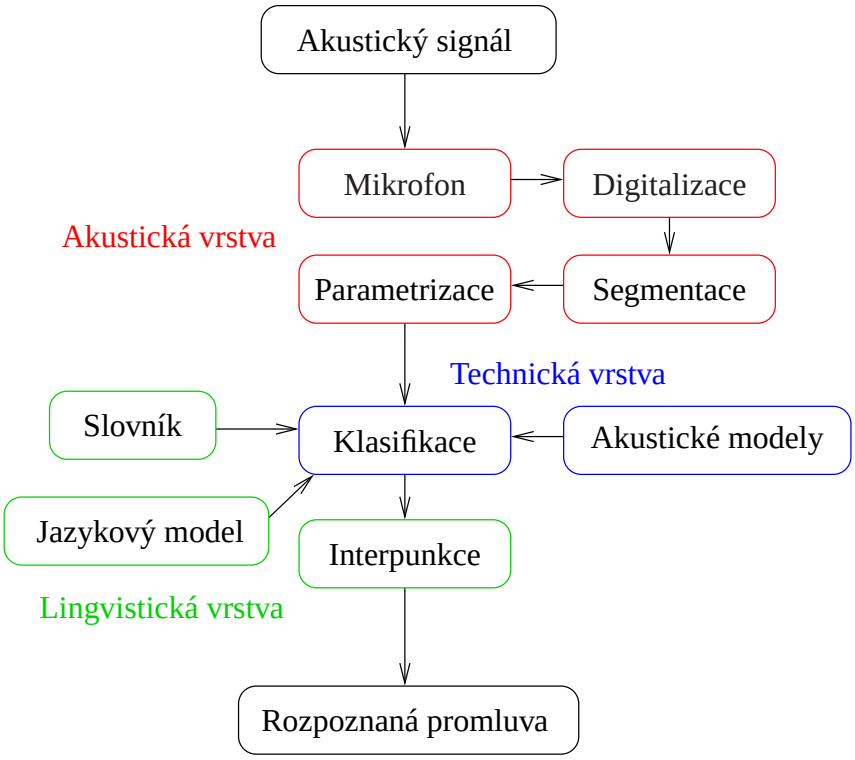
Lingvistická vrstva vystihuje zákonitosti jazyka, který je rozpoznáván, a tím pomáhá technické vrstvě v efektivnějším prohledáváním variant přiřazení sekvence hlásek a nalezením nejlepší promluvy. Lingvistická vrstva též vkládá interpunkci do rozpoznané promluvy, aby se zvýšila čitelnost výstupu rozpoznávače.

Tato práce zahrnuje komplexně pojatý problém lokalizace lingvistické vrstvy rozpoznávače mluvené češtiny. Postupy uvedené v této práci souvisí s vývojem skutečného systému rozpoznávání mluvené češtiny. V průběhu vytváření práce byl tento systém nasazen při přepisu zpravodajských pořadů. Jako modelový příklad jiné aplikace je uvedena úloha diktování lékařských zpráv.

2 Principy automatického rozpoznávání řeči

Před rozpoznáváním řeči je nutné nahraný signál obsahující řeč vhodně předzpracovat, aby obsahoval pouze informace podstatné pro rozpoznávání řeči. Toto předzpracování je v této práci nazýváno akustickou vrstvou. V rámci akustické vrstvy je odstraňován šum a potlačována různorodost mluvčích tak, aby bylo rozpoznávání na mluvčím minimálně závislé. Nahraný akustický signál je digitalizován ve zvukové kartě. Digitalizace spočívá ve vzorkování signálu a následné kvantizaci vzorků pomocí analogově-číslicového převodníku. Vzorky jsou dále rozděleny na krátké segmenty o délce 25 ms, které se nazývají framy. Sousední framy se vzájemně překrývají. Délka framu se volí tak, aby bylo možné považovat signál v rámci framu za stacionární. Dalším krokem zpracování signálu je parametrizace, která převede framy na příznaky splňující následující požadavky. Pomocí příznaků by mělo být možné jednoduše identifikovat jednotlivá slova ve slovníku. Zároveň by měly potlačit vliv různých řečníků (výška hlasu, síla signálu). Dále by příznaků nemělo být mnoho a měly by být jednoduše vypočítatelné. Nejpoužívanějšími příznaky jsou MFCC příznaky [3] pro nízkou citlivost na šum v signálu. Tyto příznaky jsou taktéž používány v rozpoznávačích použitých v této práci. Existuje množství přístupů pro zpracování signálu pro rozpoznávání řeči. Přehled základních metod lze nalézt v [3].

Rozpoznávání řeči se dělí na dvě základní úlohy a to rozpoznávání izolovaných slov a rozpoznávání spojité řeči. Rozpoznávání izolovaných slov je lehčí varianta. Cílem rozpoznávání izolovaných slov je přiřadit nahranému zvuku právě jedno nejpravděpodobnější slovo. Součástí každého rozpoznávače izolovaných slov musí být detektor začátku a konce slova. Tento detektor využívá znalosti energie signálu reprezentované jedním z příznaků vytvořených v akustické vrstvě rozpoznávání. Rozpoznávání spojité řeči se k dané nahrávce snaží najít nejpravděpodobnější sekvenci slov. Není předem známo, kolik slov nahrávka obsahuje, v jakém pořadí jsou slova vyslovena a ani hranice slov není známa. Tato úloha má exponenciální složitost. Základní schéma rozpoznávání řeči je na obrázku 1.



Obrázek 1: Etapy rozpoznávání mluvené řeči

3 Východiska

Tato práce je úzce spjata s vývojem systému pro automatický přepis televizních a rozhlasových pořadů a výsledky práce jsou v tomto systému uplatněny. Při vývoji rozsáhlého sytému pro rozpoznávání mluvené češtiny bylo třeba odpovědět na řadu koncepčních i dílčích otázek, vyřešit řadu dílčích úloh, implementovat je

do modulů a tyto moduly správně propojit. Vzhledem k praktickému nasazení pak bylo též nutné řešit úlohy efektivní a paralelní správy slovníku a jazykového modelu a možnosti jejich časové adaptace.

Otázky, na které bylo třeba najít odpovědi:

Jak velký musí být slovník, aby dostatečně pokrýval češtinu? Je zřejmé, že pokud není slovo ve slovníku, není možné, aby jej rozpoznávač rozpoznal. Pokud se v promluvě vyskytne slovo, které není ve slovníku, udělá rozpoznávač chybu tím, že jej zamění za jiné podobné slovo ve slovníku. Často však rozpoznávač zamění chybějící slovo a jeho okolí sekvencí slov ze slovníku, což způsobí více chyb. Je zřejmé, že pro inflektivní jazyky s velikým počtem slov nebude možné pracovat s kompletním slovníkem všech slov, což je dáno zejména vysokými výpočetními nároky při používání velmi velkého slovníku. Podobná slova ve slovníku mají i podobné akustické modely, což vede k častým chybám v rozpoznávání. Veliké slovníky je též obtížné spravovat, a proto obsahují množství chyb jako jsou překlepy nebo špatné fonetické přepisy.

Z jakých zdrojů tento slovník tvořit? Velký slovník vyžaduje velké množství textu z dané aplikační oblasti. Z tohoto textu je odvozen jak slovník tak i jazykový model. Nejpřístupnějším zdrojem dat pro přepis zpráv jsou webové portály zpravodajských pořadů. Použití webu jako zdroje dat s sebou přináší mnohé problémy. Je nutné vytvořit dostatečně robustní programy schopné pracovat 24 hodin denně, 365 dní v roce, neboť množství textu vytvořeného za jeden den není příliš veliké a navíc se často opakuje v různých zdrojích. Při stažení stránky je třeba zkontrolovat, jestli obsahuje požadované informace a provést extrakci podstatných dat, kterých může být na celé stránce i méně než třetina.

Jak předzpracovat výchozí text? Nasbíraný text obsahuje množství zkratek a číslovek, které je nutné rozepsat do tvaru více podobného jejich výslovnosti. Tento úkol není pro inflektivní jazyky jednoduchý, neboť přepis některých zkratek a číslovek je nutné vytvořit ve správném tvaru, což je někdy možné až po analýze okolí slova. Některé zkratky je naopak vhodné přidat do slovníku tak, jak jsou, a vytvořit pouze alternativní výslovnosti.

Je vhodné přidat do slovníku i slovní spojení? Slovní spojení ve slovníku je v mnoha inflektivních jazycích spíše problém, který pouze zvětšuje slovník a ředí data pro jazykový model. Na druhé straně je zřejmé, že krátká slova způsobují vyšší chybovost než slova dlouhá, proto je výhodné je spojit se sousedními slovy a vytvořit tak jedno slovní spojení zapsané ve slovníku jako jedno slovo. Tato úloha je spíše úlohou nalezení vhodného kritéria pro výběr slovních spojení.

Jak efektivně vytvořit výslovnost ke slovům? Slovník obsahuje veliké množství slov a jejich manuální fonetická transkripce je v přijatelné době nerealizovatelná. Proto je třeba použít a implementovat fonologická pravidla, která provedou automatickou fonetickou transkripci. V češtině se však vyskytuje množství slov cizího původu, na která nejsou česká fonologická pravidla aplikovatelná. Nejčastější problémy tvoří přepis slabik di, ti a ni. Pro tyto slabiky je třeba nalézt další pravidla tak, aby slova správně přepsaná českými pravidly nebyla poškozena a nová pravidla opravila co nejvíce chyb.

Jak upravit výstup rozpoznávače, aby byl co nejvíce čitelný? Výstup rozpoznávače je tvořen sekvencí mezerami oddělených slov, což je značně nečitelné. Automatická interpunkce a správná první velká písmena významně zvyšují čitelnost. Interpunkce je z části závislá na intonaci, tudíž na vstupním signálu. Při rozpoznávání je však automatická interpunkce posledním článkem, a tudíž od signálu velmi vzdálena.

Jak důležitá je pravidelná aktualizace slovníku a jazykového modelu? Je zřejmé, že se témata ve zprávách v čase mění. Je tudíž nutné provádět občasné aktualizace slovníku a jazykového modelu. Hlavní otázkou je jak často, neboť i tato operace zabírá čas, který může být při méně častých úpravách využit efektivněji. Aktualizaci slovníku není totiž možné provádět zcela automaticky z důvodu velkého množství překlepů vybraných frekvenční analýzou za kandidáty na přidání.

Jak lze adaptovat lingvistickou vrstvu pro jinou aplikační oblast? Pokud je již k dispozici rozsáhlý systém pro přepis zpráv, je žádoucí, aby mohl být co nejsnadněji použit i v jiných aplikacích. Otázkou je, co bude nutné provést pro jeho adaptaci a kolik to bude stát.

4 Dílčí úlohy

Cílem této práce je tvorba lingvistické vrstvy pro rozpoznávač řeči s tím, že veškeré kroky jsou maximálně automatizovány. Výsledky výzkumu jsou aplikovány na rozpoznávač izolovaných slov [4] a spojité řeči [2] vyvíjené v Laboratoři počítačového zpracování řeči technické univerzity v Liberci. Oba rozpoznávače jsou primárně určeny pro rozpoznávání češtiny, čímž je také demonstrována lokalizace jazykového modelu a slovníku, a tedy snížení jedné z překážek masového rozšíření hlasových technologií.

Tvorba lingvistické vrstvy pro rozpoznávač mluvené češtiny zahrnuje mnoho akcí, z nichž některé jsou plně automatizovatelné, některé jen částečně a některé může udělat pouze manuálně specialista, například přepis lékařských zkratek. Při

automatizaci jednotlivých akcí je nutné aplikovat jak hrubou výpočetní sílu, tak i heuristické informace a metody umělé inteligence. Hlavní úkoly při tvorbě lingvistické vrstvy jsou:

Tvorba textového korpusu: Pro rozpoznávání zpráv z televize a rádia [5] lze sbírat novinové články z webových stránek. Sběr dat z webu lze zajistit robustním programem automaticky prohledávajícím zvolené stránky. Při vytváření slovníku pro lékařský diktovací systém [6] je nutné z dat vypustit osobní informace pacientů, což komplikuje sběr dat.

Čištění a normalizace nasbíraných dat: Nasbíraná data obsahují číslicemi psané číslovky a zkratky, které je nutno rozepsat. U číslic se tak zmenší počet různých slov a zlepší se jazykový model. U zkratek se zjednoduší fonetický přepis. Expanze zkratek a číslovek není triviální, neboť je nutné vygenerovat správný tvar (pád), což nelze provést vždy automaticky. V některých případech pomůže automatická morfologická analýza [7]. Speciální zkratky mohou přepsat jen specialisté, kteří je používají. Čištění je operace výrazně závislá na jazyce a konkrétním zdroji textů.

Výběr slov do slovníku: Do slovníku se ze získaných textů vybírají nejčetnější slova jazyka. Se snižující se četností výskytu slov přibývá překlepů, cizích a nesmyslných slov. Čeština jako inflektivní jazyk obsahuje mnoho tvarů slov, někdy i správné tvary mohou být méně četné než překlepy. Proto nelze proces výběru slov do slovníku plně automatizovat.

Vytvoření fonetické transkripce slov ve slovníku: Fonetická transkripce definuje napojení slova na akustické modely. Pro češtinu existuje soubor pravidel, která platí pro většinu českých slov. Cizí slova je nutné většinou přepisovat ručně. Jiné jazyky, například angličtina, mohou mít fonetickou transkripci obtížněji algoritmizovatelnou. Pro implementaci fonologických pravidel se používají produkční pravidla, stavové automaty nebo neuronové sítě.

Vytvoření jazykového modelu: Používaný rozpoznávač řeči používá jazykový model ve formě dvojic sousedních slov. Pro slovník obsahující 300000 slov je teoreticky možných 300000^2 slovních dvojic je proto nutné zabývat se implementací počítání těchto dvojic, aby je bylo možné umístit do paměti běžně dostupných počítačů.

Adaptace jazykového modelu: Při rozpoznávání televizních zpráv dochází v průběhu času ke změně témat. Proto je nutné aktualizovat slovník i jazykový model, což zahrnuje všechny předchozí akce, ale s menším množstvím dat a větším množstvím šumu v datech (překlepů). Adaptace jazykového

modelu spočívá ve vhodném kombinování různých existujících jazykových modelů tak, aby výsledný model měl minimální perplexitu na testovacích datech.

Automatická interpunkce: Výstupem automatického rozpoznávače spojité řeči je mezerami oddělený proud slov. Pro zvýšení čitelnosti tohoto výstupu je nutné provést automatickou interpunkci zvýrazňující konce vět. Automatická interpunkce kromě akustické informace využívá též informaci z jazykového modelu.

Detailní analýza výsledků rozpoznávání: Pro efektivní zvyšování úspěšnosti rozpoznávání je výhodné vědět, která slova jsou nejčastěji špatně rozpoznávána. Běžně používaná metoda vyhodnocování výsledků rozpoznávání počítá slova, která jsou zaměněná, vložená, či vypuštěná. V případě výskytu sekvence chyb není běžnou metodou zjišťováno, které slovo je vložené a které zaměněné, je pouze zjištěno, že jedno je vložené a jedno zaměněné.

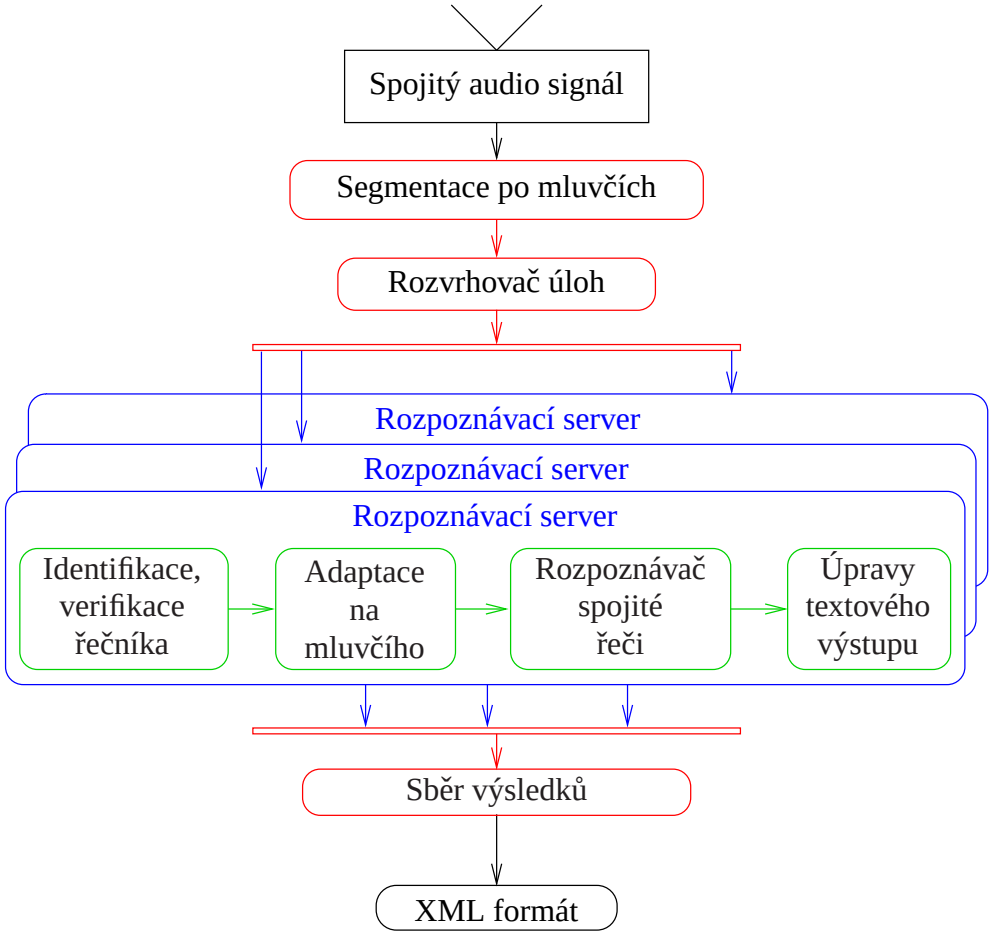
5 Systém automatické transkripce televizních a rozhlasových pořadů

V Laboratoři počítačového zpracování řeči Technické univerzity v Liberci byl vyvinut systém pro automatický přepis televizních a rozhlasových pořadů. Tento systém je velmi modulární, což umožňuje provádět množství různých experimentů. Systém je implementován tak, že rozpoznávače běží na několika počítačích najednou a úloha rozpoznávání je distribuována, čímž se zrychlí provádění experimentů na reálných datech. Zrychlením provádění experimentů také dochází k rychlejšímu vývoji v oblasti rozpoznávání řeči, neboť je možné provádět i experimenty, které nebyly realizovány díky obtížně predikovatelným výsledkům a velkým časovým nárokům.

Tento systém je nasazen v komerční sféře na přepis televizních a rozhlasových pořadů. Schéma systému je uvedeno na obrázku 2.

6 Tvorba textového korpusu

Vytváření jazykového modelu a slovníku vyžaduje shromáždění velkého množství textu. Čím je větší slovník, tím více textu je potřeba. Nasbíraný text by měl být z oblasti, kde bude rozpoznávač řeči používán, aby slovník rozpoznávače pokryl co nejvíce nejčetnějších slov a jazykový model zachytil a spolehlivě odhadl hodnoty bigramů či trigramů. Je všeobecně známé, že často vyskytujících se slov



Obrázek 2: Systém pro přepis televizních a rozhlasových pořadů

je málo a řídce vyskytujících se slov je mnoho. To ukazuje i fakt, že přibližně polovina slov ze 2.4 milionu různých slov vyskytujících se ve 3.5 GB nasbíraných novinových článků byla viděna právě jednou.

6.1 Zdroje dat

Nejvíce veřejně přístupných textů se v dnešní době nalézá na webových stránkách. Pro přepis televizních a rozhlasových zpráv jsou nejvhodnější webové portály denního tisku, protože jsou snadno dostupné a nejvíce se obsahově přibližují vysílaným zprávám. Přepisy televizních a rozhlasových zpráv je nejběžnější dnešní aplikace rozpoznávačů spojité řeči kvůli vysoké kvalitě studiových nahrávek a snadnosti přístupu.

Pro automatizaci stahování webových stánek existuje množství nástrojů. Zr-

cadlo webového portálu lze jednoduše vytvořit například programem *wget*. Pokud je nutné další zpracování nasbíraných dat je výhodnější použít specializovanou knihovnu pro nějaký programovací jazyk, například knihovna *LWP* pro jazyk *Perl*, která je dobře popsána v [8]. *LWP* umožňuje transformaci webových stránek do stromové struktury a tím i snadné vyhledávání relevantních odkazů na další stránky.

Jako zdroj textových dat pro přepisy televizních a rozhlasových zpráv lze využít služeb firem, které zprávy přepisují. Těchto dat je mnohem méně a jsou drahá, neboť se jedná o ruční přepisy. Na druhou stranu však nejlépe reflektují jazyk, který je používán ve zprávách.

Získat data pro lékařský diktovací systém je oproti novinovým článkům výrazně složitější a je jich mnohem méně. Lékařské zprávy často obsahují osobní údaje, které jsou chráněny zákonem proti zveřejnění. Osobní údaje není vždy možné spolehlivě eliminovat, proto nejsou tyto texty přístupné.

6.2 Normalizace textového korpusu

Cílem normalizace získaných textů je přiblížení psané formy řeči formě vyslovované především expanzí zkratek a přepisem číslic. Normalizace se též snaží snížit počet různých zápisů slov, například „gymnázium, gymnasium” jsou stejná slova a zbytečně zvyšují velikost jazykového modelu a slovníku. Převodník těchto slov na unifikovanou formu byl vytvořen ručně [9]. Normalizace textu je prováděna ad-hoc. Normalizace korpusu pro rozpoznávač spojité řeči je prováděna v následujících krocích:

1. Expanze zkratek, které nejsou skloňovány, například: apod., tzn. Tyto jedno či dvouslovné zkratky jsou expandovány, neboť jsou expandovány i ve slovníku. Zkratky obsahující více slov jako s. r. o., v. o. s expandovány nejsou.

2. Expanze číslic následovaných slovem letý, letou, . . . , např. 50letý na padesátiletý.

3. Přepis zkratky tzv. na takzvaná, takzvaný, . . . Jde o velmi četnou zkratku s netriviálním přepisem, neboť je její přepis skloňován.

4. Standardizace, kdy jsou začátky vět označeny speciálním tagem.

5. Přepis zkratky hod. na hodin, hodina, . . . Přepis je netriviální, neboť slovo hodina je nutné skloňovat.

6. Přepis datumů [1], například *4. ledna* na *čtvrtého ledna*.

7. Expanze číslic s předložkou s využitím znalosti mluvnických kategorií následujícího slova, například *ve 4. patře* na *ve čtvrtém patře*.

8. Expanze číslic vyjadřujících čas ze spojení: V X hodin, například *v 5 hodin* na *v pět hodin*.

9. Expanze zbylých číslic. Tato normalizace zahrnuje pravidla pro přepis desetinných čísel a číslovky částečně vyjádřené slovem, například *7 milionů* se přepíše na *sedm milionů*.

10. Aplikace převodníku jednoduchých slov. Pomocí ní se sníží počet alternativních textových variant slov. Například slova *benzín* a *benzin* znamenají stejnou věc a píší se podobně.

11. Expanze číslic před slovem krát, například *5krát* na *pětkrát*.

12. Drobné úpravy na základě vizuální inspekce korpusu. Pravidla úprav jsou vytvářena na základě vizuální inspekce korpusu a výsledků rozpoznávače. Tyto úpravy zahrnují například přepsání *př. n. l.* na *před naším letopočtem*.

Výsledný korpus je převeden na malá písmena.

6.3 Vliv normalizace na úspěšnost rozpoznávání

Pro zjištění vlivu normalizace korpusu na úspěšnost rozpoznávání byl proveden experiment, kdy byly originální texty postupně normalizovány. Z korpusů jednotlivých kroků normalizace byly vytvořeny jazykové modely. Experiment byl proveden na databázi TV2005. Výsledky jsou uvedeny v tabulce 1.

Z výsledků je patrné největší zvýšení úspěšnosti rozpoznávání při označování začátku vět a oddělení nealfanumerických znaků od slov. P-hodnota testu statistické významnosti zlepšení mezi originálními texty (pořadí operace 0) a poslední úpravou (pořadí operace 13) je 4.6e-09. Textový korpus zahrnuje v současné době 3.5 GB textových souborů, převážně článků z denního tisku. Korpus obsahuje přibližně 519 milionů slov, z toho 2375859 různých.

[1] V této práci je používáno nespisovné skloňování slova *datum*, aby byla eliminována podobnost s tvary slova *data*, které se zde též vyskytuje. Stejné tvary těchto slov by mohli vést k mylnému pochopení textu

Název operace	pořadí operace	úspěšnost rozpoznávání (Acc)
Originální texty	0	78.64 %
Jednoduché zkratky	1	78.63 %
X-letý	2	78.50 %
Tzv.	3	78.65 %
Standardizace	4	79.90 %
Hod.	5	79.98 %
Datum	6	80.05 %
Číslice s předložkou	7	80.11 %
V X hodin	8	80.09 %
Ostatní číslice	9	80.03 %
Převodník	10	80.18 %
X krát	12	80.08 %
Ruční úpravy	13	80.04 %

Tabulka 1: Vliv normalizace na úspěšnost rozpoznávání

7 Principy výběru slov do slovníku

Slovník je důležitou součástí rozpoznávače řeči, neboť rozpoznávač pracuje pouze se slovy ve slovníku. Pokud není slovo ve slovníku, nemůže být rozpoznáno. Slovník je vytvářen z velkého množství textu vybráním nejčetnějších slov, tím se dosáhne vysokého pokrytí zdrojového textu slovníkem. Text, ze kterého je vytvářen slovník, musí obsahovat témata, pro která je rozpoznávač navrhován. Pro požadované pokrytí je velikost slovníku závislá na jazyku, pro který je slovník vytvářen. Angličtina obsahuje málo slov, proto stačí slovník o desítkách tisíc slov. Mnoho slovních tvarů jazyka zvyšuje velikost slovníku. Aby se dosáhlo stejného pokrytí českého textu jako anglického, je nutné použít větší slovník. Pro 99% pokrytí anglického či španělského textu je třeba slovník o velikosti 65 tisíc slov [10], [11]. Příklad slovníku je uveden v tabulce 2.

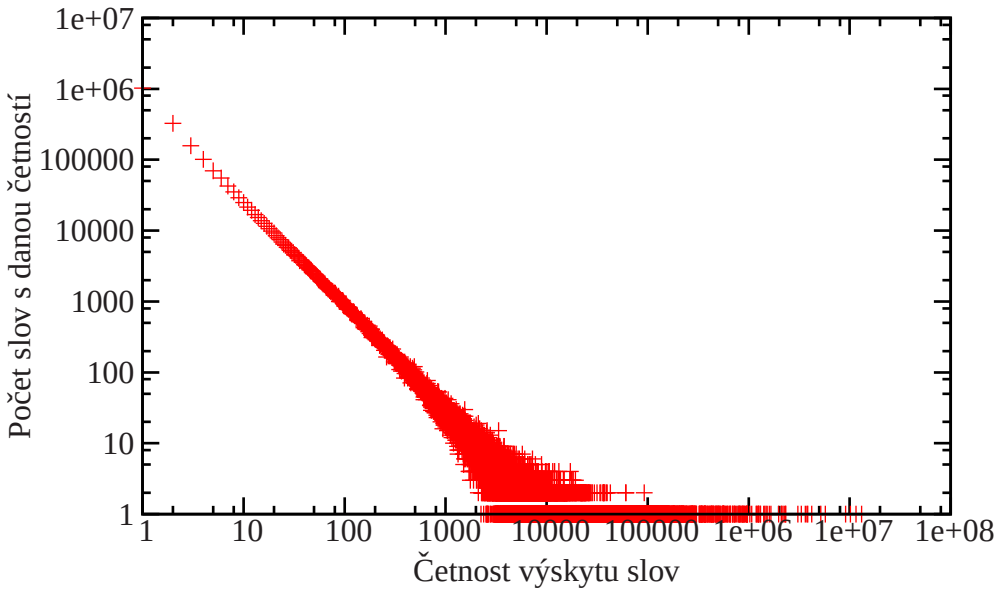
8 Charakteristiky slovníku pro rozpoznávač řeči

Slovník pro rozpoznávání řeči odvozen z textového korpusu obsahujícího 3.5 GB textu, převážně článků z denního tisku. Korpus obsahuje přibližně 519 milionů slov, z toho 2375859 různých. Histogram četností slov v textovém korpusu, viz obrázek 3, ukazuje, že četných slov je jen málo, zatímco různých slov s občasným výskytem je mnoho, což je důvod používání velkých slovníků pro rozpoznávání češtiny.

Pokrytí textového korpusu různě velikými slovníky je uvedeno v tabulce 3. Ze

Slovo	výslovnosti
absolventi	apsolveňťi
absolventka	apsolventka
absolventská	apsolvencká
absolventské	apsolvencké
absolventského	apsolvenckého
absolventském	apsolvenckém
absolventský	apsolvenckí
absolventských	apsolvenckíX apsolvenckíh
účinkujícím	účiNkujícím
účinkujícími	účiNkujícími
účinnost	účinost účinozd
účinnosti	účinosťi

Tabulka 2: Příklad slovníku



Obrázek 3: Histogram četností výskytu slov v textovém korpusu. Graf je vykreslen v logaritmických souřadnicích.

slovníku používaném v rozpoznávači spojité řeči obsahujícím 312 tisíc slov byly odvozeny menší slovníky na základě četnosti výskytů slov v textovém korpusu.

Počet slov ve slovníku	Počet pokrytých slov	Pokrytí
25000	447807165	86.23 %
50000	474605302	91.39 %
75000	486883723	93.76 %
100000	494064808	95.14 %
125000	498796835	96.05 %
150000	502145401	96.70 %
175000	504637124	97.18 %
200000	506561303	97.55 %
225000	507514229	97.73 %
250000	509313731	98.08 %
275000	510327772	98.27 %
300000	511175175	98.44 %

Tabulka 3: Pokrytí textového korpusu různě velikými slovníky

9 Fonetická transkripce

Při rozpoznávání řeči je důležité vzájemné spojení textové a akustické formy slova. Pro každé slovo je nutné mít jeho akustický model, který je porovnáván s akustickým signálem přicházejícím do mikrofonu. Pro velké slovníky není možné vytvořit akustický model pro každé slovo zvlášť. Proto jsou vytvářeny akustické modely pro menší stavební jednotky slova, například fonémy, a ty jsou poté spojovány. Foném je nejmenší jednotka řeči, která může rozlišovat jednotlivá slova [12]. Fonémů je podstatně méně než slov. Je možné používat i jiné větší stavební jednotky slov, ale vždy je nutné volit kompromis mezi počtem jednotek, složitostí přepisu textu na jednotky a dostatečným množstvím dat, ze kterého jsou akustické modely natrénovány.

Pro zápis fonémů je možno použít mezinárodní fonetickou abecedu (IPA) popsanou v [13]. Pro češtinu byla vypracována abeceda PAC [14] přehledněji vystihující česká fonetická pravidla. Česká fonetická abeceda je i implementačně výhodnější. Rozpoznávače řeči používané v této práci pracují s modely fonémů, kterých je 40 [14]. Dále jsou přidány modely nejběžnějších šumů a hluků [15].

Fonetická transkripce je přepis textové podoby slova na sekvenci fonémů. V každém jazyce existují fonologická pravidla jak provádět fonetickou transkripci (vyslovovat slova). V některých jazycích, jako je angličtina existuje veliké množství pravidel. Naopak v češtině nebo němčině je pravidel mnohem méně a je možné je jednodušeji implementovat.

Fonetická transkripce češtiny není jen pouhý přepis písmen na odpovídající fonémy. Často dochází ke koartikulaci, kdy je písmeno přepsáno na foném v závislosti na jeho okolí. V tomto případě může být jedno písmeno přepsáno na ně-

kolik různých fonémů nebo úplně vypuštěno z přepisu.

Fonologická pravidla mohou být implementována ve formě produkčních pravidel [16], [17], rozhodovacích stromů [18], konečných automatů [19] nebo neuronové sítě [20, 21].

Pro češtinu bylo fonetiky vypracováno množství fonologických pravidel [12] přepisujících hlásky na fonémy. Fonologická pravidla jsou ve formě produkčních pravidel a popisují jak přepsat hlásky na fonémy v závislosti na jejich okolí.

Česká fonologická pravidla fungují spolehlivě pro česká slova. Cizí slova, zejména ta, kde se vyskytují slabiky *di*, *ti*, *ni* bývají často špatně přepisována na *ďi*, *ťi*, *ňi*. Například slovo *antimon* je přepsáno podle českých pravidel na *ant'imon* nikoli na *antymon*. Některá slova jsou přepisována podle českých i cizích pravidel současně, například slovo *antirasisti* má být přepsáno na *antyrasisťi*, ale podle českých fonologických pravidel by bylo přepsáno na *ant'irasisťi*. V příkladech je použit foném *y*, přestože není v PAC. Cílem je čitelnost příkladů. Ve skutečné fonetické transkripci není rozdíl mezi *y* a *i*, proto se obě zapisují jako *i*

Řešení správného přepisu cizích slov spočívá v zavedení výjimek. Výjimky se při fonetické transkripci aplikují jako první. Standardní fonologická pravidla jsou aplikována jako druhá. Výjimek však rychle přibývá a stávají se nepřehlednými, což může vést k poškození správné fonetické transkripce slov, která byla správně přepsána standardními fonologickými pravidly.

Jiným řešením fonetického přepisu cizích slov je odvození nových pravidel ze známých přepisů slov. Tím se i sníží počet výjimek. V této práci jsou odvozována nová fonologická pravidla ve formě produkčních pravidel. Originální sada pravidel je převzata z [16]. Nová pravidla mají původní sadu rozšířit, proto jsou ve stejné formě. Zabrání se tím reimplementaci původních pravidel. Nová fonologická pravidla jsou odvozována pomocí gramatické evoluce přímo do požadovaného formátu. Odvozování nových fonologických pravidel je uvedeno v [22].

9.1 Nová fonologická pravidla

Jak již bylo uvedeno, jsou fonologická pravidla odvozována jako produkční pravidla ve formátu

$$\text{písmeno} \rightarrow \text{foném}/\text{prefix}\ \_\text{postfix},\ \text{krok}, \tag{1}$$

kde prefix a postfix jsou sekvence písmen předcházející a následující přepisované písmeno. Krok označuje, kolik následujících písmen má být při fonetické transkripci přeskočeno. Krok umožňuje přepsat několik písmen najednou. Při odvozování nových fonologických pravidel jsou počet kroků, foném a přepisované písmeno fixní. Tím je zjednodušeno učení nových pravidel a zároveň je tak možné se soustředit na případy, které jsou nejčastěji chybně přepsány jako jsou slabiky di, ti, ni.

Nová pravidla by měla pokrýt co nejvíce špatně přepsaných slov a zároveň by neměla bránit aplikaci původních pravidel, pokud je jimi slovo správně přepisováno. Fonetický přepis je prováděn v následujících krocích.

1. Aplikuj výjimky.
2. Aplikuj nová pravidla.
3. Aplikuj originální pravidla.

Pravidla jsou uspořádána od nejspecifičtějších aplikovatelných na málo případů po nejobecnější aplikovatelná na libovolné písmeno bez ohledu na jeho kontext. Pokud je nějaké pravidlo aplikováno, je písmeno přepsáno a další pravidla se již na něj neaplikují. V opačném je na písmeno aplikováno následující obecnější pravidlo.

Při odvozování nových fonologických pravidel jsou výjimky ignorovány. Učené pravidlo je aplikováno jako první. Následně jsou aplikována ostatní pravidla.

9.2 Trénovací a testovací data

Trénovací a testovací množiny jsou vytvořeny pro každou trojici {písmeno, foném, krok} zvlášť. Hledá se jen prefix a postfix. Trénovací a testovací vzorky jsou vybrány ze slovníku obsahujícího 200000 slov. Všechna slova obsahující přepisované písmeno jsou vybrána ze slovníku a roztříděna do třech skupin. První skupina obsahuje slova, která jsou správně přepsána pomocí originálních fonologických pravidel. Druhá skupina je tvořena slovy, která by mohla být správně přepsána, kdyby bylo aplikováno nějaké nové pravidlo, které je hledáno. Třetí skupina jsou slova, u nichž nelze jednoduše odhadnout, zda jejich přepis nové pravidlo opraví. Toto rozdělení lze provést plně automaticky tak, že se pro všechna slova špatně přepsaná originálními pravidly provede i alternativní přepis hledaným pravidlem s prázdným prefixem i postfixem. Pokud je mezi alternativami přepis shodný s přepisem ze slovníku, pak je možné fonetický přepis opravit nějakým novým pravidlem. Trénovací a testovací množiny jsou vytvořeny z prvních dvou skupin a to tak, že dvě třetiny jsou trénovací a jedna třetina testovací.

9.3 Experimenty a výsledky

Všechny experimenty probíhaly s populací tvořenou 500 jedinci. Během evoluce bylo vyhodnoceno 50500 jedinců. Rodiče, na které byly aplikovány genetické operátory byli vybíráni turnajovou selekcí mezi třemi jedinci. Nová populace byla

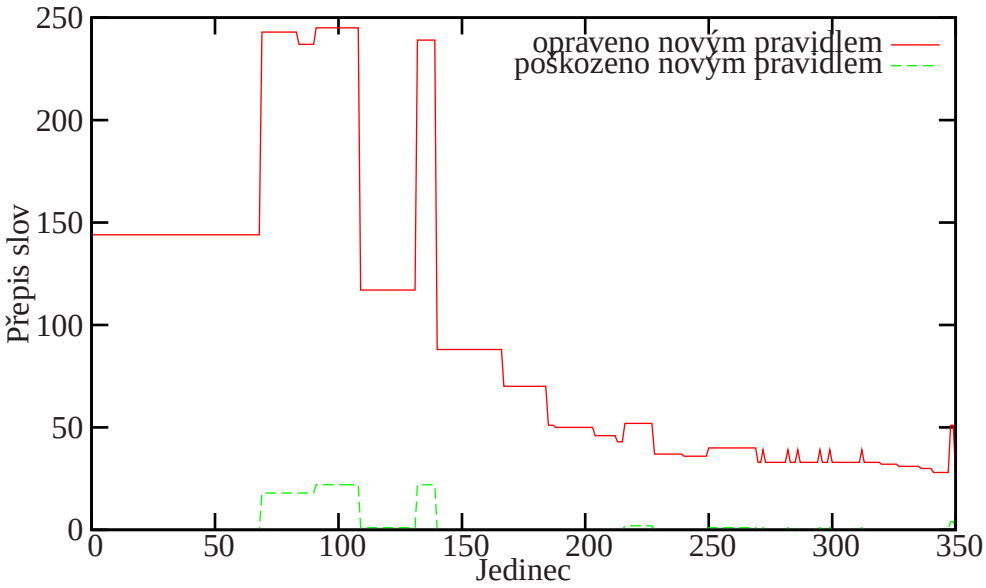
vytvářena metodou „steady state“ [23] tak, že 80 % jedinců zůstalo a 80 % nových jedinců bylo vytvořeno křížením, ostatní mutací. Diverzita populace byla udržována metodou LICE [24].

Fitness funkce neboli kritérium výběru jedince je popsáno níže. Trénovací množina je rozdělena na slova správně přepsaná originálními fonologickými pravidly C a slova, jejichž automatická transkripce může být opravena novým pravidlem. Nechť B je počet slov s transkripcí opravenou novým pravidlem a $\bar{C}$ je počet pravidel z C, jejichž přepis je novým pravidlem poškozen. Fitness je pak

$$f = B - w\bar{C}, \tag{2}$$

kde $w \in \langle 0, \infty \rangle$ je váha penalizující transkripci poškozenou novým pravidlem. Pro všechny experimenty je $w = 6$.

V rámci experimentů jsou hledána pravidla pro nejproblematičtější slabiky di, ti, ni. Výsledná fonologická pravidla byla ručně vybírána z poslední populace, a to taková pro která bylo $\bar{C} = 0$, tedy žádný přepis nebyl novým pravidlem poškozen. 350 nejlepších jedinců poslední populace je ukázáno na obrázku 4



Obrázek 4: 350 nejlepších jedinců poslední populace

Bylo nalezeno 38 nových pravidel, která byla přidána k 248 originálním pravidlům.

Slovník s 200000 slovy byl přepsán pomocí originálních fonologických pravidel s výjimkami a s novou sadou fonologických pravidel bez použití výjimek. Úspěšnost přepisu je uvedena v tabulce 4. Počet slov, která jsou opravena jednotlivými pravidly není lehké přesně určit, neboť slovo může být opraveno dvěma pravidly zároveň.

	Správně přepsáno	úspěšnost
Originální pravidla s výjimkami	185237	93 %
Přidána nová pravidla	189807	95 %

Tabulka 4: Experimentální výsledky s novými fonologickými pravidly

Počty opravitelných a opravených fonetických přepisů jsou uvedeny v tabulce 5.

	di	ti	ni	celkem
Opravitelné chyby	3746	1392	2021	7159
Opravené chyby	-	-	-	4570

Tabulka 5: Opravitelné a opravené chyby fonetické transkripce

9.4 Vyhodnocení

Výsledky ukazují, jak lze téměř automaticky vylepšit fonetickou transkripci. Uvedený přístup znovu neobjevuje všeobecně známá fonologická pravidla. Gramatická evoluce umožnila najít pravidla v takovém formátu, aby je bylo možné použít v existujícím systému automatické fonetické transkripce. Jak bylo předpokládáno, nová nalezená pravidla jsou velmi specifická a aplikovatelná na menší počet slov než originální fonologická pravidla.

10 Slovní spojení ve slovníku

Na základě analýzy rozpoznávaných promluv bylo zjištěno, že krátká slova jsou často špatně rozpoznána. Krátká slova jsou ignorována, rozpoznána jako šum, nebo přidána jako předpona či přípona následujícího nebo předcházejícího slova. Dlouhá slova jsou většinou rozpoznána správně. Slovní spojení krátkého frekventovaného slova a jeho častého následníka, či předchůdce může zvýšit úspěšnost rozpoznávání, neboť je toto spojení chápáno rozpoznávačem jako jedno dlouhé slovo. Slovní spojení jsou již do slovníku rozpoznávačů spojité řeči přidávána [5, 25] ručně. Cílem této sekce je zjistit vliv plně automatického přidávání slovních spojení do slovníku na úspěšnost rozpoznávání.

Přirozená slovní spojení, která se v některých jazycích běžně vyskytují, mohou způsobovat problémy tím, že zvětšují velikost slovníku a zvyšují řídkost textového korpusu, ze kterého je počítán jazykový model. Tyto problémy se týkají zejména jazyků, kde se nová slova běžně vytvářejí spojením existujících slov jako

například v němčině nebo finštině. Několik postupů, jak rozbíjet tato spojení bylo publikováno v [26], [27].

Slovní spojení automaticky přidávaná do slovníku jsou tvořena ze slov již ve slovníku existujících, čímž je eliminováno riziko vložení překlepu, či nesmyslného slova. Slovní spojení mohou být vybírána buď na základě vzájemné informace, nebo četnosti výskytu spojení v textovém korpusu.

10.1 Míry pro výběr slovních spojení

Kritérium pro výběr vhodného slovního páru musí splňovat následující požadavky:

- Slovní spojení musí obsahovat alespoň jedno krátké slovo. Slovní spojení dlouhých slov přispívá pouze k řídkosti textového korpusu. Dlouhá slova nejsou cílem optimalizace rozpoznávače.
- Slovní spojení musí být četné, aby se pouze nezvětšoval slovník a řídkost dat.
- Slova ve slovním spojení musí být četná, neboť četná slova jsou spíše česká než cizí a je možné aplikovat automatický fonetický přepis s nižším rizikem nesprávného přepisu.

Jako krátká slova jsou chápána slova mající maximálně 3 znaky a minimální četnost výskytu každého slovního spojení je stanovena na 30.

10.1.1 Vzájemná informace

Vzájemná informace je často používána k výběru kolokací. Kolokace jsou slova, která se často vyskytují spolu a zřídka zvlášť. Vzájemná informace je definována následovně:

$$PMI = \log\left(\frac{p(w_1, w_2)}{p(w_1)p(w_2)}\right), \tag{3}$$

kde $p(w_1, w_2)$ je pravděpodobnost sekvence slov w_1 a w_2, $p(w_1)$ je pravděpodobnost slova w_1 jako předchůdce a $p(w_2)$ je pravděpodobnost slova w_2 jako následníka.

10.1.2 Četnost výskytu slovního spojení

Četnost výskytu slovního spojení je nejjednodušší způsob výběru slovního spojení. Četnost výskytu splňuje požadované vlastnosti na kriteriální funkci a je počítána při vytváření jazykového modelu.

10.2 Experimenty

Experimenty byly prováděny na databázi COST278. Základní úspěšnost rozpoznávání pro slovník bez slovních spojení bylo 74.48 %. Základní úspěšnost rozpoznávání se slovníkem s manuálně vybranými 1731 slovními spojeními bylo 75.80 %. P-hodnota při testování s ručně vybranými spojeními oproti slovníku bez spojení je 1.1e-04. Při manuálním výběru slovních spojení byl brán ohled na: kolokace, slova objevující se často spolu a zřídka zvlášť, běžná spojení předložek a následujícího slova a časté slovní páry s nestandardní fonetickou transkripcí.

Slovní spojení s nejvyšší hodnotou vzájemné informace (PMI) a s nejvyšší četností výskytu byla přidána do slovníku. Výsledky jsou uvedeny v tabulce 6.

Přidaných spojení	úspěšnost rozpoznávání (Acc)	
	PMI	četnost výskytu
1000	74.59	75.40
2000	74.55	75.73
3000	74.55	75.95
4000	74.50	76.20
5000	74.64	76.05
6000	74.70	75.89
7000	74.67	75.78
8000	74.68	76.04
9000	74.64	76.28
10000	74.60	76.33

Tabulka 6: Výsledky rozpoznávání se slovními spojeními vybranými na základě vzájemné informace PMI a četnosti výskytu.

Výběr slovních párů pomocí PMI nepřinesl zlepšení v úspěšnosti rozpoznávání. To je způsobeno tím, že kolokace vybrané pomocí PMI nejsou dostatečně četné. Četnost výskytu se ukázala být vhodnějším kritériem pro výběr slovních párů. Přidání 10000 slovních párů zlepšilo úspěšnost téměř neznatelně. Zlepšení není statisticky významné na hladině významnosti 5 %.

Následující experimenty ukazují případy, kdy je slovních spojení přidáno více. Tabulka 7 ukazuje případ, kdy je přidáno více slovních spojení. Stagnace a mírné snižování úspěšnosti rozpoznávání je patrné od 45000 přidaných slovních spojení. Pro 45000 přidaných slov je p-hodnota rovna 4.0e-07 pro zamítnutí hypotézy o stejných výsledcích jako pro slovník s ručně přidanými slovními spojeními.

Přidaných spojení	úspěšnost rozpoznávání (Acc)	
	četnost výskytu	četnost výskytu s předložkou na 1. místě
10000	76.33	76.99
15000	76.82	77.68
20000	77.13	77.57
25000	77.37	77.65
30000	77.43	77.77
35000	77.57	77.88
40000	77.43	77.90
45000	77.69	77.94
50000	77.46	77.91
55000	77.45	77.90

Tabulka 7: Více slovních spojení přidaných na základě četnosti výskytu.

10.3 Vyhodnocení

Experimentální výsledky potvrdily, že přidáním vhodných slovních spojení lze zvýšit úspěšnost rozpoznávání z 74.48 % na 77.94 %, i když jsou spojení přidávána plně automaticky. Je také patrná saturace v počtu přidávání slov, kdy více jak 45000 přidaných slov již nepřispívá ke zvýšení úspěšnosti rozpoznávače.

Výběr pomocí četnosti výskytu slovního spojení v textovém korpusu byl pro uvedenou úlohu vhodnější, neboť PMI nevybírá dostatečně četná slovní spojení. Vzájemná informace může pomoci při ručním výběru takových slovních spojení, kdy je fonetická transkripce spojení běžnými fonologickými pravidly nesprávná, zejména u cizích slov.

Zvýšení úspěšnosti rozpoznávače přidáním slovních spojení má však za následek zvětšení slovníku a jazykového modelu, neboť jsou přidávána spojení tvořená nejčetnějšími slovy. Doba rozpoznávání je kvůli většímu jazykovému modelu delší.

11 Analýza výstupu rozpoznávacího systému

Výsledky rozpoznávání řeči jsou nejčastěji vyjádřeny úspěšností rozpoznávání, nebo mírou chybovosti. Vyhodnocování rozpoznávání spojité řeči je oproti rozpoznávání izolovaných slov složitější v tom, že kromě špatně rozpoznaných slov (substituce s) mohou být některá slova rozpoznávačem ignorována (delece d) a jiná přidána (inzerce i) oproti referenčnímu textu. Porovnávání referenčního textu a rozpoznané řeči se provádí zarovnáváním, které je založeno na dynamickém programování.

V následujících ukázkách jsou uvedeny skutečné chyby nalezené ve výstupu rozpoznávače. Pokud dojde k záměně jediného slova, pak je toto slovo akusticky velmi podobné, například:

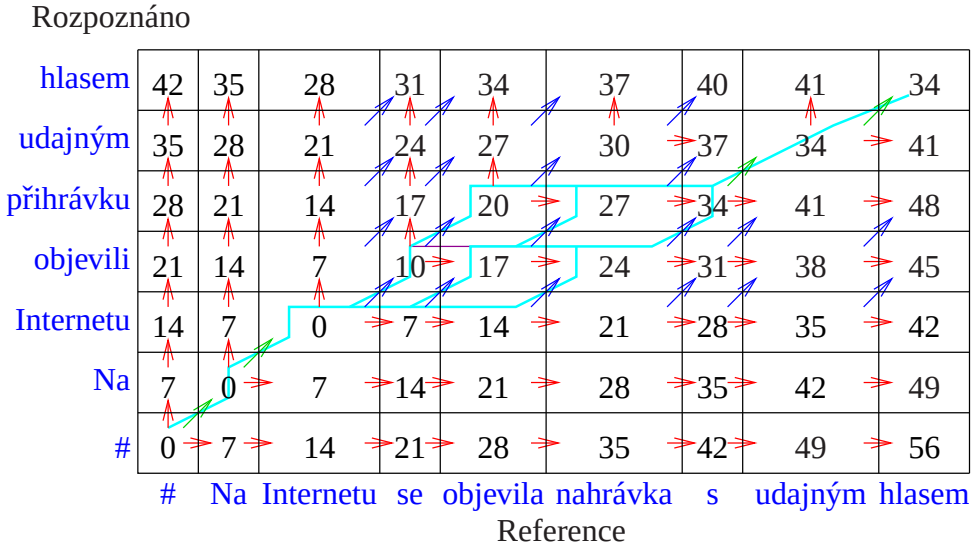
Reference:	Jiří Paroubek **odmítl** návrhy ODS komentovat.
Rozpoznáno:	Jiří Paroubek **odmítnul** návrhy ODS komentovat.

Případ chyby delece–substituce:

Reference:	**My jsme** v odpoledních hodinách zadrželi celkem …
Rozpoznáno:	**Slezsko** v odpoledních hodinách zadrželi celkem …

11.1 Zarovnávání textů

Na obrázku 5 je uveden postup zarovnávání referenční věty „Na Internetu se objevila nahrávka s údajným hlasem.“ a rozpoznané věty „Na Internetu objevili přihrávku údajným hlasem.“ Horizontální a vertikální šipky označují deleci a inzerci. Diagonální šipky označují substituci nebo hit, pokud se slova v příslušném řádku a sloupci shodují. Je též vyznačena nejlevnější cesta s cenou 34.



Obrázek 5: Zarovnávání textů pomocí dynamického programování.

Nejlevnější cesty jsou: $hhddsshh$, $hhssddhh$, $hhdsdshh$, $hhsdsdhh$, $hhdssdhh$, $hhsddshh$. Je vidět, že pokud jsou vedle sebe inzerce a substituce či delece a substituce, pak na pořadí nezáleží, jsou rovnocenné. Úspěšnost rozpoznávání v příkladu je dle NIST [28]

$$Acc = \frac{8 - 0 - 2 - 2}{8} = 50\%.$$

11.2 Detailní analýza

Často je dobré vědět, která slova jsou nejčastěji špatně rozpoznána, aby mohla být cíleně a efektivně zvyšována úspěšnost rozpoznávání.

Více cest se stejnou cenou je v tomto případě nežádoucí, neboť bychom chtěli vědět, které slovo bylo substituováno a které vloženo, či eliminováno. Řešením je upravit ceny inzerce, substituce a delece tak, aby kratší slovo bylo spíše delece nebo inzerce a delší slovo bylo spíše substituce. Úprava cen inzerce, delece a substituce je provedena dle následujících vztahů

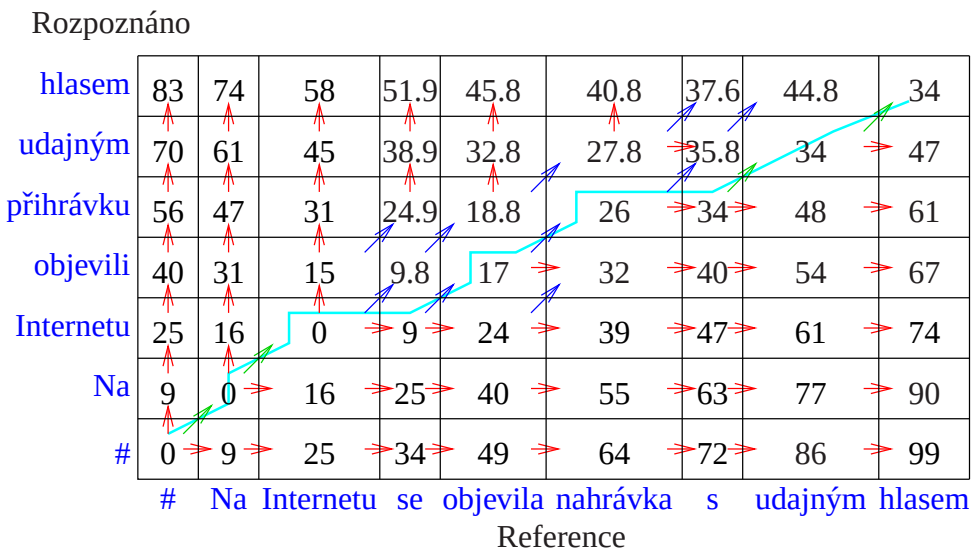
$$c_n(i) = c(i) + l(i), \quad (4)$$

$$c_n(d) = c(d) + l(d), \quad (5)$$

$$c_n(s) = c(s) - \frac{1}{l_d}, \text{ když } l_d > 0,$$
$$= c(s) - 2 \text{ jinak}, \quad (6)$$

kde $c()$ a $c_n()$ jsou původní, respektive nové ceny přechodů, $l()$ je délka slova představující inzerce nebo deleci a l_d rozdíl délek substituovaných slov, $c(i) = c(d) = 7$, $c(s) = 10$.

Upravená tabulka z předchozího příkladu uvedeném v sekci 11.1 je ukázána na obrázku 6.



Obrázek 6: Zarovnávání textů pomocí dynamického programování s eliminací více cest.

Vhodně zvolené ceny přechodů eliminovaly všechny cesty až na jednu. Výsledek je tedy jednoznačný $hhdssdhh$, což odpovídá představě o nejlepším zarovnání delecí a substitucí v tomto příkladu.

11.3 Nejčetnější chyby rozpoznávání

Tato sekce uvádí nejčetnější chyby rozpoznávání spojitého rozpoznávače řeči se slovníkem 312 tisíc slov na databázi TV2005. Rozpoznané promluvy obsahují 275 inzercí, 357 delecí, 1316 substitucí a 9769 slov v referenčním textu, což dává úspěšnost rozpoznávání 80.06%.

Tabulka 8 ukazuje nejčetnější chyby spojitého rozpoznávače řeči. Je patrné, že nejvíce chyb je způsobeno krátkými slovy, které vytvářejí inzerce nebo delece.

Četnost výskytu	chyba
56	inzerce „a“
31	delece „a“
31	delece „je“
30	delece „v“
22	delece „to“
21	delece „i“
14	inzerce „v“
14	inzerce „z“
13	inzerce „i“

Tabulka 8: Nejčetnější chyby spojitého rozpoznávače řeči.

Celkový počet delecí a inzercí slov, která jsou maximálně 2 znaky dlouhá, je 431, což je 68 % všech inzercí a delecí. Kdyby se takto krátké inzerce a delece nevyskytovaly, pak by úspěšnost rozpoznávání stoupla na 82.12%.

V češtině není odlišena výslovnost písmen *i* a *y*. Akusticky je psaní *i* a *y* naznačeno pouze v ryze českých slovech u slabik di, ti, ni. Nejčetnější chyby způsobené nesprávným *i* nebo *y* jsou uvedeny v tabulce 9. V tabulce jsou všechna *i* a *y* nahrazena *i*.

Četnost výskytu	chyba
1	demonstrovali
1	skončili
1	museli
1	milí
1	dostali
1	ti
1	vyhrožovali
1	jezdili
1	pokusili

Tabulka 9: Substituce způsobené y/i.

Nesprávně rozpoznaných *i* a *y* je 20. Takováto chybovost je způsobena převážně jazykovým modelem. Bez těchto substitucí by úspěšnost rozpoznávání byla 80.3 %.

11.4 Zhodnocení

Nově navržená metoda detailní analýzy výsledků lépe přiřadí inzerce, delece a substituce ke konkrétním slovům, čímž je umožněno cílené zlepšování rozpoznávače. Metoda však stále nemusí přiřadit vždy takové typy chyb, jaké bychom očekávali, například sekvenci delece–substituce–inzerce. Proto jsou počty chyb uvedených v tabulkách přibližné počty skutečných chyb.

12 Automatická interpunkce

Většina rozpoznávačů řeči produkuje sekvenci mezerami oddělených slov. Interpunkce vytváří výstup rozpoznávače čitelnější pro čtení. Interpunkce je také důležitá pro další zpracování textu, jako je získávání informací z rozpoznaného textu, strojový překlad, morfologická analýza, atd.

Automatická interpunkce se snaží najít konce vět a vložit do nich tečky a čárky v souvětí. K odhadnutí správné pozice interpunkce v češtině je třeba kombinovat informace z akustické části promluvy, jazykového modelu a morfologické analýzy. Detailní morfologická analýza je však závislá na znalosti pozic interpunkčních znamének. Morfologická analýza může být částečně nahrazena jazykovým modelem. V této práci je použit morfologický analyzátor Jana Hajiče [7].

Z literatury je patrné, že dosavadní systémy provádějící automatickou interpunkci kombinují znalost průběhu základní frekvence (F0), n-gramového jazykového modelu, délky fonémů [29] a případně i morfologických značek [30]. Průběh F0 je po částech linearizován a jsou z něj extrahovány různé příznaky, například sklon lineárních úseku. Článek [30] vychází z [29], je ale zaměřen na češtinu.

Automatická interpunkce je v této práci založena na automaticky nalezených produkčních pravidlech, která jsou naučena pro tečky a čárky zvlášť.

Rozpoznávač řeči používaný v této práci [2] je schopen rozpoznat také některé hluky [15] jako je ticho, nádech, atd. Informace o hlucích je použita místo akustické informace, čímž je umožněna automatická interpunkce výstupu rozpoznávače bez znovupoužití rozpoznávaného signálu. Pozorováním výstupů rozpoznávače bylo zjištěno, že hluky potřebnou akustickou informaci pro účely automatické interpunkce zachovávají.

12.1 Automatické vkládání teček

Pravidla pro vkládání teček jsou odvozena z rozpoznaných šumů (ticho, nádech), které rozpoznávač vkládá do svého výstupu. Tyto šumy jsou označeny čísly 0 až 5 a pomlčkou [15]. Příklad nahrávky s rozpoznaným šumem je uveden na obrázku 7.

Výstup rozpoznávače:	...podle ní nerespektuje soukromí lidí **3** i ministr zahraničí ho vidí jako chybu
Vložená interpunkce:	...podle ní nerespektuje soukromí lidí**.** I ministr zahraničí ho vidí jako chybu.

Obrázek 7: Nahrávka s rozpoznaným šumem

Sekvence šumů indikující interpunkci je hledána gramatickou evolucí [31]. Délka hledaných sekvencí šumu nebyla limitována. Cílem bylo najít takové sekvence, aby přesnost umístění teček byla maximální. Pokud je vložená tečka na místě zarovnaného rozpoznaného přepisu, pak je umístěna přesně. Ostatní umístění teček, či jejich vynechání je považováno za chybu.

Populace gramatické evoluce čítala 500 jedinců. Turnajová selekce byla použita pro výběr rodičů a steady state selekce pro vytváření nové populace. Diverzita populace byla udržována metodou LICE [24]. Učená pravidla jsou produkční pravidla následujícího formátu:

$$\textbf{pokud}\ (\text{sekvence šumů}),\ \textbf{pak}\ \text{napiš tečku místo sekvence} \tag{7}$$

Duplicitní tečky jsou odstraněny po aplikaci pravidel na celý výstup rozpoznávače.

12.2 Automatické vkládání čárek

Pravidla pro vkládání čárek jsou založena na jazykovém modelu a znalosti morfologických kategorií slov, které jsou zjištěny morfologickým analyzátorem [7]. Pravidla jsou odvozena z textového korpusu.

Pravidla pro vkládání čárek jsou automaticky odvozena z textového korpusu. Formát pravidel je:

$$\textbf{pokud}\ (\text{sekvence slov}),\ \textbf{pak}\ \text{napiš čárku před sekvenci} \tag{8}$$

Kvůli velkému množství různých slov a ještě většímu množství slovních spojení je kompletní prohledávání sekvencí slov téměř nemožné. Proto jsou sekvence slov omezeny pouze na obvyklá spojovací slova, spojky, zájmena, příslovce a

předložky, což je založeno na pozorování korpusu. Slovní druh je určen morfologickým analyzátorem. Protože rozpoznávač pracuje s omezeným slovníkem, je možné slovní druhy určit jednou pro slova ve slovníku, jiná slova se ve výstupu rozpoznávače nemohou objevit.

12.3 Aplikace pravidel

Dva typy pravidel pro vkládání teček a čárek byly naučeny odděleně. Je proto nutné vyřešit konflikty, kdy může být aplikováno více pravidel najednou.

Jako první jsou aplikována pravidla vkládající tečky. Duplicitní tečky a zbylé šumy jsou odstraněny. Text již obsahuje jen tečky a slova.

12.4 Experimenty

Experimenty byly prováděny na testovací části akustických dat. Před vyhodnocováním byly rozpoznané promluvy zarovnány s referenčními přepisy. Následně byla provedena automatická interpunkce výstupu rozpoznávače. Výstupy automatické interpunkce a zarovnávání byly porovnávány.

Výsledky jsou uvedeny ve 4 mírách: úspěšnosti inzerce (Acc), precision (P), recall (R) a F-measure (F) definované [32]:

$$F = \frac{2RP}{R+P}. \tag{9}$$

V prvním experimentu nejsou aplikována žádná pravidla pro vkládání interpunkce, a proto slouží jako baseline. Výsledky jsou uvedeny v tabulce 10.

Úspěšnost (Acc)	Precision (P)	Recall (R)	F-measure (F)
88.27 %	100.00 %	88.27 %	93.77 %

Tabulka 10: Žádná interpunkce není vložena, baseline

V následujícím experimentu jsou pravidla pro vkládání teček a čárek aplikována odděleně vždy na čistý výstup rozpoznávače. Tabulka 11 ukazuje výsledky pro tento případ.

Typ pravidel	Úspěšnost (Acc)	Precision (P)	Recall (R)	F-measure (F)
Pouze tečky	90.10 %	75.06 %	90.10 %	81.90 %
Pouze čárky	90.35 %	77.17 %	90.35 %	83.24 %
Tečky i čárky	92.05%	77.05%	92.05%	83.88%

Tabulka 11: Tečky a čárky jsou aplikovány samostatně

12.5 Zhodnocení

Uvedené experimenty ukazují významné zlepšení oproti ponechání výstupu rozpoznávače jako mezerami oddělený proud slov. Byla dosažena úspěšnost vkládání interpunkce 92.05%. Čitelnost výstupu rozpoznávače může být dále vylepšena kapitalizací písmen po tečkách.

13 Závěr

Tvorba lingvistické vrstvy systému automatického rozpoznávání mluvené češtiny je v této práci pojata jako komplexní problém. V průběhu práce bylo vytvořeno množství programů, postupů a vylepšení umožňující automatizaci adaptace slovníku a jazykového modelu.

V disertační práci jsou diskutovány problémy týkající se různých zdrojů textových dat, jejich získávání a čištění. Byly vytvořeny robustní autonomní programy schopné získávat data 24 hodin denně 356 dní v roce. Zároveň byly uvedeny postupy normalizace textu jak pro všeobecné texty z novinových zpráv, tak i pro speciální lékařské texty, kterých je většinou málo a obsahují mnoho chyb a cizích slov. Jsou uvedeny metody identifikace cizích slov umožňující aplikaci správných fonologických pravidel. Byl experimentálně prokázán pozitivní vliv různých normalizací na úspěšnost rozpoznávání. Během práce byl zvětšen textový korpus o více než 100 %.

Další část díla je zaměřena na slovník a fonetickou transkripci. Je uvedena závislost pokrytí českého textového korpusu na velikosti slovníku. Dále je uvedena vlastní metoda vylepšení fonetické transkripce spočívající v natrénování nových fonologických pravidel, která jsou následně přidána k existujícím fonologickým pravidlům. Nová pravidla jsou natrénována pomocí Gramatické evoluce. Výhodou uvedené metody je, že neobjevuje již známá pravidla. Nová naučená pravidla jsou ihned připravena k aplikaci. Poslední část kapitoly týkající se slovníku se zabývá přidáváním slovních spojení do slovníku. Jde o jednoduchý a téměř bezpracný způsob zvýšení úspěšnosti rozpoznávání. Slovní spojení jsou vybírána na základě vhodné míry. Vhodnost různých měr je experimentálně ověřena. Úspěšnost rozpoznávání byla touto metodou zvýšena z 74.48 % na 77.94 %.

V kapitole zabývající se jazykovým modelem jsou diskutovány otázky efektivní implementace výpočtu jazykového modelu s velkým slovníkem tak, aby jej bylo možné spočítat na běžně dostupných počítačích v přijatelném čase. Jsou uvedeny vlastní implementace výrazně zrychlující výpočet jazykového modelu oproti dosavadnímu programu používaném v Laboratoři počítačového zpracování řeči. Stejná kapitola uvádí výsledky experimentů zjišťujících vliv velikosti slovníku a interpunkce na úspěšnost rozpoznávání. Je též uveden průběh nalézání nových

bigramů při výpočtu jazykového modelu, ze kterého je patrné, že pro slovník obsahující 312 tisíc slov stále existuje množství bigramů, jejichž hodnota je odhadnuta nepřesně pro nedostatek dat. K přesnějšímu odhadu málo četných bigramů je však potřeba velké množství dat. Je proto nutné sbírat další texty do textového korpusu.

Další kapitola se zabývá detailní analýzou výsledků rozpoznávání. Abychom mohli efektivně zlepšovat rozpoznávač, je nutné vědět, které chyby jsou při rozpoznávání nejčetnější. V kapitole je uvedena vlastní modifikace běžně používané metody vyhodnocování výsledků. Pomocí uvedené modifikace je možné přesněji určit, která slova jsou rozpoznávačem vložena, vypuštěna, či zaměněna za jiné. Jsou zde též uvedeny a kvantizovány nejčetnější chyby rozpoznávačů a chyby vzniklé díky příčestí minulému a chyby psaní „y“ a „i“.

V kapitole zabývající se adaptací jazykového modelu jsou provedeny experimenty týkající se tématické a časové adaptace jazykového modelu. Především experimentů ukazujících vliv přidávání nových textů na úspěšnost rozpoznávání je v literatuře velmi málo. Ve většině publikací je uveden pouze vliv adaptace na perplexitu která, jak se v literatuře ukazuje, má malý vztah ke skutečné úspěšnosti rozpoznávání. Z provedených experimentů vyplývá, že k údržbě kvalitního jazykového modelu není třeba častých aktualizací. Občasné přidání aktuálních dat se pozitivně projeví na úspěšnosti rozpoznávání. Je též zřejmá nutnost přidávání nových slov do slovníku, aby tato slova mohla být rozpoznávána.

Poslední kapitola se zabývá úpravou textového výstupu z rozpoznávače s cílem zvýšit čitelnost tohoto výstupu. V kapitole je uvedena vlastní modifikace existujících metod automatického vkládání interpunkce. Publikovaná metoda je oproti ostatním metodám schopna odvodit pozice interpunkce pouze z výstupu rozpoznávače, a to díky informacím o různých šumech, které výstup rozpoznávače obsahuje. Byla dosažena 92.05% úspěšnost automatické interpunkce.

Většina vytvořených programů je aktivně používána jak v Laboratoři počítačového zpracování řeči, tak je i součástí komplexních komerčních produktů Laboratoře počítačového zpracování řeči.

Reference

[1] Jan Nouza, Tomáš Nouza, and Petr Červa. A multi-functional voice-control aid for disabled persons. In *Proceedings of the SPECOM 2005*, Patras, Greece, 2005.

[2] Jan Nouza, Jindřich Žďánský, Petr David, Petr Červa, Jan Kolorenč, and Dana Nejedlová. Fully automated system for czech spoken broadcast transcription with very large (300k+) lexicon. In *Proceedings of the Interspeech 2005*, Lisbon, Portugal, 2005.

[3] Xuedong Huang and Alex Acero Hsiao-Wuen Hon. *Spoken Language Processing: a guide to theory, algorithm, and system development*. Prentice Hall PTR, Upper Saddle River, New Jersey 07458, 2001. ISBN 0-13-022616-5.

[4] Jan Nouza and Tomáš Nouza. A voice dictation system for a million-word czech vocabulary. In *Proceedings of the of ICCCT 2004*, ISBN 980-6560-17-5, pages 149–152, Austin, USA, 8 2004.

[5] Jan Nouza, Dana Nejedlova, Jindrich Zdansky, and Jan Kolorenc. Very large vocabulary speech recognition system for automatic transcription of czech broadcast programs. In *Proceedings of the ICSLP 2004*, October 2004.

[6] Jan Kolorenč and Tomáš Klimovič. Cardiology language model for voice dictation. In *Proceedings of the 14th Czech-German Workshop*, pages 93–97, Prague, September 2004. ISBN 80-86269-11-6.

[7] Jan Hajič, Eva Hajičová, Petr Pajas, Jarmila Panevová, Petr Sgall, and Barbora Vidova Hladká. Prague dependency treebank. CDROM LDC2001T10, Linguistic Data Consortium, University of Pennsylvania, 2001.

[8] Sean M. Burke. *Perl & LWP*. O'Reilly, 2002. ISBN 0-596-00178-9.

[9] Dana Nejedlová, Jindra Drábková, Jan Kolorenč, and Jan Nouza. Lexical, phonetic, and grammatical aspects of very-large-vocabulary continuous speech recognition of czech language. In *Proceedings of the Electronic Speech Signal Processing 2005*, pages 224–231, Prague, Czech Republic, September 2005. ISBN 3-938863-17-X.

[10] Jean-Luc Gauvain, Lori Lamel, Gilles Adda, and Mich'ele Jardino. The limsi 1998 hub-4e transcription system. In *Proceedings of the DARPA Broadcast News Workshop*, Herndon, VA, 1999.

[11] Gerhard Backfried and Roser Jaquemot Caldes. Spanish broadcast news transcription. In *Proceedings of the EUROSPEECH-2003*, pages 1561–1564, 2003.

[12] Zdena Pálková. *Fonetika a fonologie češtiny*. Karolinum, Praha, 2 edition, 1997.

[13] International Phonetic Association. Report on the 1989 kiel convention. *Journal of the Phonetic Association*, 19(12), 1989.

[14] Jan Nouza, Josef Psutka, and Jan Uhlíř. Phonetic alphabet for speech recognition of czech. *Radio Engineering*, 6(4):16–20, December 1997.

[15] Jindřich Žďánský and Martin Kroul. Semi-automatic non-speech events database formation. In *Proceedings of the 8th International Student Conference on Electrical Engineering - POSTER 2004*, May 2004.

[16] Marek Volejník. Fonetická transkripce psané a mluvené češtiny pro účely automatického zpracování řeči. Master's thesis, Technická univerzita v Liberci, Fakulta mechatroniky a mezioborových inženýrských studií, 1999.

[17] Johnson and Mark. A discovery procedure for certain phonological rules. In *Proceedings of the Tenth International Conference on Computational Linguistic*, pages 334–347. Stanford, 1984.

[18] Rilley and D. Michael. A statistical model for generating pronounciation networks. In *Proceedings of the IEEE ICASSP-91*, pages 737–740, 1991.

[19] Terrence J. Sejnovski and Charles R. Rosenberg. Parallel networks that learn to read aloud. In *Cognitive Science*, volume 1598, pages 179–211, 1986.

[20] Daniel Gildea and Daniel Jurafsky. Automatic induction of finite state transducers for simple phonological rules. In *Proceedings of the Meeting of the Association for Computational Linguistics*, pages 9–15, 1995.

[21] Dana Nejedlová. Fonetická transkripce češtiny pomocí třívrstvé neuronové sítě. Technical report, Technická univerzita v Liberci, Laboratoř zpracování řeči, Liberec, 200.

[22] Jan Kolorenč. Evolving phonological rules using grammatical evolution. In *Proceedings of the 8th International Student Conference on Electrical Engineering–POSTER 2004*, Prague, 5 2004. [CD-ROM].

[23] Vladimír Mařík, Olga Štěpánková, Jiří Lažanský, and kolektiv. *Umělá inteligence 3*. Academia. ISBN 8020004726, EAN 9788020004727.

[24] Jan Kolorenč. Získávání znalostí z dat pomocí gramatické evoluce. Master's thesis, České vysoké učení technické v Praze, Fakulta elektrotechnická, 2004.

[25] Wayne Ward, Holly Krech, Xiuyang Yu, Keith Herold, George Figgs, Ayako Ikeno, Dan Jurafsky, and William Byrne. Lexicon adaptation for lvcsr: speaker idiosyncracies, non-native speakers, and pronunciation choice. In *Proceedings of the Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology (PMLA)*, pages 83–88, 2002.

[26] Andre Berton, Pablo Fetter, and Peter Regel-Brietzmann. Compound words in large-vocabulary german speech recognition systems. In *Proceedings of the ICSLP 96*, 1996.

[27] Roeland Ordelman, Arjan van Hessen, and Franciska de Jong. Compound decomposition in dutch large vocabulary speech recognition. In *Proceedings of the Eurospeech 2003*, September 2003.

[28] NIST. Matched pairs sentence-segment word error (mapsswe) test. online⟨http://www.nist.gov/speech/tests/sigtests/mapsswe.htm⟩.

[29] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gökhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000.

[30] Jáchym Kolář, Josef Psutka, and Jan Švec. Automatic punctuation annotation in czech broadcast news speech. In *Proceedins of 9-th International Conference Speech and Computer (SPECOM 2004)*, St. Petersburg, Russia, 2004.

[31] Conor Ryan, J. J. Collins, and Michael O' Neill. Grammatical evolution: Evolving programs for an arbitrary language. In *Proceedings of the First European Workshop on Genetic Programming*, volume 1391, pages 83–95, Paris, 14-15 1998. Springer-Verlag.

[32] Information retrieval. Wikipedia. online ⟨http://en.wikipedia.org/wiki/Information_retrieval⟩.