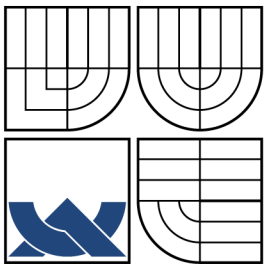
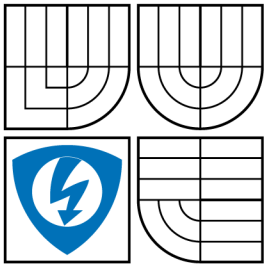


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND
COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

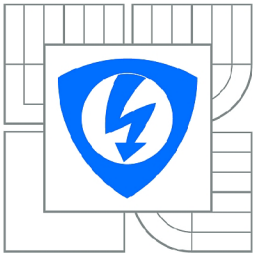
FONETICKÁ TRANSKRIPCE ČESKÉHO JAZYKA
PHONETIC TRANSCRIPTION OF CZECH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — MARTIN ZEDEK
AUTHOR

VEDOUCÍ PRÁCE — Ing. PETR SYSEL, Ph.D.
SUPERVISOR

BRNO 2014



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Bakalářská práce

bakalářský studijní obor
Teleinformatika

Student:	Martin Zedek	***ID:***	107702
Ročník:	3	***Akademický rok:***	2013/2014

NÁZEV TÉMATU:

Fonetická transkripce českého jazyka

POKYNY PRO VYPRACOVÁNÍ:

Nastudujte používané metody zápisu zvukové podoby jazyka pomocí zvolených sad znaků. Zaměřte se především na specifika českého jazyka a pravidla jeho přepisu. Vytvořte v prostředí Matlab nebo v jazyce C program pro automatický přepis psaného textu do fonetického zápisu. Implementujte i případné asimilace znělosti a podobné výjimky. Dále program upravte tak, aby umožňoval i zpětný převod z fonetického přepisu do psaného textu. Ve sporných případech se pokuste použít algoritmy na kontrolu pravopisu. Funkci programu ověřte na několika testovacích promluvách.

DOPORUČENÁ LITERATURA:

[1] Psutka, J.; Müller, L.; Matoušek, J.; Radová, V. Mluvíme s počítačem česky. 1. vydání. Praha: Academia, 2006. 752 s. ISBN 80-200-1309-1
[2] Deller, J. R.; Hansen, J. H. L.; Proakis, J. G. Discrete-Time Processing of Speech Signals. New York: IEEE Press, 2000. ISBN 0-7803-5386-2

Termín zadání:	10.2.2014	***Termín odevzdání:***	4.6.2014

Vedoucí práce: Ing. Petr Sysel, Ph.D.
Konzultanti bakalářské práce:

doc. Ing. Jiří Mišurec, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Cílem práce je vytvoření skriptu pro automatický přepis českého jazyka do fonetické podoby a zpět. V práci jsou porovnány jednotlivé fonetické abecedy a popsány jejich výhody a nevýhody. Pro tuto práci byla nakonec zvolena česká fonetická abeceda (ČFA). Následně jsou uvedeny nejdůležitější pravidla pro spojení samohlásek a souhlásek a dále hlavní asimilační pravidla. Skript byl vytvořen v prostředí MATLAB. Funkční je převod do fonetické formy s využitím popsaných pravidel. Zpětný převod není plně odladěn a je nutno výsledek zpřesnit užitím programu ASPELL pro korekci pravopisu.

KLÍČOVÁ SLOVA

matlab, skript, fonetika, asimilace, český jazyk, transkripce do ČFA, transkripce z ČFA, aspell

ABSTRACT

The aim is to create a script for automatic transcription of Czech language phonetic forward and backward. The thesis compares the different phonetic alphabet and describes their advantages and disadvantages. For this thesis was eventually selected Czech Phonetic Alphabet (ČFA). The following are the most important rules for connection of vowels and consonants and the main assimilation rules. The script was created in MATLAB environment. The function is to convert the phonetic form using the rules described. Backward conversion is not fully debugged and it is necessary to refine the result using Aspell program for correcting spelling.

KEYWORDS

matlab, script, phonetics, assimilation, Czech language, transcription into ČFA, transcription out of the ČFA, aspell

ZEDEK, Martin *Fonetická transkripce českého jazyka*: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2014. 36 s. Vedoucí práce byl Ing. Petr Sysel, Ph.D.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Petru Syslovi, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Brno ....

(podpis autora)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic
http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této bakalářské prďż"ci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno ....
(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

ÚVOD

Téma této práce jsem si vybral především proto, že mluvená řeč je základní a nejpřirozenější způsob komunikace mezi lidmi a je důležité zporstředkovat tuto komunikaci za pomoci moderní technologie, což může výrazně usnadnit život. V současné době je však nemožné plnohodnotně komunikovat s počítačem bez jakýchkoliv omezení, protože existuje mnoho problémů, které ještě stále čekají na svá řešení.

V první části se budu okrajově zabývat historií řeči a jejího studia z hlediska fonetiky a fonologie. Dále se zaměřím na vznik fonetických abeced, na jejich jednotlivé druhy, význam a užití. Především se zaměřím na mezinárodní abecedy IPA a SAMPA, dále pak vysvětlím základy české fonetické abecedy a správného přepisu výslovnosti. V závěru kapitoly připomenu důležitost českých řečových jednotek (samohlásky a souhlásky), jejich charakteristiky a dělení.

V další části se budu věnovat pravidlům fonetické transkripce češtiny a základním vlastnostem její automatické transkripce. Vyjmenuji a popíši symboly užívané v pravidlech fonetické transkripce. Dále proberu způsoby spojování samohlásek v češtině, spojení samohlásek se souhláskami a výslovnost jednotlivých skupin hlásek. Poté popíši základní asimilační znělosti, včetně jejich vyjímek, asimilační artikulaci, slabikotvorné souhlásky a zjednodušení některých souhláskových skupin.

V poslední části popíši mnou vytvořený program pro přepis textu do fonetického přepisu, včetně různých podpodmínek popsaných ve druhé kapitole. Dále bude popsán program pro zpětný převod z fonetického znění do běžného textu. Pro přehlednost připojím i pár vývojových diagramů.

Cílem této práce je shrnutí pravidel správné výslovnosti českého jazyka a jeho zápisu do fonetického tvaru pro práci na počítači a vytvoření programu pro obousměrný přepis textu mezi běžným zobrazením a fonetickým zápisem.

1 ZÁKLADNÍ ŘEČOVÉ JEDNOTKY A FONETICKÉ ABECEDY

Jazyk přestavuje velice složitý systém komunikace. Je vlastní pouze lidským bytostem, a přitom každá kultura má jiné jazykové kategorie. Existuje ve dvou základních podobách, v podobě mluvené a psané. Účelem jazykového vyjadřování je především sdělovat myšlenky a zprostředkovávat komunikaci mezi lidmi. Mluvená forma jazyka vznikla daleko dříve a je také přirozenou a nejčastěji užívanou formou komunikace. Výhodou mluvené řeči je, že se všichni účastníci komunikace mohou věnovat i jiným činnostem. Nevýhodou však je pomíjivost informace. Oproti tomu psaná forma umožňuje uchovávat myšlenky a informace na delší dobu, což je její hlavní výhodou. Není však tak osobní, jako forma mluvená, a navíc vyžaduje schopnost umět číst a psát.

Je velice důležité dokázat zaznamenat mluvenou podobu jazyka psanou formou tak, aby co nejlépe vyjadřovala zvukovou podobu mluvy. V dnešní moderní době je kladen důraz na zaznamenávání informací v elektronické podobě, což znamená správně zaznamenávat i fonetickou transkripci jazyka a výslovnost. Tato práce se bude věnovat fonetické transkripci češtiny a české výslovnosti.

1.1 Základní řečové jednotky

Řeč je poslopnost zvuků, které vznikají v artikulačním ústrojí člověka. Způsobem jakým se jednotlivé zvuky vytváří, se zabývá fonetika. Ta studuje celkový proces vytváření řeči, který začíná v dechovém ústrojí, dále pokračuje v hlasovém ústrojí a nakonec se vytvoří základní hlasivkový tón v artikulačním ústrojí. Soubor takto vytvořených zvuků se nazývá hláska, která je považována za základní řečovou jednotku. Posloupnost hlásek vyslovovaných člověkem je definována jako řeč. Fonetický zápis řeči se zapisuje do hranatých závorek [ ]. Jednotlivé hlásky se na fonetické úrovni zkoumají nezávisle na konkrétním jazyku, fonetika se nezabývá významem vytvářené řeči.[3]

Zatímco fonetika se věnuje zkoumání řečových zvuků, fonologie tyto zvuky zkoumá z hlediska systémové stavby daného konkrétního jazyka a vytváří lingvistické jednotky zvané fonémy. Foném a hláska jsou tedy naprosto odlišné pojmy. Foném je nejmenší lingvistická jednotka, schopná rozlišovat významové jednotky, např. slova. Neexistuje nezávisle, vytváří množiny jednotek, které bývají definovány na principu tzv. minimálních párových slov (tvoří dvojice slov, které se liší pouze jedním jediným fonémem). Jeden foném může být vyjadřován více hláskami, které mohou změnit význam slova, třeba hlásky [a] a [á] v češtině jsou dva fonémy /a/

a /á/, které udělají rozdíl ve významu mezi slovy *past* a *pást* nebo [z] a [ž] ve slovech *zrát* a *žrát*, ale například hlásky [m] a [ɱ] jsou v češtině jeden foném /m/ protože, když zaměníme jeden za druhý tak se nezmění význam slova, například tramvaj. Různé jazyky používají různé fonémy tzn. inventáře fonémů. Ty fonémy, které jsou nejvíce využívány tvoří centrum inventáře (v češtině tvoří toto centrum inventáře většina samohlásek). Některé fonémy jsou užívány minimálně, v češtině jde o /g/ a /ó/. Fonologický záznam řeči se zapisuje mezi lomítka //.

1.2 Fonetické inventáře a abecedy

Souhrn všech odlišných fonetických jednotek jazyka je nazýván fonetický inventář. jsou v něm zahrnuty všechny jednotlivé zvuky řeči, které jsou dále děleny. Jednotlivé inventáře se od sebe liší velikostí a podrobností. Jednotky každého inventáře se popisují pomocí různých fonetických symbolů, které tvoří fonetické abecedy a díky kterým je možné zachytit a zapsat řeč. Přepis výslovnostní formy řeči pomocí fonetické abecedy se nazývá fonetická transkripce.

1.2.1 Mezinárodní fonetická abeceda IPA

Za mezinárodní standart pro fonetický zápis ve všech jazycích je považována mezinárodní fonetická abeceda IPA (International Phonetic Alphabet). Tento systém vytvořili britští a francouzští učitelé jazyků pod záštitou Mezinárodní fonetické asociace, ta byla založena roku 1886 v Paříži[1]. Tvoří ji fonetické znaky, které nejsou závislé na jazyce a jejich výslovnost je možno zapisovat jednotným způsobem. Různé jazyky mají rozdílný počet fonémů, tudíž nevyužijí všechny symboly abecedy IPA. Navíc pro popis některých jazyků se používají i jiné abecedy (zejména pro slovanské jazyky). Mluvená řeč má různou délku a intenzitu hlásek, rytmus, melodii a hlasitost promluvy. Prozatím však pro tyto charakteristické znaky neexistuje žádná komplexní mezinárodní abeceda a obecně se tedy používá pouze omezené množství doplňujících symbolů. IPA odlišuje trvání hlásek pomocí čtyř úrovní délek samohlásek (příklad uveden pro písmeno [a]): velmi krátké [ă], krátké [a], středně dlouhé [aˑ] a dlouhé [aː]. Dále úrovně výšky hlasu na velmi vysokou, vysokou, střední, nízkou a velmi nízkou, také zavádí symboly zvýšení [↑] nebo snížení [↓] výšky hlasu. Mimo jiné lze také označit melodii promluvy. Dále zavádí značky pro přízvuk, což znamená význačné postavení určité slabiky nebo určitého úseku, který obsahuje jeden hlavní přízvuk. Rozlišuje dva druhy přízvuků, hlavní [ˈ] a vedlejší [ˌ]. V neposlední řadě IPA obsahuje značky pro segmentaci promluvy na kratší [|] a větší [‖] úseky.

1.2.2 SAMPA

Symboly mezinárodní fonetické abecedy IPA se velice složitě zapisují do počítače, proto bylo nutné vytvořit jednodušší systém zadávání fonetických symbolů, který sloužil jako základ pro vznik nové fonetické abecedy SAMPA (Speech Assessment Methods Phonetic Alphabet) v roce 1989. Ta je velmi snadno zobrazitelná v počítači a provádí kódování symbolů IPA na 7-bitové tisknutelné znaky ASCII. Zápis v této abecedě musí být především jednoznačný a srozumitelný. nejprve vznikla pro fonetický přepis šesti evropských jazyků (angličtinu, němčinu, dánštinu, holandštinu, francouzštinu a italštinu), poté přibyla norština a švédština (roku 1992). V rozmezí let 1993-1996 byly přidány jihoevropské a severské jazyky[2]. Pro ostatní jazyky byla vydána doporučení, aby veškeré znaky abecedy IPA shodující se s malými písmeny latinky, byly v abecedě SAMPA totožné a zbylé byly kódovány znaky ASCII 33-126. Tato doporučení ale není nutno striktně dodržovat. Při dodržení daných doporučení by se české krátké samohlásky abecedou SAMPA zapsaly /a/, /E/, /I/, /O/, /U/. Pro zpřehlednění je lepší je psát jako /a/, /e/, /i/, /o/, /u/. Bohužel díky takovýmto odchylkám v různých jazycích, dochází ke ztrátě nandnárodního charakteru a každá verze je spjata s určitým jazykem. Roku 2006 byla SAMPA deklarována pro dvě desítky evropských jazyků (mezi nimiž byla i čeština) a pár dalších (např. kantonština, hebrejština, thajština). Roku 1995 profesor Londýnské univerzity John C. Wells vymyslel rozšíření SAMPA abecedy. Tato nová verze byla označena jako X-SAMPA (eXtended SAMPA). Jejím cílem je umožnit kódování veškerých znaků IPA abecedy na lehce tisknutelné ASCII znaky, díky čemuž by X-SAMPA dovolila zapsat fonetickou reprezentaci jakéhokoliv známého jazyku. Nejvíce používané symboly jsou shodné se symboly abecedy SAMPA a zapisují se jedním znakem. Méně používané symboly jsou zapsány jedním znakem, doplněným zpětným lomítkem. Diakritika, jež není obsažena v abecedě SAMPA, se značí podtržítkem za symbolem k němuž se vztahuje a je následován znakem určujícím druh diakritiky.

1.2.3 Ostatní fonetická abecedy

Existují i jiné abecedy, než výše popsané. Většinou se jedná o abecedy přizpůsobeny konkrétnímu jazyku, tak aby byl fonetický zápis co nejlépe a nejjednodušeji čitelný. Jednotlivé symboly se mohou zapisovat jedním, či více znaky. Pokud se jedná o více znaků, je nutno jednotlivé fonetické symboly oddělovat mezerou. Mezi tyto abecedy patří například abeceda KLATTBET pro americkou angličtinu, jež se využívá pro syntézu řeči, nebo také abeceda MRPA (Machine Readable Phonetic Alphabet) pro britskou angličtinu. Nejzajímavější je abeceda WORLDBET, která stejně jako abeceda SAMPA, obsahuje veškeré znaky abecedy IPA, reprezentované znaky

ASCII. S pomocí abecedy WORLDBET je tudíž možné zapsat všechny řečové zvuky různých světových jazyků.

1.2.4 České fonetické abecedy

Český jazyk lze zamozřejmě interpretovat pomocí standartních mezinárodních abeced IPA a SAMPA. Z důvodu složité zobrazitelnosti se tudíž pro zpracovaní české mluvy nehodí. Mezinárodní fonetickou abecedu dokonce většinou nepoužívají ani čeští fonetici, protože tato abeceda jako taková není příliš vhodná pro popis slovanských jazyků [4]. Oficiální SAMPA pro češtinu, umožňující její přepis do vhodné formy pro strojové zpracování, byla uvedena teprve před pár lety. Česká fonetická abeceda (ČFA) pro vyjádření hlásek užívá jeden a více znaků. Každá takto reprezentovaná hláska je vždy oddělena mezerou. Díky tomu je tato abeceda problematicky čitelná. Z toho důvodu se zavedla tzv. zjednodušená česká fonetická abeceda (ZČFA), která pro větší přehlednost obsahuje i českou diakritiku. Proto je pak fonetický přepis téměř totožný jako původní věta, a proto je bez větších problému lehce čitelný. Prvky, jež nelze zapsat pomocí znaků české abecedy, se většinou nahradí znaky převzatými z abecedy IPA, tudíž opět dochází k problému s jejím použitím na počítači. Pro srovnání jsou v tab. 1.1 uvedeny rozdílné zápisy češtiny v abecedách ZFČA, IPA, SAMPA a ČFA.

Abecedu českých fonetických jednotek lze posuzovat dle alofonické realizace fonémů. „Hrubší"abeceda je tvořena pouze fonémy, „jemnější"varianta k nim přidává i alofony (malé změny výslovnosti fonému). Tyto změny výslovnosti neovlivňují smysl vět, ale pouze mění znělost některých hlásek. Toho se užívá především pro zpřesnění syntézy řeči. Dále uvedenu některé významné alofonické fonémy, vyskytující se v českém jazyce:

1. **Zadopatrové (měkkopatrové) *n* [ŋ].** Alofon fonému /n/ se vyslovuje ve spojeních *nk* a *ng* (např. *slánka* nebo *mango*).
2. **Retozubné *m* [ɱ].** Alofon fonému /m/ se vyslovuje ve spojeních *mv* a *mf* (např. *tramvaj* nebo *nymfa*).
3. **Neznělé *ř* [r̊].** Varianty fonému /ř/ vznikají v případě, že v sousedství *ř* je neznělá hláska nebo se *ř* nachází před pauzou a *mf* (např. *břich* nebo *zvěř*).
4. **Slabikotvorné *r* [r̩].** Varianty fonému /r/ vznikají, když se v okolí *r* nachází souhláska nebo se nachází na konci slova po souhlásce (např. *mrkev* nebo *lotr*).
5. **Slabikotvorné *l* [l̩].** Podobně jako u fonému [r̩] vzniká, když se v okolí *l* nachází souhláska nebo se nachází na konci slova po souhlásce (např. *vlk* nebo *sedl*).
6. **Slabikotvorné *m* [m̩].** Tento alofon se vyskytuje zřídka a opět vzniká, pokud se v okolí *m* nachází souhláska nebo se *m* nachází na konci slova po souhlásce

(např. *Žamberk* nebo *osm*).

7. **Znělé *ch*** [ɣ]**.** Vzniká spodobou znělosti před znělou párovou souhláskou. V českém jazyce se tak často nenachází. Může být nahrazen fonémem /h/. Příkladem znělého *ch* je *kdybych vzal* a může být nahrazeno fonémem /h/, tzn. /kdybyh vzal/.
8. **Ráz** [ʔ]**.** Vzniká, když se hlasivky uvedou velmi prudce do pohybu, tzn. při tvrdém hlasovém začátku. V českém jazyce se objevuje při spisovné výslovnosti před samohláskou na začátku slova (např. [ʔanakonda], dále se užívá uvnitř fonetického slova na morfologickém švu (např. [náʔústek] nebo [vʔoku]).

1.3 České řečové jednotky

V následujících podkapitolách jsou podrobněji popsány české řečové jednotky, které se dělí do dvou velkých skupin. První skupinou jsou samohlásky, nebo-li vokály, Druhá velká skupina se nazývá souhlásky, nebo-li konsonanty. Znalost akustických a artikulačních vlastností řečových jednotek se ve značné míře využívá v počítačové syntéze řeči. Při popisu české mluvené řeči se využívají jak fonologický, tak i fonetický popis mluvy. Fonetický popis (tzn. hlásky) se však využívá pouze pro zdůraznění některých speciálních vlastností jednotlivých variant fonémů.

1.3.1 Samohlásky

Česká řeč obsahuje celkem 10 samohlásek, které se však dají rozdělit na krátké (/a/, /e/, /i/, /o/, /u/) a dlouhé (/á/, /é/, /í/, /ó/, /ú/) (y, ý považujeme jako i, í a ů je stejné jako ú)[3]. Jejich artikulace je shodná, avšak liší se především trváním. Samohlásky jsou v mluvené řeči velice důležité protože tvoří jádro slabiky a také jsou nositelkami tónové kvality a formulují estetické vyznění mluvené řeči.

Důležitou skupinou jsou dvojhlásky, nebo-li diftongy. V českém jazyce se však objevují pouze u slov přejatých (/a͡u/, /e͡u/), jedinou vyjímkou je česká dvojhláska /o͡u/. Názory, zda považovat dvojhlásky za jeden foném, nebo za spojení dvou fonémů, se rozcházejí. Proto se při fonetickém popisu jazyka používají obě varianty. V některých jazycích existuje redukovaná (neutrální, neurčitá) samohláska (šva) [ə], která je považována za střední středovou samohlásku. V českém jazyce se však objevuje vyjímečně a nepovažuje se za samostatný foném. Lze ji nalézt pří hláskování souhláskových písmen b [bə], t [tə], d [də], nebo při nedbalé výslovnosti samohlásek a při nespisovném hláskování některých zkratek (např. PNG [pənəgə], které se má správně vyslovit [péengé]).

1.3.2 Souhlásky

Druhá významná skupina českých řečových jednotek jsou souhlásky, které se v českém jazyce dělí na 27 souhláskových fonémů: /b/, /c/, /č/, /d/, /ď/, /f/, /g/, /h/, /ch/, /j/, /k/, /l/, /m/, /n/, /ň/, /p/, /r/, /ř/, /s/, /š/, /t/, /ť/, /v/, /z/, /ž/, /dz/, /dž/ a několik významných alofonů. Narozdíl od samohlásek, které jsou důležité pro estetické vyznění řeči, správná a přesná výslovnost souhlásek je základem pro srozumitelnou řeč. Akustické signály souhlásek mají naprosto odlišný charakter, než akustické signály samohlásek, proto lze poměrně jednoduše rozeznat shluky souhlásek od samohlásek. Někdy je však obtížné rozlišit od sebe navzájem jednotlivé souhlásky.

1.3.3 Slabiky

Základními zvukovými jednotkami při popisu mluveného jazyka jsou slabiky. Díky svým výhodným vlastnostem jsou významné především v úloze počítačové syntézy řeči. Seskupují se do nich fonémy a hlásky a jejich délka může být různá (od jednoho fonému do více než pěti). Nejčastěji se v českém jazyce využívají slabiky o dvou až třech fonémech. Posloupnost jedné a více slabik tvoří slovo, které v českém jazyce může obsahovat i více než deset slabik. Příkladem může být nejdelší české slovo *nejneobhospodařovávatelnějšími*, většinou se objevují slova kratší, vyjímkou nejsou slova neslabičná (některé předložky a příklonky)[3].

Slabiku lze rozdělit na slabičné jádro a svahy slabiky. Nalézt přesnou hranici mezi slabikami není vždy jednoduché a jednoznačné. Nejčastěji se slabiky vytvářejí podle typu souhláska-samohláska např. [`te-le-vi-ze`], [`ho-di-ny`]. V případě výskytu dvou souhlásek mezi samohláskami vzniká hranice mezi souhláskami např. [`žid-le`], [`vej-ce`]. Hranice slabiky bývá často umístěna v morfologickém švu, tato hranice však není vždy jednoznačná, protože cit pro morfologický šev se může lišit např. [`ne-jíst`], [`o-sa-hat`].

Tab. 1.1: Srovnání fonetických abeced češtiny. Tabulka převzata z publikace [3]

	ZČFA	IPA	SAMPA	ČFA	Příklad		ZČFA	IPA	SAMPA	ČFA	Příklad
vokály	i	ɪ	i	i	lis	plozivy	p	p	p	p	pec
	e	ɛ	e	e	pes		b	b	b	b	bratr
	a	a	a	a	sad		t	t	t	t	tuk
	o	ɔ	o	o	kov		d	d	d	d	dům
	u	ʊ	u	u	sukně		ť	c	c	tj	děti
	í	iː	i:	ii	víno		ď	ɟ	J\	dj	dítě
	é	ɛː	e:	ee	lék		k	k	k	k	kost
	á	aː	a:	aa	sál		g	ɡ	g	g	tygr
	ó	oː	o:	oo	kód	nazály	m	m	m	m	muž
	ú	uː	u:	uu	růže		n	n	n	n	víno
diftogy	o͡u	o͡ʊ	o_u	ow	bouda		ň	ɲ	J	nj	laňka
	a͡u	a͡ʊ	a_u	aw	auto	afrikáty	c	t͡s	t_s	c	cena
	e͡u	e͡ʊ	e_u	ew	eunuch		č	t͡ʃ	t_S	ch	oči
frikativy	f	f	f	f	fík		d͡z	d͡z	d_z	dz	podzim
	v	v	v	v	vítr		d͡ž	d͡ʒ	d_Z	dzh	džbán
	s	s	s	s	sůl	významné alofony	ŋ	ŋ	N	ng	tango
	z	z	z	z	koza		ŋ̍	ŋ̍	F	mg	tramvaj
	š	ʃ	S	sh	škola		ɣ	ɣ	G		abych byl
	ž	ʒ	Z	zh	žena		ř̥	r̥	Q\	rsh	tři
	x͡	x	x	x	chata		r̩	r̩	r=		krk
	h	ɦ	h\	h	hůl		l̩	l̩	l=		vlk
	l	l	l	l	vlak		m̩	m̩	m=		osm
	r	r	r	r	rok		ʔ	ʔ	?		„ráz“
	ř	r̝	P \	rzh	moře		ə	ə	@		„šva“
	j	j	j	j	jev						

2 FONETICKÁ TRANSKRIPCE ČEŠTINY

Úkolem fonetické transkripce je pomocí symbolů fonetických abeced přesně popsat a zapsat jednotlivé zvuky mluvené řeči. Dle zaměření práce či požadavků můžeme volit jednotlivé abecedy, pro zahraniční či světové práce je vhodná IPA nebo SAMPA, pro naši národní formu se více hodí ČFA a ZČFA. Pokud budeme pracovat na počítači, je vhodné volit ze dvojíce abeced SAMPA a ČFA.

Dále lze volit abecedu dle toho, jak přesný přepis chceme vytvářet. V případě volné transkripce užijeme abecedu na bázi fonémů, výsledkem pak bude fonologická transkripce. Pro zpřesnění můžeme tuto fonologickou transkripci rozšířit o významné alofony. Nejpřesnější transkripce vznikne abecedou, která obsahuje „jemné"alofony. Pro přesné a kvalitní zpracování mluvené řeči je tedy nutné pracovat až na úrovni alofonů. Problémem je, že fonémy se mluvené řeči nevyskytují osamoceně, ale jsou spojovány do slabik a slov. Tím dochází ke změnám jednotlivých fonémů, vyjímečně může dojít ke vzniku nového fonému či vypuštění stávajícího. Fonetická transkripce tudíž musí umět popsat i tyto změny v mluvě tak, aby je byla schopná zaznamenat.

Při popisu mluvené řeči se nejčastěji zajímáme o spisovnou výslovnost. Stejně jako jsou pravidla pro zápis spisovného jazyka (pravopis), tak jsou i pravidla pro zápis mluvené podoby. Uznávaná norma spisovné výslovnosti se odborně nazývá ortoepie (je to souhrn obecně platných zásad a pravidel spisovné výslovnosti). Spisovná výslovnostní forma však není jednotná a existuje více výslovnostních stylů, které mají i různá pravidla. Nejčastěji se spisovná výslovnost rozděluje do tří stylů:

1. **Styl vybraný („vyšší").**
   Užíva se nejčastěji při oficiálních projevech nebo při horších akustických podmínkách. Jeho znakem je pomalejší tempo a pečlivá výslovnost.
2. **Styl základní („neutrální").**
   Jde o běžnou a nejčastější formu, se kterou se setkáme při přednáškách a jednáních. Nejčastěji se vyskytuje u učitelů a hlasatelů zpravodajství. Tento styl také dále využiji v praktické části této práce.
3. **Styl zběžný („nižší").**
   Je spisovnou formou hovorové řeči, užívané nejvíce u sportovních komentátorů při živých přenosech. Charakteristické pro tento styl je rychlé tempo a časté zjednodušování souhláskových spojení.

2.1 Automatická fonetická transkripce

Ke zpracování jazyka nám nestačí pouze analýza psaného textu. Je důležité znát i mluvenou podobu jazyka. Fonetická transkripce tudíž slouží ke dvoum důležitým činnostem, jedním je rozpoznání mluveného slova, druhým využitím je tvorba syntetické řeči ze zapsaného textu. Fonetický přepis má daná specifická pravidla, obecně se nazývají fonetická (fonologická) pravidla. Následující struktura a definice je převzata z [3]:

JESTLIŽE	řetězci znaků A bezprostředně předchází řetězec znaků C a je bezprostředně následován řetězcem znaků D,
PAK	se A přepíše na řetězec znaků B.

Pro jednoduchost budu dále toto pravidlo zapisovat ve tvaru:

$A \rightarrow B / C\_D$

Řetězec A, C a D pak reprezentují posloupnost v běžném psaném textu, řetězec B pak představuje posloupnost symbolů z použité fonetické abecedy. Na základě obecně formulovaných zásad pro výslovnost českých slov, byla vytvořena základní produkční pravidla pro přepis českého textu do řetězce hlásek. Navíc existuje fonetický slovník vyjímek, který obsahuje slova přejatá či cizího původu, jelikož takováto slova by bylo velice těžké těmito pravidly popsat. Zpracování textu by mělo následovat přibližně v těchto krocích:

1. Text lze zpracovávat buď zleva do prava, anebo zprava doleva. Druhá varianta je vhodnější pro řešení problému s vícenásobnou asimilací znělosti.
2. Dále zjistíme, zda lze na znaky uplatnit vyjímku z fonetického slovníku vyjímek. Zde záleží na tvaru slovníku (zda obsahuje celá slova, nebo pouze jejich části). Pokud se tedy znaky (slova) shodují, tak tuto vyjímku využijeme.
3. Pokud nelze využít žádnou vyjímku, tak vybereme vhodné produkční pravidlo. Ne vždy je kontext jednoznačný, takže můžeme aplikovat více pravidel a výsledkem nám je více variant fonetické transkripce.
4. V případě že zůstane znak, na který nelze použít vyjímka a ani využít žádne pravidlo, tak tento znak opíšeme, případně jej přizpůsobíme do zvolené fonetické abecedy.

2.2 Pravidla fonetického přepisu

Pro lepší čitelnost budu volit pro zápis výslovnosti využívat zjednodušenou českou fonetickou abecedu, jejíž znaky jsou popsány v tab. 1.1. Dále popsaná pravidla vychází z publikace [3], použité symboly pro popis pravidel jsou shrnuty v tab. 2.1.

Tab. 2.1: Pomocné symboly pro zápis fonetické transkripce češtiny. Tabulka převzata z publikace [3]

Značka	Popis	Příklad
V	samohlásky a dvojhlásky	([a],[á],[e],[é] . . . ,[u],[ú],[a͡u],[e͡u],[o͡u])
K	souhlásky	([b],[c],[č],[d] . . . ,[z],[ž])
ZPK	znělé párové souhlásky	([b],[d],[ď],[g],[v],[z],[ž],[h],[d͡z],[d͡ž],[ř])
NPK	neznělé párové souhlásky	([p],[t],[ť],[k],[f],[s],[š],[c͡h],[c],[č],[r̥])
JK	jedinečné souhlásky	([m],[n],[ň],[l],[r],[j])
′	hlavní přízvuk	[′jedu ′domú]
|	hranice mezi slovy	[dům | bil]
#	pauza	[# . . . řekl # že příjde . . . #]
-	vnitřní předěl (prefixový šef)	[v-lese],[na-jíst] atd.
+	vnitřní předěl (sufixový šef)	[buť+me] apod.
¬	znělostní protějšek	¬[b]=[p],¬[p]=[b],¬[d]=[t],¬[t]=[d]...
NP	neslabičné předložky	(k, s, v, z)
JPZ	jednoslabičné předložky	(bez, nad, ob, od, pod, před)
*	libovolný symbol	
⊘	prázdný symbol	

2.2.1 Základní pravidla

Mezi jednoduchá základní (bezkontextová) pravidla patří ta, která lze aplikovat na přepis českých písmen bez ohledu na kontext. Jedná se o písmena, která nemají stejný zápis v zjednodušené české fonetické abecedě, tedy *y*, *ý*, *ů* a *ch*. Tato pravidla je vhodné použít až po všech pravidlech pracujících s měkkými *i*, *í*, aby nedošlo k chybné transkripci. Zbylé české hlásky se pro zjednodušení pouze opíší stejným znakem. Tvar těchto pravidel je takovýto:

$$\mathrm{ch} \rightarrow \mathrm{c\overset{\frown}{h}} / \_$$

$$\text{ů} \rightarrow \texttt{ú}/\ \_$$
$$\text{y} \rightarrow \texttt{i}/\ \_$$
$$\text{ý} \rightarrow \texttt{í}/\ \_.$$

Příklad: *pyl* [pil], *netopýr* [netopír], *nůž* [núž], *chlap* [chlap].

2.2.2 Přepis dvojhlásek

V případě dvou sousedících samohlásek se výslovnost určuje podle toho, zda jsou tyto samohlásky součástí jedné slabiky či nikoliv. Jedná-li se o dvojhlásky, pak je přepis následující:

$$\text{ou} \rightarrow \texttt{o͡u}/\ \_$$
$$\text{au} \rightarrow \texttt{a͡u}/\ \_$$
$$\text{eu} \rightarrow \texttt{e͡u}/\ \_.$$

Příklad: *kousat* [kousat], *aut* [aut], *pneumatika* [pneumatika].

2.2.3 Spojení samohlásky a souhlásky

Takováto spojení se vyskytují na hranici slabik i uvnitř slabik samotných. Při výslovnosti obecně nedochází ke změnám, s vyjímkou hlásek [d], [t], [n] ve spojení s [i].

$$\text{d} \rightarrow \texttt{d'}/\ \_(\text{i, í})$$
$$\text{t} \rightarrow \texttt{t'}/\ \_(\text{i, í})$$
$$\text{n} \rightarrow \texttt{ň}/\ \_(\text{i, í}).$$

Příklad: *díra* [d'íra], *protiva* [prot'iva], *nic* [ňic].

K té samé změně dochází i v případě *dě*, *tě*, *ně*, *mě*.

$$\text{dě} \rightarrow \texttt{d'e}/\ \_$$
$$\text{tě} \rightarrow \texttt{t'e}/\ \_$$
$$\text{ně} \rightarrow \texttt{ňe}/\ \_$$
$$\text{mě} \rightarrow \texttt{mňe}/\ \_.$$

Příklad: *děda* [d'eda], *stěna* [st'ena], *němý* [ňemý], *měch* [mňech].

U souhlásek [b], [p], [v] ve spojen s [ě] dochází k následujícímu přepisu:

$$\text{ě} \rightarrow \texttt{je}/\ (\text{b, p, v})\_.$$

Příklad: *oběd* [objed], *pěna* [pjena], *věnec* [vjenec].

2.2.4 Výslovnost souhláskových skupin

Tyto skupiny jsou v češtině poměrně běžné a vyskytují se uvnitř jedné slabiky i na hranici slabik. Na hranici slabik se může vyskytovat uvnitř slov (např. *skr-čit*), na hranici slov, dále mezi slovy bez pauzy (např. *byl bych šel*), a nebo mezi kmenem slova a jeho předponou (např. *za-kopnout*). Přičemž dochází k asimilaci (sblížení) vlastností přilehlých souhlásek, což zjednodušuje výslovnost a tvoří ji plynulejší.

2.2.5 Asimilace znělosti

Podstatou asimilace znělosti je změna znělosti souhlásek uvnitř souhláskové skupiny [3]. Naše artikulační ústrojí provádí asimilaci (spodobu) znělosti automaticky, aniž bychom si to uvědomovali. Při výslovnosti dochází k drobným změnám znělosti celé skupiny a zjednodušení artikulace. Poté se párové dvojíce liší pouze znělostí. Existují dva způsoby, pomocí níchž můžeme vyrovnat znělost. Prvním je **regresivní (zpětná) asimilace**, kdy je pro výslednou znělost skupiny důležitá poslední souhláska (např. *sklidit* [`zklidit`]). V češtině se nejčastěji užívá tato regresivní asimilace. Druhým způsobem je **progresivní (postupná) asimilace**, jejíž výsledná znělost se řídí podle první souhlásky (např. *shodit* [`schodit`]). K této asimilaci dochází v češtině jen vyjímečně.

Z uvedených způsobů vyrovnání znělosti lze formulovat pravidla pro výslovnost souhlásek. Při spojení dvou a více souhlásek dochází ke změně znělosti podle poslední souhlásky ze skupiny. Dochází k tomu uvnitř slov, na hranici slov vyslovovaných bez pauzy i na vnitřním předělu. Uvnitř slova vyvolávají znělost párové souhlásky, na hranici slov řídí asimilaci samohláska, ráz, nebo párová souhláska.

Ke spodobě znělé párové souhlásky na neznělou dochází i před pauzou (po pauze se v češtině výslovnost souhlásek nemění) [3]

$$ZPK \rightarrow \neg ZPK/\ \_(NPK, \text{-}NPK, \text{-}ʔ, |NPK, |JK, |V, |ʔ, |\#)$$
$$NPK \rightarrow \neg NPK/\ \_(ZPK, \text{-}ZPK, |ZPK).$$

Příklad: *plechovka* [`plechofka`], *hněv* [`hňef`], *hrb* [`hrp`].

V dalších případech ke znělosti nedochází a dané souhlásce zůstává její znělost (popř. neznělost).

Příklad: *taška* [`taška`], *sleva* [`sleva`], *směna* [`smňena`].

Pravidla pro asimilaci obsahují tyto vyjímky:

1. Výslovnost skupiny *sh*, již lze vyslovovat zněle i nezněle, což závisí na zvoleném typu asimilace. Zpětná asimilace se převážně používá na Moravě, kdežto postupná se spíše užívá v Čechách. Obě tyto výslovnosti jsou považovány za

spisovné s vyjímkou slov *shora*, *shůry* a *shluk* a jejich variant, kde je spisovná pouze znělá výslovnost

$$\text{sh} \rightarrow \text{zh}/\ \_$$
$$\text{sh} \rightarrow \text{s}\underset{\smile}{\text{ch}}/\ \_.$$

2. Znělost fonému /v/, který tvoří znělostní dvojíci s fonémem /f/, se řídí též vyjímkou z pravidel. V okolí souhlásek nezpůsobuje změnu znělosti předchozí neznělé souhlásky. Dokonce na hranici slov vyvolá spodobu znělosti poslední znělé párové souhlásky na konci předchozího slova [3].

$$\text{v} \rightarrow \texttt{f}/\ \_NPK$$
$$NPK \rightarrow NPK/\ \_\ (\text{v, -v,}|\text{v})$$
$$ZPK \rightarrow \neg ZPK/\ \_\ |\text{v}.$$

3. Zvláštní postavení mezi českými souhláskami má foném /ř/, který nemá na fonologické úrovni neznělý protějšek. V tabulce 2.1 jsme jej mezi párové souhlásky zařadili, protože na fonetické úrovni existuje znělostní pár: znělé [ř] a neznělé [$\underset{\circ}{\text{ř}}$]. Platí pro ně podobné vlastnosti jako pro ostatní znělostní páry, ale mohou podléhat jak regresivní, tak i progresivní spodobě [3]

$$\text{ř} \rightarrow \underset{\circ}{\text{ř}}/\ NPK\ \_.$$

4. Končí-li kmen slova znělou párovou souhláskou a přípona začíná jedinečnou souhláskou (např. *buďme, hleďme, snažme se* apod.), pak dochází k nejistotám ve výslovnosti. Doporučuje se využívat výslovnosti s neznělým protějškem na konci kmene, i když ani výslovnost se znělou souhláskou není chápána jako nespisovná [3]

$$ZPK \rightarrow \neg ZPK/\ \_\ + JK$$
$$ZPK \rightarrow ZPK/\ \_\ + JK.$$

5. Jednoslabičné předložky zakončené zakončené znělou párovou souhláskou (*JPZ*, viz tab. 2.1) si při vnitřním předělu zachovávají svou znělost, následuje-li nejen znělá párová, ale i jedinečná souhláska. Předložka *přes* se navíc vyslovuje, jako by byla psána se `z` [3]

$$ZPK\_1 \rightarrow ZPK\_1/\ JPZ\ \_(\text{-}ZPK\_2, \text{-}JK)\ .$$

6. Neslabičné předložky (*NP*) se obecně před párovou souhláskou na začátku následujícího slova řídí pravidly asimilace. Výslovnost neslabičných předložek před slovem začínajícím jedinečnou souhláskou nebo souhláskou [v] už je poněkud složitější [3]. Speciálně předložku *s* lze vyslovovat buď jako [s], nebo jako [z], při spojení s osobními zájmeny ji však vyslovujeme pouze [s]

$$\begin{aligned}
\text{z} &\rightarrow \texttt{z}/ \mid _(\text{-}JK, \text{-v}) \\
\text{k} &\rightarrow \texttt{k}/ \mid _(\text{-}JK, \text{-v}) \\
\text{v} &\rightarrow \texttt{v}/ \mid _(\text{-}JK, \text{-v}) \\
\text{s} &\rightarrow \texttt{s}/ \mid _(\text{-}JK, \text{-v}) \\
\text{s} &\rightarrow \texttt{z}/ \mid _(\text{-}JK, \text{-v}).
\end{aligned}$$

2.2.6 Asimilace artikulační

Obdobně jako u asimilace znělosti, tak i zde je postihnuta celá skupina sousedních hlásek. Rozdílem je, že se souhlásky nepřizpůsobují znělosti, ale právě artikulaci. Vyrovnávají se tudíž artikulační rozdíly hlásek stojící vedle sebe. Této asimilace se účastní párové i jedinečné souhlásky. V češtině se často uplatňuje v rychlé mluvě, ale na rozdíl asimilace znělosti k ní dojít buďto může, nebo také nemusí. Existuje i řada nespisovných variant artikulační asimilace, ale těmto se v této práci věnovat nebudu. Shrnu tudíž jen některé nejvíce užívané varianty: [3]

1. Předodásňová hláska [n] se uvnitř slova před [k], [g] vyslovuje jako měkkopatrová varianta [ŋ]. Výslovnost s [n] se v tomto případě považuje za nespisovnou

   $$\text{n} \rightarrow \text{ŋ}/\ \_(\text{k}, \text{g}).$$

   Příklad: *banka* [baŋka], *tango* [taŋgo].

2. Podobně se obouretná hláska [m] uvnitř slova před [v], [f] může vyslovovat jako retozubná varianta [ɱ]. Poznamenejme, že oproti předchozímu případu je frekvence spojení [ɱv] a[ɱf] v češtině velice nízká a navíc se jako spisovná považuje i nezměněná výslovnost (tj. s [m])

   $$\begin{aligned}
   \text{m} &\rightarrow \text{ɱ}/\ _(\text{v}, \text{f}) \\
   \text{m} &\rightarrow \texttt{m}/\ _(\text{v}, \text{f}).
   \end{aligned}$$

   Příklad: *tramvaj* [traɱvaj] i [tramvaj], *nymfa* [miɱfa] i [nimfa].

3. Předodásňové hlásky [t] a [d] se mohou uvnitř slova před tvrdopatrovým [ň] nahradit změkčenými tvrdopatrovými [t’], [d’]

   $$\begin{aligned}
   \text{t} &\rightarrow \texttt{t'}/\ _\text{ň} \\
   \text{t} &\rightarrow \texttt{t}/\ _\text{ň} \\
   \text{d} &\rightarrow \texttt{d'}/\ _\text{ň} \\
   \text{d} &\rightarrow \texttt{d}/\ _\text{ň}.
   \end{aligned}$$

   Příklad: *špatně* [špat’ně] i [špatně], *vodník* [vod’ník] i [vodník].

4. Obdobně se může předodásňové [n] uvnitř slova před tvrdopatrovými [t’], [d’] zaměnit za změkčené tvrdopatrové [ň]

$$\text{n} \rightarrow \text{ň} / \_(\text{ť}, \text{ď})$$
$$\text{n} \rightarrow \text{n} / \_(\text{ť}, \text{ď}).$$

Příklad: *puntík* [puň’ík] i [punt’ík], *anděl* [aňd’el] i [and’ěl].

3 PRAKTICKÁ REALIZACE

K realizaci mého skriptu jsem využil program `Matlab` ve verzi 7.11.0. Jedná se o přepis textu z `txt` souboru do české fonetické abecedy. Pro přepis jsou aplikována základní pravidla, pravidla pro spojení samohlásky a souhlásky, asimilace znělosti a některé nejvíce užívané případy asimilace artikulační. Jejich konkrétní podmínky jsou shrnuty v podkapitole 2.2. Výstupem je nový `txt` soubor, který by dále šlo například za použití syntetizéru řeči využít na převod textu do audio formátu (mluvené řeči).

Součástí skriptu je i funkce pro zpětný převod (tj. pro převod z české fonetické abecedy do klasického uceleného textu). Výsledkem zpětného převodu je opět textový soubor, který je nutno nechat zkontrolovat a opravit na pravopis pomocí programu `GNU Aspell` ve verzi 0.50.3.

Skript byl vytvářen postupně a rozšiřován o další podmínky a vyjímky. Proto není plně odladěn a některé případy by bylo možno popsat a ošetřit v rámci již existujících podmínek.

3.1 Převod do ČFA

Vlastní skript se spouští přímo v programu `Matlab` pomocí hlavního skriptu `transkripce.m` a je rozdělen na více částí.

V prvotní verzi byly jednotlivá písmena a znaky přepisovány podle základních pravidel a nebyly brány v potaz přilehlé znaky. V tomto kroku byl tedy primárně řešen přepis písmen s diakritikou (např. [`áíůšř`]) a dále písmen [`y`] a [`ch`]. Vzorové slovo *šedivě* by tak bylo přepsáno pouze na /`shedive`/, což není dostatečné.

Proto byla postupně implementována další pravidla, aby se dosáhlo co nejpřesnějšího fonetického přepisu. Po aplikaci pravidel pro spojení samohlásek a souhlásek (viz. odstavec 2.2.3) již došlo k zpřesnění do tvaru /`shedjivje`/ .

Přidáním asimilačních pravidel nakonec došlo k zpřesnění výstupního textu do takového tvaru, kdy se nejvíce podobá vyslovené řeči.

3.1.1 Popis programu

V první části se pouze řeší asimilace znělosti (jsou volány skripty `asim1.m` a `asim2.m`) a dochází ke zmenšení případných velkých písmen pomocí funkce `lower`. Způsob kontroly asimilace bude popsán dále v textu, v této části je umístěn duplicitně z důvodu kontroly vícesouhláskových skupin. Vstupní soubor `test.txt` je načten

do proměnné `fid` a postupně je tedy dle pravidel kontrolována a upravována asimilace v závislosti na znělosti, či neznělosti souhláskových skupin. Výsledný mírně upravený text je uložen do dočasného souboru `docas.tmp`.

Tento soubor je poté použit jako zdroj vstupních dat pro druhou část skriptu, kde je načten do proměnné `fid` a dále k vytvoření výstupního souboru `reseni.txt` do něhož se poté upravený text zapisuje.

Vybrané hlásky jsou rozděleny do skupin podle tabulky 2.1 a to do samohlásek (značených jako `sam`), jedinečných souhlásek (`jed`), znělých souhlásek (`zne`) a neznělých souhlásek (`nez`). Schéma toho programu je naznačeno na obrázku 3.1.

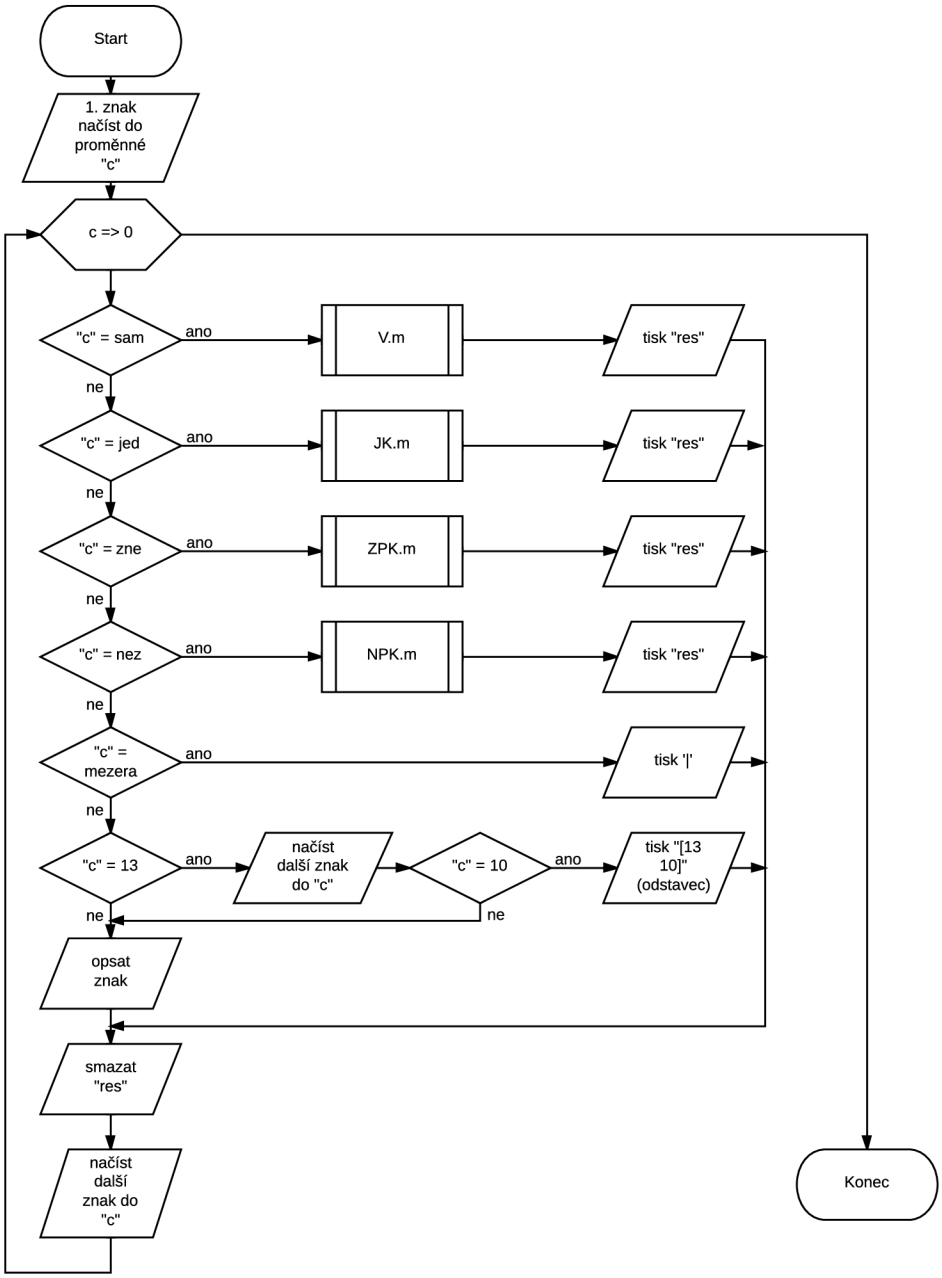
Pomocí funkce `fgets` je načten první znak z proměnné `fid` do proměnné `c`. Poté je zaveden cyklus, který se provádí dokud proměnná `c` není rovna nule. Tato proměnná je následně přejata do hlavní části programu `tran.m`, kde je znak v ní obsažen porovnán se sérií podmínek určujících zda znak náleží do samohlásek či různých druhů souhlásek, pokud ano je volána podfunkce řešící daný případ. V jiném případě se zkoumá zda je daný znak mezera, nebo zda se jedná o odsazení řádku. Pokud znak nesplní žádnou s podmínek, tak je pouze opsán (znaky `,.:` apod.). Výsledné znaky či skupiny znaků jsou zapsány do proměnné `res`, která je poté připisována do proměnné `fout`. Následuje smazání proměnné `res`, je načten následující znak do `c` a cyklus se opakuje. Po dokončení cyklu, je obsah proměnné `fout` vypsán a uložen do výstupního souboru `reseni.txt`.

`sam`

Jedná-li se o samohlásku, tak dojte k převzetí proměnné `c` do podfunkce `V.m`. V této části je ošetřen přepis krátkých a dlouhých samohlásek. Navíc je zde kontrolováno zda se nejedná o dvojhlásku. V případě [a], [e], [o] je načten další znak, pokud se jedná o [u] je tato dvojíce přepsána dle pravidla v tabulce 1.1. Pokud se jedná o jiný symbol, tak je předchozí samohláska zapsána a nový znak je navrácen do podprogramu `tran.m`, kde bude rozhodnuto jak bude s tímto znakem naloženo.

`jed`

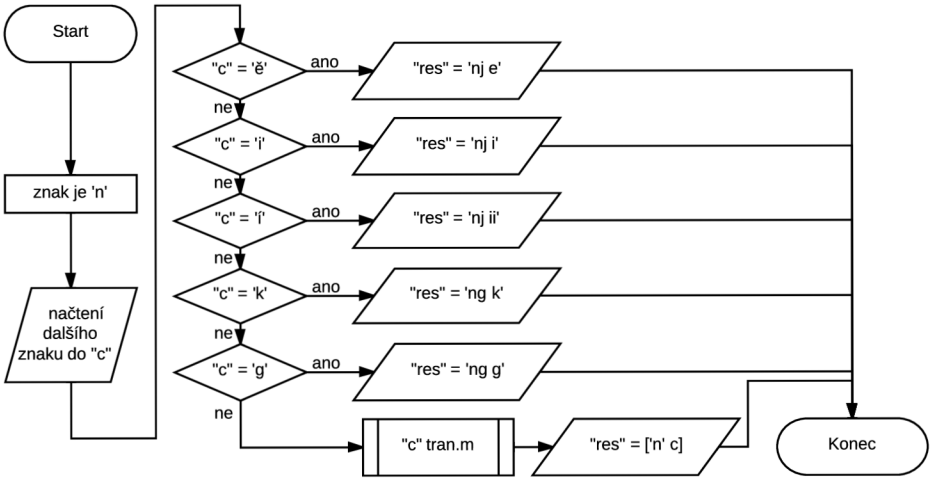
Pokud se jedná o jedinečnou souhlásku je volána podfunkce `JK.m`, kde je nejvíce řešen případ hlásek [m] a [n]. V těchto případech je opět načten následující znak a dle pravidel asimilací a spojení se samohláskami je vypsán příslušný znak, či skupina znaků. Příkladem budiž schéma na obrázku 3.2.



Obr. 3.1: Schéma programu `tran.m`

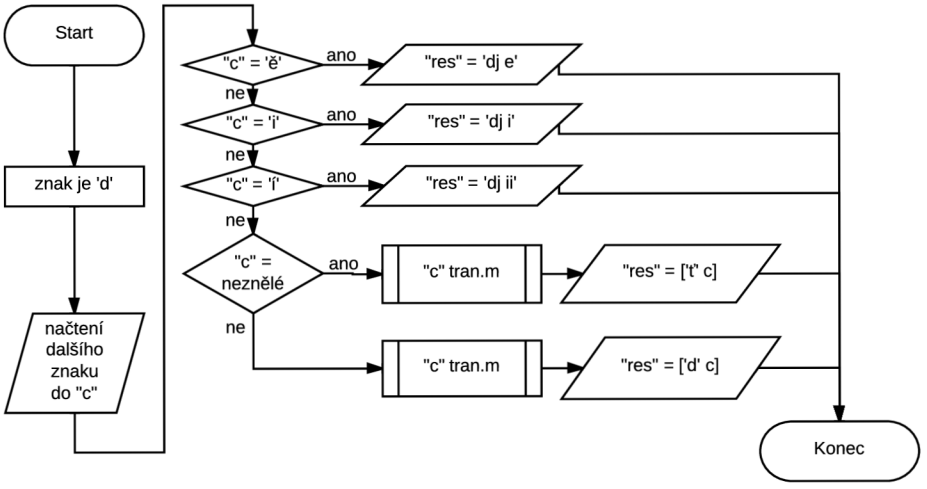
`zne`

V tomto případě je spuštěna podfunkce `ZPK.m` a obdobně jako u souhlásek jedinečných je i u znělých souhlásek, vyžaduje-li podmínka, načten další znak. Buďto je tedy znak přepsán podle pravidla pro daný znak, nebo je kontrolován následující



Obr. 3.2: Přepis hlásky [n]

znak a podle něj je případně upraven přepis i předchozího znaku. Nejčastěji se jedná o změkčení souhlásky, je-li následována znaky [i], [í], [ě], případně asimilací znělosti (viz. kapitola 2.2.5). Podrobnější schéma pro hlásku [d] ([děda je přepsáno na [d'eda]) je zobrazeno na obrázku 3.3.



Obr. 3.3: Přepis hlásky [d]

`nez`

Neznělé souhlásky volají podfunkci `NPK.m` a chovají se obdobně jako znělé. Také u nich dochází v některých případech ke změkčení souhlásek a řídí se podobnými asimilačními pravidly jako znělé souhlásky (samozřejmě vztahujícími se k neznělým). Proto i zde je v určitých případech nutno načíst následují znak, aby byl výsledný přepis co nejvíce foneticky správný.

3.2 Zpětný převod z ČFA

Zpětný převod z české fonetické abecedy zpět do klasického textu je opět řešen v rámci programu `Matlab`. Hlavním skriptem pro tuto část je soubor `zpet.m`. Výsledný textový soubor bohužel nezohledňuje rozdíly například mezi [ú] a [ů], či [i] a [y], neboť jsou tyto dvojice znaků ve fonetické formě reprezentovány stejnými symboly, a nelze je jednoduše zpětně rozlišit. Obdobný problém je i asimilace, kdy není patrné zda se jedná o skupinu souhlásek upravenou asimilací či nikoliv.

Z tohoto důvodu je výsledný soubor opraven a zkontrolován na pravopis programem `GNU Aspell`, který bude více popsán v kapitole 3.3.

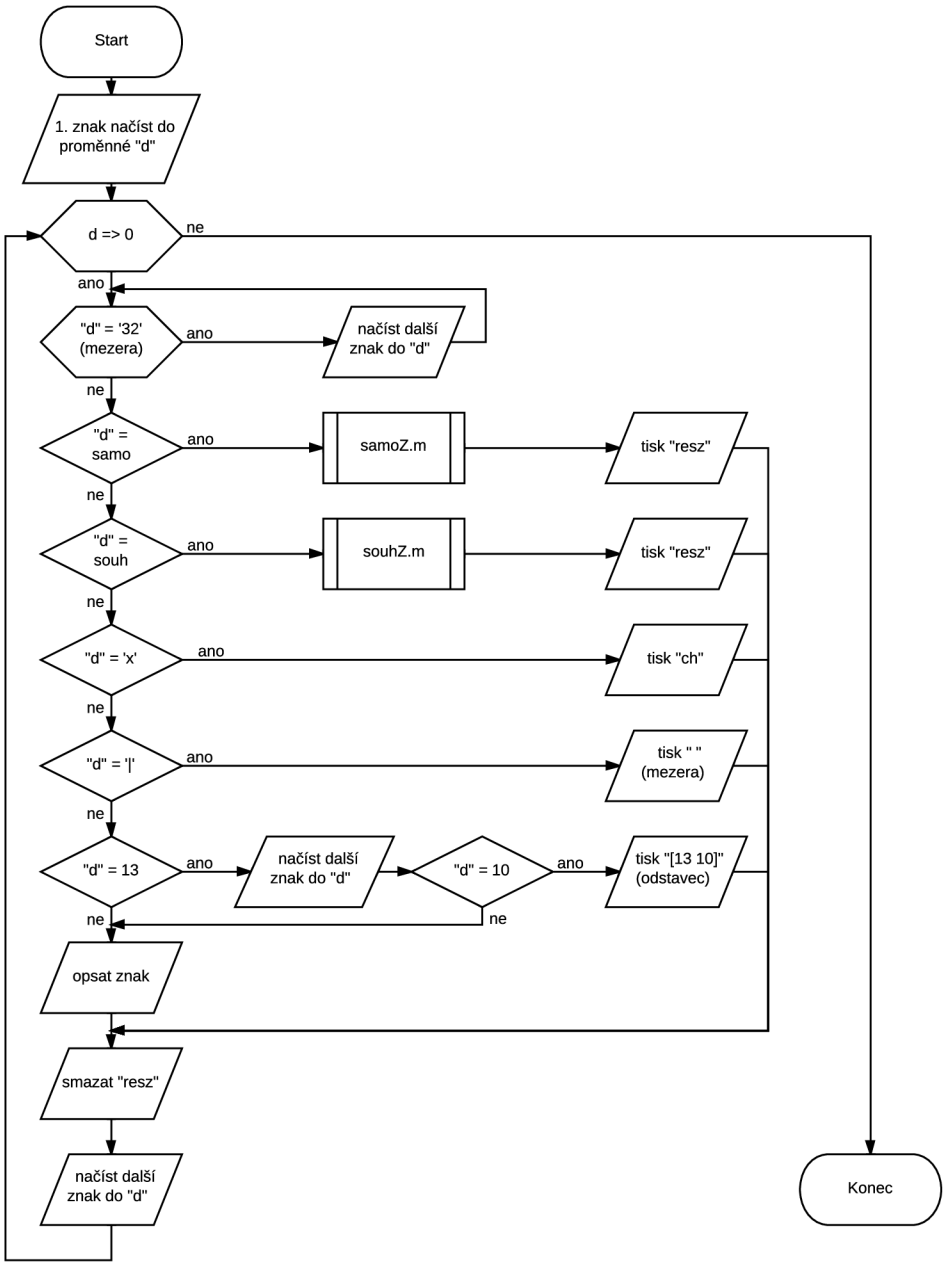
3.2.1 Popis zpětného programu

Jako vstupní soubor je nastaven soubor `reseni.txt`, umístěn v adresáři programu a jehož obsah je načten do proměnné `fidz`. Dále je vytvořen výstupní soubor `zpet.txt` do kterého je zapisován zpětně upravený text.

Opět, pomocí funkce `fgets`, je načten první znak z proměnné `fidz` do proměnné `d`. Poté je zaveden cyklus, který se provádí dokud proměnná `d` není rovna nule. Tato proměnná je poté přejata do hlavní části programu `step1.m`. Zde je vložen malý cyklus, který v případě, že se jedná o prázdný znak (mezeru), tak je automaticky do proměnné `d` následující znak. Písmena, jež byla nějakým způsobem změněna, jsem rozdělil do dvou proměnných samohlásky (značených jako `samo`) a souhlásky (`souh`). Jedná-li se o znak spadající do jedné z těchto proměnných, pak je volána její podfunkce (popsáno níže). V jiném případě je zjištěno, zda se jedná o znaky značící odsazení odstavce, nebo znak oddělující jednotlivá slova. Pokud se nejedná ani o jeden z těchto případů, pak je znak přímo opsán.

Výsledné znaky či skupiny znaků jsou tentokrát zapsány do proměnné `resz`, která je následně připisována do proměnné `foutz`. Poté je proměnná `resz` vymazána, je načten následující znak do `d` a cyklus se opakuje. Po dokončení cyklu, je obsah proměnné `foutz` vypsán a uložen do výstupního souboru `zpet.txt`.

Schéma toho programu je naznačeno na obrázku 3.4.



Obr. 3.4: Schéma zpětného programu [`step1.m`]

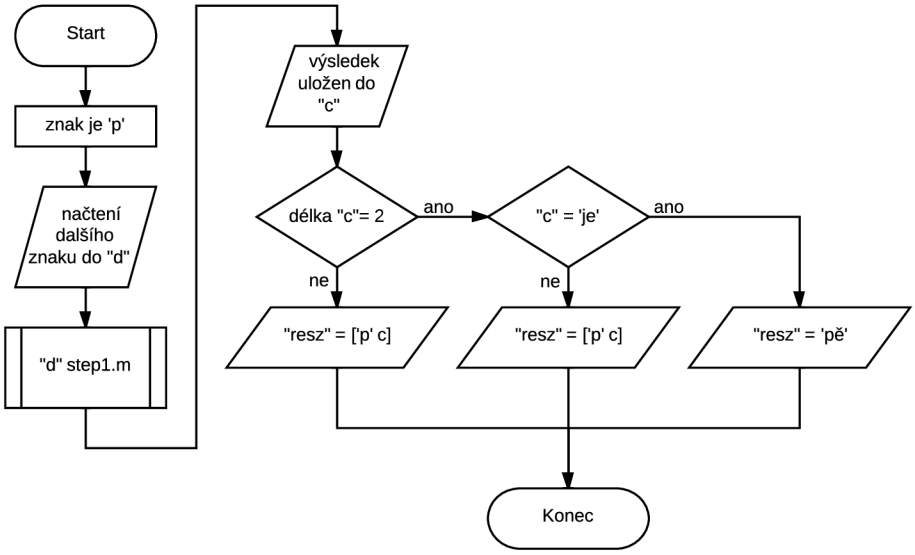
`samo`

Pokud se jedná o některou ze samohlásek, je volána podfunkce `samoZ.m`, která za pomoci dalšího znaku určí, zda se jedná o krátkou či dlouhou samohlásku, případně

dvojhlásku. Příklad: z [a], [aa], [aw] vytvoří [a], [á], [au].

souh

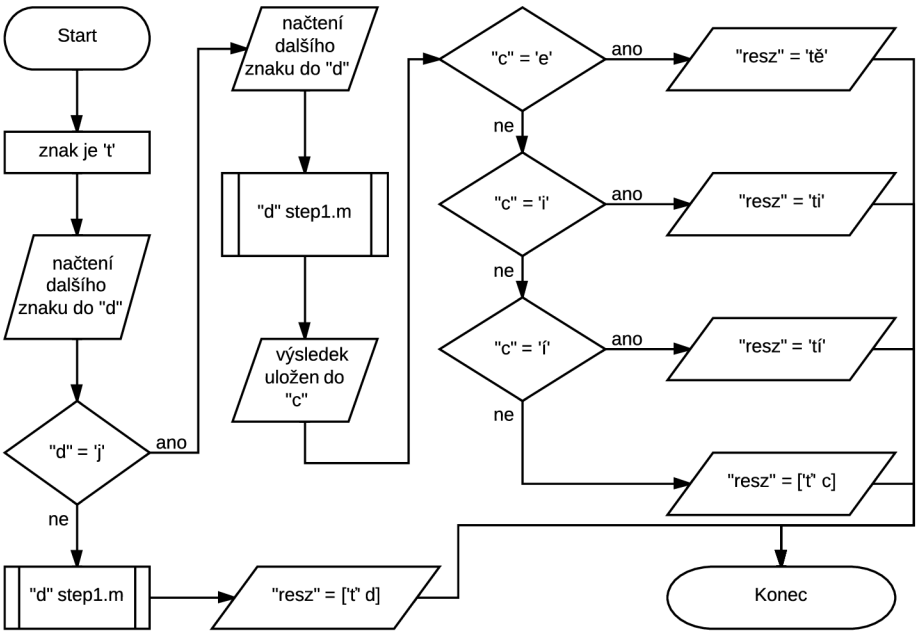
Spadá-li znak do vybraných souhlásek, pak je tento znak převzat do podfunkce souhZ.m. Zde jsou postupně zkoumány následující znaky a podle nich je určeno, jakým způsobem bude daná posloupnost znaků přepsána. Na obrázku 3.5 je naznačeno schéma zpětného přepisu hlásky [p]. Pro ilustraci je zjednodušené schéma zpětného přepisu hlásky [t] znázorněno na obrázku 3.6.



Obr. 3.5: Zpětný přepis hlásky [p]

3.3 Program ASPELL

GNU Aspell je volně šiřitelný Open Source program pro kontrolu pravopisu. Lze jej použít jako knihovnu slov, nebo jeho pomocí přímo provést kontrolu. Jeho hlavní výhodou je, že dokáže navrhnout případné návrhy za chybně napsané slovo. Dále je jeho slovník univerzální pro dokumenty v různých stylech kódování (UTF-8, CP1250 a jiné). Neposlední výhodou je inteligentní nakládání s osobními slovníky, kdy program časem přednostněji navrhuje správnou náhradu. Program je volně ke stažení z webové stránky http://aspell.net/win32/, kde se nachází instalátor v aktuální verzi 0.50.3 a dále na dvě desítky slovníků pro různé jazyky, včetně češtiny.[5]



Obr. 3.6: Zpětný přepis hlásky [t]

3.3.1 Popis korekce programem `Aspell`

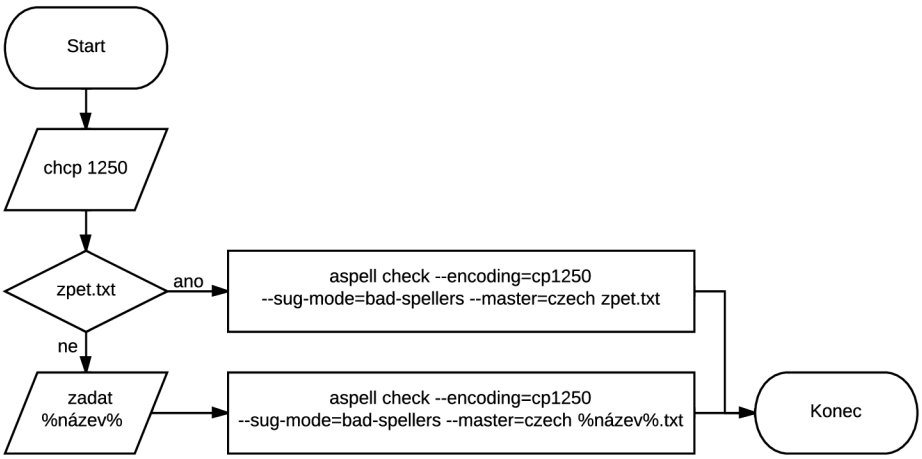
Po instalaci tohoto programu je třeba do složky `data`, v instalačním adresáři programu, nahrát soubor `cp1250.dat`. Ten zajistí správný přepis a zobrazení znaků u souborů v kódování CP1250.

Ve složce `bin` je umístěn soubor po zpětné transkripci (v mém případě soubor `zpet.txt`). Spouštění a nastavení parametrů probíhá přes příkazový řádek. pro zjednodušení a zrychlení jsem vytvořil spouštěcí soubor `run.bat` (též umístěn ve složce `bin`), který nejdůležitější parametry již obsahuje.

`run.bat`

Tento soubor spustí příkazový řádek. Jeho zjednodušené schéma je znázorněno na obrázku 3.7 a je popsáno dále. Jako první v něm pomocí příkazu `chcp 1250` změní kódování, aby se i zde korektně zobrazily znaky (někdy je potřeba ve vlastnostech okna příkazové řádku změnit font písma). Dále se zeptá zda má spustit kontrolu u výchozího souboru `zpet.txt`. Pokud ano, je kontrola spuštěna příkazem `aspell check --encoding=cp1250 --sug-mode=bad-spellers --master=czech zpet.txt`, kde *aspell check* spouští kontrolu, *encoding* značí typ kódování, *sug-mode* volí

přesnost a rychlost kontroly, *master* určuje jazykovou sadu a *zpet.txt* je zvolený soubor. Pokud ne, tak zadáme název zvoleného souboru a kontrola je spuštěna analogicky, s rozdílem názvu zvoleného souboru.



Obr. 3.7: Zjednodušené schéma souboru [`run.bat`]

V obou případech se textový soubor načte do okna příkazového řádku, kde jsou postupně zvýrazněna chybná slova. U takto vybraných slov jsou navržena různá alternativní a gramaticky správná slova. Volbou příslušného čísla s klávesnice dojde k jeho nahrazení, případně můžeme slovo přeskočit, nebo nahradit jiným, které není v nabídce. Po skončení kontroly se okno zavře a soubor je přepsán opravenou verzí, původní verze je zálohována do souboru s příponou `.bak`.

4 ZÁVĚR

V této práci jsem shrnul historii řeči a jejího studia z hlediska fonetiky a fonologie. Zaměřil jsem se na vznik, význam a užití fonetických abeced. Především jsem se zaměřil na mezinárodní abecedy IPA a SAMPA, včetně jejich historie a výhod či nevýhod, a dále vysvětlil základy české fonetické abecedy a pravidel správného přepisu výslovnosti. Na konci kapitoly jsem zmínil a popsal důležité české řečové jednotky, včetně jejich charakteristik a dělení.

V další kapitole jsem popsal pravidla fonetické transkripce češtiny a základní vlastnosti její automatické transkripce. Popsal jsem symboly užívané v pravidlech fonetické transkripce. Prostudoval a popsal jsem způsoby spojování samohlásek v českém jazyce, jejich spojení se souhláskami a výslovnost určitých skupin hlásek. Závěrem jsem rozdělil a popsal základní asimilační znělosti a asimilaci artikulační, a to včetně jejich vyjímek.

V poslední kapitole jsem popsal princip mnou vytvořeného programu a přidal jsem několik vývojových diagramů. Přepis do fonetického znění obsahuje většinu nejdůležitějších pravidel, včetně základních asimilací znělosti a asimilace artikulační. Zpětný převod je poněkud složitější, jelikož nelze jednoznačně určit u kterých fonému došlo předtím k asimilaci nebo jednalo-li se o měkké či tvrdé [i]. Proto je výsledek velice hrubý a je nutno na něj použít kontrolu pravopisu, kterou obstarává na CD přiložený program Aspell. U zpětného převodu se tedy nachází prostor pro zlepšení a zpřesnění výstupního textu. Přidáním vyjímek, pomocných proměnných nebo vložením slovníku vyjímek by šlo docílit přesnějšího přepisu tak, aby již nebylo potřeba využít program Aspell.

LITERATURA

[1] WIKIPEDIE. *Mezinárodní fonetická abeceda*. [online]. poslední aktualizace 3. 1. 2014 [cit. 2. 6. 2014]. Dostupné z URL: <http://cs.wikipedia.org/wiki/Mezin%C3%A1rodn%C3%AD_fonetick%C3%A1_abeceda>.

[2] WELLS, J.C., et al. *SAMPA computer readable phonetic alphabet*. [online]. 1997, poslední aktualizace 25. 10. 2005 [cit. 2. 6. 2014]. Dostupné z URL: <http://www.phon.ucl.ac.uk/home/sampa/>.

[3] PSUTKA, Josef, et al. *Mluvíme s počítačem česky*. 1. vyd. Praha: Academia, 2006. 746 s. ISBN 80-200-1309-1.

[4] PALKOVÁ, Zdena. *Fonetika a fonologie češtiny*. 1. vyd. Praha: Karolinum, 1994. 366 s. ISBN 80-706-6843-1.

[5] ATKINSON, Kevin. *GNU Aspell 0.61-cvs: Table of Contents*. [online]. 2004, poslední aktualizace 4. 7. 2013 [cit. 1. 6. 2014]. Dostupné z URL: <http://aspell.net/0.61/man-html/index.html>.

SEZNAM PŘÍLOH

.1 Obsah Přiloženého CD

- `xzedek02.pdf` – text bakalářské práce,
- `\matlab` – adresář se zdrojovými soubory,
  - `transkripce.m` – hlavní spustitelný skript pro převod textu do ČFA,
  - `zpet.m` – hlavní spustitelný skript pro zpětný převod z ČFA do textu,
- `\aspell` – hlavní adresář programu Aspell,
  - `\bin` – adresář pro umístění souboru a spuštění kontroly,
    - `run.bat` – spouštěcí soubor kontroly
- `\install` – adresář s instalačními soubory Aspell,
  - `README.txt` – návod na instalaci.