TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

METODY MĚŘENÍ PODOBNOSTI JAZYKŮ

Diplomová práce

Studijní program: N2612 – Elektrotechnika a informatika
Studijní obor: 1802T007 – Informační technologie

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

Liberec 2014

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

METHODS FOR LANGUAGE SIMILARITY MEASUREMENT

Diploma thesis

Study programme: N2612 – Electrical Engineering and Informatics
Study branch: 1802T007 – Information Technology

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

Liberec 2014

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií
Akademický rok: 2013/2014

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Radek Šafařík**
Osobní číslo:	**M12000230**
Studijní program:	**N2612 Elektrotechnika a informatika**
Studijní obor:	**Informační technologie**
Název tématu:	**Metody měření podobnosti jazyků**
Zadávající katedra:	**Ústav informačních technologií a elektroniky**

Zásady pro vypracování:

1. Seznamte se s problematikou měření podobnosti jazyků a existujícími metodami využívajícími paralelní korpusy i nezávislé texty.
2. Ze stránek Evropské unie stáhněte texty zásadních dokumentů a vytvořte tzv. paralelní korpus, který použijete pro experimenty s textově závislým porovnáním.
3. Z internetových stránek periodik vycházejících v různých evropských jazycích vytvořte dostatečně reprezentativní vzorek použitelný pro textově nezávislé měření podobnosti.
4. Implementujte různé metody měření podobnosti jazyků a otestujte je na výše zmíněných textech. Najděte řešení pro vhodné mapování specifických znaků jednotlivých jazyků (i těch, které nepoužívají latinku).
5. U slovanských jazyků zkuste vzít v úvahu při měření podobnosti též výslovnost (automaticky vygenerovanou na základě nejvýznamnějších výslovnostních pravidel).

Rozsah grafických prací:	**Dle potřeby dokumentace**
Rozsah pracovní zprávy:	**cca 40 - 50 stran**
Forma zpracování diplomové práce:	**tištěná/elektronická**

Seznam odborné literatury:

[1] **Nouza J., Koldovský Z a Vích R. Řeč a počítač. Technická univerzita v Liberci, 2009.**

[2] **internetové zdroje zaměřené na metody měření podobnosti jazyků**

[3] **Valta J. Identifikace jazyka textového dokumentu. Diplomová práce TUL, 2012.**

Vedoucí diplomové práce:	**prof. Ing. Jan Nouza, CSc.** Ústav informačních technologií a elektroniky
Datum zadání diplomové práce:	**12. září 2013**
Termín odevzdání diplomové práce:	**16. května 2014**

V. Kopecký

prof. Ing. Václav Kopecký, CSc.
děkan

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
vedoucí ústavu

V Liberci dne 12. září 2013

Prohlášení

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

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

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

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

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Poděkování

Rád bych poděkoval vedoucímu práce, panu Prof. Ing. Janu Nouzovi, CSc., za odborné vedení práce a za užitečné návrhy a připomínky, které mi pomohly při práci.

Dále bych rád poděkoval dhr. dr. A.A. Barentsenovi za poskytnutí paralelního korpusu ASPAC, který pomohl při měření podobnosti především slovanských jazyků.

Abstrakt

Tato diplomová práce je zaměřena na měření podobnosti evropských jazyků v jejich psané a mluvené formě. Pro měření podobnosti v mluvené formě se práce zaměřuje pouze na slovanské jazyky. Práce začíná teoretickým úvodem do komparativní lingvistiky a popisuje základní pojmy a principy hledání podobnosti jazyků. Jako zdroj dat pro měření podobnosti jsou použity různé textové korpusy a slovníky, které jsou dále podrobně popsány. Dále práce popisuje postupy pro předzpracování textů, vytvořený systém pro fonetickou transkripci textů do mezinárodní fonetické abecedy, metody použité pro měření podobnosti a výsledné vyhodnocení naměřených dat. Pro měření podobnosti jsou použity metody pracující s textovými řetězci a množinami znaků.

Klíčová slova

měření podobnosti jazyků, komparativní lingvistika, paralelní korpus, fonetická transkripce

Abstract

This diploma thesis is focused on similarity measurement of European languages in their written a spoken form. Thesis is focused only on slavic languages for similarity measurement in spoken form. Thesis begins with a theoretical introduction to the comparative linguistics and describes the basic concepts and principles of language similarity measurement. Different text corpuses and dictionaries are used as a source of data for measuring similarity which are further described in detail. Thesis describes procedures for preprocessing texts, created system for phonetic transcription of texts into International Phonetic Alphabet, methods used for similarity measurement and the final evaluation of the measured data. Methods working with text strings and sets of characters are used for similarity measurement.

Keywords

language similarity measurement, comparative linguistics, parallel corpus, phonetic transcription

Obsah

Seznam tabulek

Seznam grafů

Úvod

Cílem této diplomové práce je měření podobnosti jazyků, úloha spadající do oboru komparativní lingvistiky. Jelikož se jazyky vyvíjely ze společných předků, takzvaných protojazyků, anebo se při vývoji jinak ovlivňovaly, existují mezi těmito jazyky jisté podobnosti, které jsou různými metodami měřitelné. Tyto nalezené podobnosti lze dále využít v různých úlohách.

Jednou z aplikačních oblastí, která byla motivací pro vznik této práce, je problematika automatického rozpoznávání řeči. Ta je v posledních letech velmi aktuální a postupně vznikají systémy pro přepis mluvené řeči pro různé jazyky. V okamžiku, kdy je systém vyvinut pro konkrétní jazyk a chceme ho modifikovat pro jiný, míra podobnosti obou jazyků hraje významnou roli při volbě efektivní strategie použité pro modifikaci systému. U jazyků s vyšší mírou podobnosti lze totiž ve větší míře využít již hotový akustický model a při tvorbě výslovnostního slovníku lze též převzít řadu již připravených pravidel a modulů. V Laboratoři počítačového zpracování řeči na TUL je snaha využít této podobnosti pro efektivní adaptaci již vyvinutého systému rozpoznávání češtiny na další, zejména slovanské jazyky.

Tato práce se zaměřuje na porovnání podobnosti evropských jazyků v jejich psané podobě a u slovanských jazyků i v mluvené podobě. Porovnání je prováděno na textových korpusech a slovnících. Pro porovnání mluvené podoby jazyka je potřeba provést fonetickou transkripci textových dat.

Jelikož je v Evropě používáno několik abeced a každý jazyk má navíc své varianty, je dalším úkolem této práce vytvořit mapování mezi různými abecedami a mezi různými znakovými variantami v těchto abecedách.

Pro porovnání je třeba navrhnout a otestovat metody, které budou dávat použitelné výsledky. Tyto metody pro měření podobnosti pracují na principu porovnávání jednotlivých vět a slov. Jedná se o metody pro práci s řetězci a s množinami znaků.

1. Úloha měření podobnosti jazyků a její problémy

Úloha hledání podobnosti mezi jazyky je obecně úlohou komparativní lingvistiky. Jelikož je lingvistika spíše humanitním oborem, nesnaží se komparativní lingvistika při srovnávání jazyků o empirické měření, které by vyjádřilo číselnou podobnost, ale snaží se hledat společné rysy mezi jazyky, na základě kterých dále pracují jiné lingvistické úlohy. Hledání podobnosti bývá lingvistou prováděno ručně na základě jeho znalostí a zkušeností, kdy tak může být využito široké spektrum metod pro porovnání. Tato práce se však zabývá automatizací těchto metod a zpracování velkého množství dat pomocí výpočetní techniky. [1][2][3]

V podstatě základní úlohou komparativní lingvistiky je hledání takzvaného protojazyka, nebo naopak zjišťování, zdali porovnávané jazyky vycházejí z tohoto společného protojazyka. Protojazyk, nebo také prajazyk, je dávný jazyk (opravdový nebo hypotetický), ze kterého se vyvinuli dnešní jazyky. Protojazyk je základem jazykových skupin jako například praslovanština pro slovanské jazyky, starověká latina pro románské jazyky atd. Komparativní lingvistika si klade za cíl na základě podobností v různých jazycích hledat společné rysy a pomocí nich vytvořit tento protojazyk. Ten je z většiny případů pouze rekonstrukcí, jak mohl takový jazyk vypadat, jelikož neexistují důkazy, že tento jazyk existoval. To ale například neplatí v případě starověké latiny, která dala vzniknout skupině románských jazyků a o které se dochovala spousta záznamů a v jisté podobě se používá do dnes jako moderní latina. Je tak možné zpětně porovnat jak se jazyky vyvíjely a otestovat takto metody komparativní lingvistiky.

Porovnání je většinou prováděno mezi dvěma jazyky, případně i více, a z těchto porovnání se postupně sestavuje strom vedoucí k nalezení protojazyka. Pomocí těchto metod se zjišťuje takzvaná genetická příbuznost jazyků a následuje jejich seskupování do jazykových skupin a rodin. Jazykové skupiny seskupují jazyky (slovanské, germánské, románské, keltské…), které mají relativně mladého předka, jenž může být znám. Kdežto jazykové rodiny (indo-evropské jazyky, semitské jazyky, altajské jazyky…) seskupují jazyky a jazykové skupiny, které mají velmi dávného předka, o němž neexistují žádné důkazy, a je čistě hypotetický. Například čeština je příbuzná s ruštinou i s němčinou, ale s každým jazykem v jiné míře podle toho, kdy se tyto jazyky oddělily od společného předka. S ruštinou patří do stejné skupiny slovanských jazyků, zatímco s němčinou už jen do rodiny indo-evropských jazyků, takže s němčinou má mnohem staršího jazykového předka.

1.1 Základní pojmy

Před vysvětlením samotných postupů bude potřeba nejdříve vysvětlit základní pojmy. Zde tedy definuji pojmy, které jsou dále v práci používány. [4][5][6][7]

Grafemika je lingvistický obor zabývající se psanou formou jazyky a jeho písmem. Základní jednotkou je **grafém**, který je reprezentován nějakým znakem. Grafemika nerozlišuje grafémy podle jejich tvaru, jako například různé druhy písma či fonty, ale rozlišuje grafémy podle toho, zda ovlivní význam slova, které je obsahuje.

Fonetika je obor na pomezí lingvistiky, fyziologie a akustiky. Zabývá se zvukovou stránkou jazyků a způsobem jak se tvoří zvuk ve zvukovém traktu při mluvě. Základní jednotkou jsou **fóny**, které se rozlišují podle způsobu jejich tvorby, tedy podle pozice rtů, jazyka, zapojení hlasivek, proudění vzduchu atd.

Fonologie se na rozdíl od fonetiky zabývá těmi zvuky, které rozlišují význam. Různé fóny, které ale mají dále v mluvě stejný význam, nazývá **alofóny**. Základní jednotkou je **foném**.

Lexikografie je obor, který se zabývá slovní zásobou jazyka a tvorbou slovníků.

Morfologie je lingvistická disciplína zabývající se tvaroslovím, tedy ohýbáním slov (skloňování, časování).

Syntax je disciplína zabývající se větnou skladbou, vztahy mezi slovy ve větě a správnou konstrukcí vět.

Sémantika je nauka o významu slov, vět a dalších jazykových struktur. Při určování významu využívá předchozích lingvistických oborů.

1.2 Postupy při měření podobnosti jazyků

Hlavními metodami komparativní lingvistiky je porovnání fonetického a fonologického systému, morfologického systému, syntaktického systému, slovníku a případně dalších lingvistických systémů. Tato porovnání, jak už bylo zmíněno, jsou prováděny zkušeným lingvistou na základě jeho znalostí a nemají nějaký pevně daný algoritmus. [1][2][3]

Základním krokem je sestavit seznam základních slov, které se vyskytují ve všech porovnávaných jazycích. Je nezbytné, aby slova měla stejný význam a dále se mohly jen porovnávat odlišnosti. Tyto slova jsou porovnávána především na fonetické úrovni, protože ta nejvíce reprezentuje daný jazyk. Problém porovnávání na grafemické úrovni je v tom, že většina národů písmo převzala od jiných národů (v Evropě především latinka), a to většinou plně nereflektuje fonetický systém daného jazyka, což z hlediska hlubšího porovnání jazyků není příliš vhodné.

Dalším krokem je hledání korespondencí mezi fonémy a případně znaky. To znamená, že se například v jednom jazyce vyberou slova, která začínají jedním konkrétním fonémem a k nim se naleznou slova stejného významu v druhém jazyce a zjistí se, jakým fonémem začínají. Pokud je použit jiný znak, ale stále jeden ve všech slovech, jedná se o korespondenci.

Například mezi angličtinou a latinou existuje korespondence mezi *d* a *t*:

Tabulka 1. Fonémové korespondence [2]

Angličtina	**t**en	**t**wo	**t**ow	**t**ongue	**t**ooth
Latina	**d**ecem	**d**uo	**d**ūco	**d**ingua	**d**ent

Dále se porovnávají fonetické změny, ke kterým mohlo dojít vlivem vývoje jazyka. Hledají se změny jako palatalizace (změkčení souhlásky, například *c* na *č*), změny znělosti (například změna *p* na *b*, či *t* na *d*). K tomu se využívá korespondencí z předchozího kroku.

Podobné postupy jsou prováděny i u dalších lingvistických systémů. Na základě těchto nalezených změn je nakonec vytvořen hypotetický systém, který by mohl odpovídat protojazyku.

Tyto popsané postupy je ovšem potřeba provádět ručně, což je velice pracné a nedá se aplikovat pro velké množství jazyků. Proto s vývojem výpočetní techniky vznikla **kvantitativně komparativní lingvistika**, která se pokouší tyto metody automatizovat a pomocí výpočetní techniky aplikovat na velké množství dat. Při automatizaci ovšem nelze využít širokých znalostí lingvisty a celkové porovnání se tak degraduje pouze na statistické vyhodnocení. Využívá tedy například různé metody pro porovnávání textových řetězců (nejvíce je asi využívána Levenshteinova vzdálenost). Déle jsou využívány různé pravděpodobností modely, které se snaží odhadnout, zda jazyk zapadá do vytvářeného stromu.

Komparativní lingvistika se také dále dělí na další podobory, které se zabývají podobnostmi v jazycích. Stručně to jsou například kontaktní lingvistika, ta se zabývá etymologií (původem slov) a přejatými slovy, kontrastivní lingvistika zabývající se rozdíly mezi jazyky v rámci výuky jazyků, jenž hledá základní rozdíly mezi mateřským jazykem studenta a vyučovaným jazykem tak, aby bylo jasné, na které rozdíly se má při výuce zaměřit. Dále například glottochronologie, která se snaží přesně datovat dobu, kdy se dva jazyky oddělily od společného předka.

1.2.1 Data pro měření podobnosti

Zatímco klasická komparativní lingvistika využívá pro porovnání malou množinu základních slov, na kterých provádí komplexní porovnání, kvantitativně komparativní lingvistika pracuje s velkým množstvím dat, jako jsou rozsáhlé textové korpusy. Na těchto objemech dat se snaží hledat podobnosti a poté statisticky vyhodnotit výslednou podobnost.

Nejvhodnější pro porovnání jsou paralelní korpusy, které obsahují stejné texty v různých jazycích, čímž je při porovnání zajištěna sémantická podobnost. V tomto případě vzniká ale problém s porovnáváním na fonetické úrovni. Je potřeba provádět fonetické transkripce, nebo porovnávat pouze na grafemické úrovni, což ale snižuje výslednou přesnost.

1.2.2 Problémy při měření podobnosti

Největším problémem, na který komparativní lingvistika naráží, jsou přejatá slova. To jest slova, která byla do jazyka převzata z jiného jazyka, ať už z jakéhokoliv důvodu. V češtině je to spousta slov přejatých z němčiny, která se ustálila v používání, ale nemají přitom žádný základ ve slovanských jazycích. Tato slova mohou ovlivnit výslednou podobnost.

Další problém při měření podobnosti na grafemické úrovni je písmo. Jak už bylo zmíněno, různé národy používají různé písmo, které většinou převzaly a neodpovídá tak jejich fonetickému systému. Při porovnávání jazyků píšících různým písmem je potřeba znaky jednoho písma převádět na znaky druhého. Tato práce se zabývá měření podobností evropských jazyků, kde jsou používány tři druhy písma, latinka, cyrilice a řecká alfabeta. Zde také vzniká z informatického hlediska problém s kódováním textů. Je potřeba, aby porovnávané texty byly zpracovávány ve stejném kódování.

Při automatizaci měření podobnosti také vzniká problém homonymie slov. Tedy kde jsou nalezena dvě identická slova, která ale mají v každém jazyce úplně jiný význam. Je to například v úplně nejjednodušším případě spojka *a* v češtině a neurčitý člen *a* v angličtině. Při automatickém měření se tomuto dá částečně předejít využitím zmíněných paralelních korpusů, kde je zajištěna sémantická podobnost a dále už záleží jen na použitých metodách.

2. Data pro měření podobnosti

Jak už bylo zmíněno v úvodu, tato práce se zabývá měřením podobnosti na textových datech. Bylo tedy zapotřebí získat dostatečné množství dat v různých evropských jazycích. Nejvhodnější jsou paralelní korpusy a slovníky, kde je zaručena sémantická podobnost textů. Jedním z vhodných kandidátů jsou dokumenty Evropského parlamentu, které jsou volně dostupné ve všech úředních jazycích EU. Ale jelikož neobsahují dostatek slovanských jazyků pro vyžadované fonetické měření podobnosti a jejich texty jsou velmi specifické, nalezl jsem i další korpusy. Dále popíši veškerá získaná data.

2.1 Paralelní korpus

Jazykový korpus je soubor textů v daném jazyce, který je počítačově zpracován do formy vhodné pro jazykový výzkum. Může se jednat například o přidání různých značek do textu pro různé lingvistické analýzy nebo pouze o seskupení velkého množství textu pro statistické úlohy. Paralelní korpus obsahuje stejné soubory textů ve více jazycích, které opět mohou nebo nemusí být označkovány, ale především mohou být zarovnány do vět či odstavců, tak aby bylo možné vzít větu v jednom jazyce a k ní najít stejnou větu v jazyce druhém. To je užitečné pro různé jazykové úlohy, například pro učení inteligentních překladačů nebo pro měření podobnosti jazyků. Je ale potřeba zdůraznit, že překlad nemusí být a také není vždy stoprocentní a tak může být do výsledku konkrétní úlohy zanesena určitá chyba. [8]

2.1.1 Zarovnání textů

Jelikož je v této práci porovnání prováděno především na základě vět a jejich slov, je tedy potřeba, aby texty korpusu byly zarovnány tak, aby si stejné věty nebo alespoň odstavce odpovídaly. Vzhledem k tomu, že zarovnání textů je samo o sobě náročná úloha a také skutečnost, že se mi podařilo nalézt už zarovnané korpusy, nebudu zde zarovnání probírat podrobně, jen pro úplnost zmíním základní principy zarovnávání paralelních textů. [9][10][11]

Principy zarovnávání

Nejlepším možným způsobem zarovnání je ruční zarovnání, při kterém odborník, který zná zarovnávané jazyky, postupně projde celý text a zarovnává větu po větě, nebo alespoň odstavec po odstavci, jako je to v případě dále popsaného korpusu ASPAC. Tento způsob ovšem nelze využít u velkého množství textů a je proto potřeba

využít výpočetní techniku a zarovnávací algoritmy, ale na úkor snížení přesnosti zarovnání.

Při strojovém zarovnání se využívají následující principy. Nejprve, jak už bylo zmíněno, je potřeba text rozdělit do vět. Základním principem je vyhledávání interpunkčních znamének na konci vět, jako je tečka, otazník, vykřičník. Problém je, že se tyto znaky mohou vyskytovat i ve větě, například za číslovkou, v přímé řeči, v závorkách a podobně. Je proto třeba rozšířit hledání například o kombinaci interpunkční znaménko, mezera a velké písmeno. Zde ale opět dochází k problému, že v některých jazycích podstatná jména začínají velkým písmenem a podobně. Pro větší přesnost je tedy potřeba zajít dále a provádět syntaktickou analýzu textu, k čemuž je ale zapotřebí slovník a gramatický model daného jazyka.

Po segmentaci textu na věty přichází na řadu významové přiřazení jednotlivých vět k sobě. Zde je základním principem přiřazování podle délky vět, jelikož se předpokládá, že dlouhá věta je přeložena opět na dlouho větu a krátká na krátkou. Tato metoda se nazývá Gale-Churchův algoritmus [10]. Někdy je ale možné, že se jedna dlouhá věta přeloží v jiném jazyce na dvě věty kratší. Pro větší přesnost je třeba přistoupit k pravděpodobnostním modelům nebo k sémantické analýze, což vyžaduje významové slovníky.

Ve většině korpusů použitých v této práci byl pro zarovnání použit algoritmus HunAlign [11]. Ten má pro zarovnání dvě možnosti. Pokud má k dispozici slovník, využije ho v kombinaci s Gale-Churchovým algoritmem. Pokud slovník nemá, vytvoří si ho na základě zarovnání textu pouze Gale-Churchovým algoritmem a poté s pomocí takto vytvořeného slovníku texty zarovná znovu v kombinaci Gale-Churchovým algoritmem.

Ukázka zarovnání českého a slovenského textu:

```
<p n="13">Za Radu Evropské unie</p>
<p n="14">B. Dopis vlády Demokratické republiky Svatý
Tomáš a Princův ostrov</p>
<p n="15">Vážení pánové,</p>
<p n="16">mám tu čest potvrdit přijetí Vašeho dopisu z
dnešního dne tohoto znění:</p>
```

```
<p n="13">V mene Rady Európskej únie</p>
<p n="14">B. List vlády Demokratickej republiky
Svätého Tomáša a Princovho ostrova</p>
<p n="15">Vážení páni,</p>
<p n="16">mám česť potvrdiť dnešným dátumom prijatie
Vášho listu tohto znenia:</p>
```

Chyby automatického zarovnání

Z principů pro zarovnání vyplývá, že nelze zarovnat texty vždy úplně přesně, a proto se občas objevují chyby zarovnání, které při měření podobnosti samozřejmě zkreslují výsledek. Je proto třeba provést výpočet podobnosti na velkém počtu vzorků, aby se chyby minimalizovaly.

Při náhodném průzkumu textů v českém a slovenském jazyce jsem narazil na několik chyb, ale v prakticky zanedbatelném množství.

Příklad špatného zarovnání v korpusu EUParl:

```
<p n=“2“>Dohoda ve formě výměny dopisů</p>
<p n=“3“>o prodloužení platnosti protokolu, kterým se
stanoví rybolovná práva a finanční příspěvek
podle Dohody mezi Evropským hospodářským společenstvím
a vládou Demokratické republiky Svatý Tomáš a Princův
ostrov o rybolovu při pobřeží Svatého Tomáše
a Princova ostrova, na období od 1. června 2005 do 31.
května 2006</p>

<p n=“2“>Dohoda</p>
<p n=“3“>vo forme výmeny listov o predížení platnosti
protokolu stanovujúceho na obdobie od 1. júna 2005
do 31. mája 2006 rybolovné možnosti a finančný
príspevok podľa Dohody medzi Európskym hospodárskym
spoločenstvom a vládou Demokratickej republiky Svätého
Tomáša a Princovho ostrova o rybolove pri pobreží
Svätého Tomáša a Princovho ostrova</p>
```

2.1.2 Získané paralelní korpusy

Pro měření se mi podařilo získat šest paralelních korpusů v různých jazycích s různě zaměřenými texty. Dohromady se jedná o 57 především evropských jazyků. Korpusy obsahují texty s různým zaměřením, použití jazyka se může tedy v každém korpusu značně lišit a bude proto zajímavé dále zjistit jak i obsah textu může mít vliv na výsledky měření. V tabulce jsou uvedeny získané korpusy se základními údaji a dále detailněji popíši jednotlivé korpusy.

Tabulka 2. Paralelní korpusy

Název korpusu	Počet jazyků	Velikost korpusu	Charakteristika korpusu
EUParl	22	9 GB	Dokumenty Evropského parlamentu
EMEA	22	24 GB	Příbalové informace k lékům
EUBookshop	36	112 GB	Příručky a knihy o EU
ASPAC	19	208 MB	Klasická literatura
KDE	47	14 GB	Lokalizační data (názvy, popisky, nápovědy)
Bible	33	170 MB	Překlady bible

EUParl

Paralelní korpus vytvořený z dokumentů Evropského parlamentu, které jsou k dispozici na internetu ve všech úředních jazycích EU. Dokumenty byly zarovnány a dále publikovány evropským výzkumným centrem Joint Research Center jako korpus JRC-Acquis. Korpus je zarovnaný na věty, případně několik vět. [12]

Texty jsou politického a hospodářského zaměření, nezastupují tedy celý jazyk ve všech směrech, a proto výsledky měření tohoto korpusu není možné aplikovat na celý jazyk, ale počítat s jistou odchylkou. V textech se často objevují konkrétní jména, názvy, zkratky, odborné termíny či výčty hodnot.

Texty jsou uloženy v xml souborech, pro každý jazyk vlastní soubor. Zarovnání je provedeno číslováním odstavců v jednotlivých souborech, které mají stejný název souboru, lišící se pouze v jazykovém kódu.

Ukázka českého textu:

> (1) Směrnice Rady 91/68/EHS [3], naposledy pozměněná rozhodnutím Komise 2001/10/ES [4], stanoví veterinární podmínky obchodu s ovcemi a kozami uvnitř Společenství.

Seznam produktů uvedených v čl. 10 odst. 2 a čl. 11 odst. 2

Kód KN | Popis |

0403 | Podmáslí, kyselé mléko a smetana, jogurt, kefír a jiné fermentované (kysané) nebo okyselené mléko a smetana, též zahuštěné nebo obsahující přidaný cukr nebo jiná sladidla nebo ochucené nebo obsahující přidané ovoce, ořechy nebo kakao: |

04031051 do 04031099 | – – – – Jogurt, ochucený nebo obsahující přidané ovoce, ořechy nebo kakao |

Ukázka zarovnání českého a slovenského textu:

Návrh
ROZHODNUTIA RADY
o podpísaní Dohody medzi Európskym spoločenstvom a Islandskou republikou a Nórskym kráľovstvom o dojednaní foriem účasti týchto štátov na činnosti Európskej agentúry pre riadenie operačnej spolupráce na vonkajších hraniciach členských štátov Európskej únie v mene Európskeho spoločenstva

Návrh
ROZHODNUTÍ RADY
o podpisu ujednání jménem Evropského společenství mezi Evropským společenstvím na jedné straně a Islandskou republikou a Norským královstvím na straně druhé o modalitách účasti Islandské republiky a Norského království v Evropské agentuře pro řízení operativní spolupráce na vnějších hranicích členských států Evropské unie

EMEA

Jedná se o korpus vytvořený z textů Evropské lékové agentury (European Medicines Agency), regulačního úřadu Evropské unie pro schvalování léčiv. [13]

Jedná se především o příbalové informace přeložené do 22 úředních jazyků EU. Texty jsou tedy farmaceutického zaměření a obsahují velké množství odborných názvů, které budou ve všech jazycích stejné nebo alespoň podobné. Texty jsou uloženy v xml souborech, vždy po dvojicích zarovnaných jazyků.

Ukázka českého a slovenského textu:

```
<tuv xml:lang="cs"><seg>Abilify je léčivý přípravek, který obsahuje účinnou látku aripiprazol.</seg></tuv>
<tuv xml:lang="sk"><seg>Abilify je liek, ktorého účinnou látkou je aripiprazol.</seg></tuv>
<tuv xml:lang="cs"><seg>Je dostupný ve formě tablet s obsahem 5 mg, 10 mg, 15 mg a 30 mg, ve formě tablet dispergovatelných v ústech (tablet, které se rozpustí v ústech) s obsahem 10 mg, 15 mg a 30 mg, jako perorální roztok (1 mg/ ml) nebo jako injekční roztok (7, 5 mg/ ml).</seg></tuv>
```

EUBookShop

Korpus sestavený z textů internetového knihkupectví, knihovny a archivu EU Bookshop, jenž obsahuje publikace institucí EU, jako jsou Evropská komise, Evropský parlament, Rada EU a další. Texty jsou k dispozici ve více než 50 jazycích. [13]

Pro účely této práce jsem vybral 36 evropských jazyků. Je ale potřeba zdůraznit, že pro některé neobvyklé jazyky není k dispozici dostatek textů a výsledek měření podobnosti proto může být značně zkreslen (jedná se například o jazyky jako skotská gaelština, bretonština, velština a podobně).

Texty jsou opět uloženy v xml souborech po dvojicích zarovnaných jazyků.

Ukázka českého a polského textu:

```
<tuv xml:lang="cs"><seg>Dosáhnout úspěchu s malým podnikem je náročné.</seg></tuv>
```

<tuv xml:lang="pl"><seg>Małym przedsiębiorstwom trudno jest odnieść sukces.</seg></tuv>
<tuv xml:lang="cs"><seg>Ať už takový podnik sami provozujete, pracujete pro něj nebo děláte obojí, můžete si asi myslet, že tou poslední věcí, které byste se měli věnovat, je bezpečnost a ochra-na zdraví při práci.</seg></tuv>

ASPAC

Paralelní korpus ASPAC (Amsterdam Slavic Parallel Aligned Corpus) jsem získal na požádání od doktora Barentsena z Katedry slovanských jazyků a kultur Fakulty humanitních studií na Amsterdamské universitě, který na něm několik let pracuje. Korpus je zaměřen především na slovanské jazyky, ale obsahuje i další jazyky jako angličtina, francouzština, řečtina a další., celkem 19 jazyků, z toho 12 slovanských. [14]

Korpus je složen z literárních děl klasických spisovatelů (jako např. Tolkien, Hemingway, Verne, Exupery, Hašek atd.) a je ručně zarovnán na odstavce velikosti maximálně pěti řádků, delší jsou rozděleny na víc odstavců. Každé dílo je vždy v ruském jazyce a minimálně jednom dalším. Texty tohoto korpusu, na rozdíl od ostatních korpusů, nejvíce reprezentují daný jazyk.

Texty jsou uloženy v souborech txt v různém kódování podle jazyka (cp-1250 až cp-1254), bylo proto potřeba v aplikaci pro měření při čtení souboru vybírat kódování dle jazyka. Zarovnání je zde provedeno po řádcích, bez žádného značení.

Ukázka českého textu:

I. kniha: V ZÁZEMÍ
1. kapitola: Zasáhnutí dobrého vojáka Švejka
do světové války
"Tak nám zabili Ferdinanda," řekla posluhovačka panu Švejkovi, který opustiv před léty vojenskou službu, když byl definitivně prohlášen vojenskou lékařskou komisí za blba, živil se prodejem psů,

ošklivých nečistokrevných oblud, kterým padělal rodokmeny.

Kromě tohoto zaměstnání byl stižen revmatismem a mazal si právě kolena opodeldokem.

Ukázka srbského textu v latince:

Prva knjiga U POZADINI

1 KAKO JE DOBRI VOJNIK ŠVEJK DOČEKAO SVETSKI RAT

- I tako nam ubiše Ferdinanda, - reče služavka gospodinu Švejku, koji je otpre nekoliko godina, pošto ga je vojna lekarska komisija definitivno proglasila za blesavog i otpustila iz vojske, živeo od prodaje pasa, gadnih, nečistokrvnih rugoba, kojima je falsifilkovao rodoslove.

Pored toga zanimanja još se baktao i sa reumom, i baš je trljao kolena opodeldokom.

KDE

Tento paralelní korpus je sestaven z lokalizačních dat linuxového grafického rozhraní KDE. Jedná se o názvy, popisky a nápovědy přeložené do 92 jazyků, z nichž jsem jich vybral 47. Texty jsou většinou krátké věty nebo několik slov, velmi se tedy blíží slovníku. Opět zde platí, že pro méně obvyklé jazyky je méně textů a tedy i výsledné měření může být zkreslené. [13]

Texty jsou v xml souborech po dvojicích zarovnaných jazyků.

Ukázka českého a chorvatského textu:

```
<tuv xml:lang="cs"><seg>Adblock zakázán</seg></tuv>
<tuv xml:lang="hr"><seg>Adblock onemogućen</seg></tuv>
<tuv xml:lang="cs"><seg>obrázek</seg></tuv>
<tuv xml:lang="hr"><seg>slika</seg></tuv>
<tuv xml:lang="cs"><seg>Blokovatelné položky na této
stránce</seg></tuv>
<tuv xml:lang="hr"><seg>Blokirajući elementi na ovoj
stranici</seg></tuv>
```

```
<tuv xml:lang="cs"><seg>Přidat filtr</seg></tuv>
<tuv xml:lang="hr"><seg>Dodaj filtar</seg></tuv>
```

Bible

Paralelní korpus sestavený z překladů bible do 100 rozličných jazyků (od arabštiny po zulštinu), z nichž jsem vybral 33 evropských jazyků. Některé překlady mohou být staršího data, a proto použitý jazyk může být dosti archaický a nezastupuje tedy úplně současně používaný jazyk. Výsledky měření se tedy mohou značně lišit od ostatních korpusů. [15]

Texty jsou uloženy v xml souborech a zarovnány podle veršů, označených atributem *id*.

Ukázka českého textu:

```
<seg id="b.GEN.1.4" type="verse">
A viděl Bůh světlo, že bylo dobré; i oddělil Bůh
světlo od tmy.</seg>
<seg id="b.GEN.1.5" type="verse">
A nazval Bůh světlo dnem, a tmu nazval nocí. I byl
večer a bylo jitro, den první.</seg>
```

Ukázka textu ve slovinštině:

```
<seg id="b.GEN.1.4" type="verse">
In videl je Bog svetlobo, da je dobra; in ločil je Bog
svetlobo od teme.</seg>
<seg id="b.GEN.1.5" type="verse">
In svetlobo je Bog imenoval dan, a temo je imenoval
noč. In bil je večer, in bilo je jutro, dan
prvi.</seg>
```

2.2 Slovníková data

Pro slovníková data jsem původně zvažoval nalézt kompletní slovník pro každou dvojici jazyků. To se ale brzy ukázalo jako velmi časově náročné. Zvolil jsem proto jinou variantu. Vyhledal jsem pět tisíc nejběžnějších slov v anglickém jazyce [16] a ty jsem pomocí webového překladače od společnosti Google [17] přeložil do 41 jazyků. Angličtina zde byla zvolena jako pivotní jazyk především z důvodu, že pokud překladač od Google nemá prostředky pro překlad přímo mezi dvěma jazyky, tak překládá nejdříve do angličtiny a poté do zvoleného jazyka. Tento způsob překladu samozřejmě není nejlepším řešením, protože může existovat více variant překladu a automaticky může být zvolena ta méně vhodná.

Měření podobnosti přes slovníky jsem zvolil na základě tzv. Swadeshiho seznamu [2], aparátu komparativní lingvistiky, jenž spočívá v sestavení seznamu 100 až 200 základních slov, které jsou přeloženy do porovnávaných jazyků a poté porovnávány různými lingvistickými metodami. V této práci se ale nezabývám metodami komparativní lingvistiky, které jsou prováděni ručně lingvistou na základě jeho znalostí, ale metodami informatickými, které lze algoritmizovat a naprogramovat. Proto jsem zvolil porovnání pěti tisíc slov, aby byly při automatickém zpracování minimalizovány odchylky měření.

Slovník obsahuje slova, jako jsou základní slovesa (být, mít, dělat, vidět), zájmena, spojky, číslovky a všední slova (matka, otec, život, práce, dítě), ale také různá další slova (chirurg, fyzika, sponzor, integrita).

2.3 Data pro textově nezávislé měření podobnosti

Pro textově nezávislé porovnání nebylo potřeba získávat další texty, ale po dohodě s vedoucím jsem tyto data vybral ze získaných korpusů. Pro textově nezávislé porovnání je využito statistické metody, a tedy obsah textů na porovnání nemá žádný vliv. Je potřeba pouze dostatečné množství dat, které korpusy obsahují.

Pro textově nezávislé porovnání jsem prošel texty a provedl frekvenční analýzu slov. Nechal jsem si tedy vypsat všechny nalezená slova s jejich četností.

3. Předzpracování textu

Před samotným výpočtem je nutné texty korpusu upravit, aby je bylo možné co nejpřesněji porovnávat. Jazyky jsou porovnávány na dvou úrovních, grafemické (textové) a fonetické (mluvené). Jelikož se práce zabývá pouze porovnáváním textových dat, bylo nutné pro porovnávání mluvené formy jazyka vytvořit systém a pravidla pro fonetickou transkripci textů, které jsou dále zpracovány stejně jako při grafemickém porovnání. Dalším úkolem bylo implementovat mapování specifických znaků jak pro text, tak pro fonetickou transkripci.

3.1 Úprava textu

Před samotným porovnáním je důležité převést všechny znaky na malá písmena a odstranit některé znaky, které jsou nezávislé na jazyce a zkreslovaly by podobnost. Jsou to především arabské číslice, která jsou ve všech evropských jazycích stejné a mají stejný význam, dále interpunkční znaménka a další různé znaky, které se v textech objevovaly jako například paragraf, zavináč, mřížka, lomítko a podobně. Seznam všech znaků, které bylo třeba odstranit, je obsažen v příloze.

3.2 Fonetická transkripce

I přesto že porovnání v této práci je prováděno na základě textů, bylo mým úkolem zaměřit se na porovnání výslovnosti jazyků. S ohledem na požadavek Laboratoře počítačového zpracování řeči a pro jednoduchost zpracování jsem se měl zaměřit pouze na slovanské jazyky, které jsou tzv. ortograficky mělké, což znamená, že jak se vyslovují, tak se i přibližně zapisují.

Pro měření podobnosti pomocí zvukových nahrávek by bylo potřeba získat dostatečné množství těchto nahrávek v různých jazycích, což by bylo mnohem složitější, než bylo získat textová data, a také metody pro měření podobnosti by se ubíraly jiným směrem mimo cíle této práce.

Měření výslovnosti je tedy prováděno pouze na textových datech, které jsou transkripcí převedeny na fonetický tvar a dále jsou porovnávány stejnými metodami jako při grafemickém porovnávání. Systém pro transkripci jsem s drobnými úpravami převzal z [18], naprogramoval pro tento systém aplikaci a dále pomocí internetových zdrojů nastudoval výslovnost 16 slovanských jazyků a vytvořil pro ně transkripční pravidla. Všechna vytvořená pravidla jsou obsažena v příloze.

Pro rozsáhlost problému jsem jako zdroje využil webové servery Omniglot [19] a anglickou Wikipedii [20], kde byla dostatečně popsána základní výslovnost.

Transkripce je provedena do Mezinárodní fonetické abecedy IPA [21]. Jedinou úpravou, ke které jsem musel přistoupit, byla změna takzvaných afrikát (t͡s, t͡ʃ - české c a č), které se zapisují dvěma znaky a ligaturou (spojníkem nad znaky). Ve výsledku jsou tedy reprezentovány třemi znaky, což by značně komplikovalo měření, a proto jsem je nahradil jedním velkým znakem, který jim přibližně odpovídá.

t͡s : C
d͡z : D
t͡ʃ : Č
d͡ʒ : Ď
t͡ʂ : Š
d͡ʐ : Ž
t͡ɕ : Ś
d͡ʑ : Ź

Výslovnost jsem neřešil do úplných detailů, vytvořil jsem pouze základní pravidla podle dostupných zdrojů. Neřešil jsem výslovnostní prvky, jako je například přízvuk, který je třeba v češtině na přesně daném místě, ale v ruštině je plovoucí a lze ho určit pouze na základě znalosti slov a kontextu. Neřešil jsem ani délku samohlásek, i když má fonologický význam (například pata a pátá v češtině), jak jsem zjistil, tak ve většině slovanských jazyků není délka vyjádřena explicitně diakritikou, jako v češtině, ale je dána přízvukem, který, jak už jsem zmínil, nelze jednoduše zjistit (např. v ruštině замок [zámъk] - hrad, замок [zamók] - zámek [22]).

Pravidla pro transkripci jsou ve tvaru:

A → B / C_D

Což značí, že řetězec A se přepíše na řetězec B, pokud mu bezprostředně předchází řetězec C a je bezprostředně následován řetězcem D. Řetězce C a D mohou být také prázdné nebo mohou obsahovat znak - pro mezeru. Dále mohou obsahovat konstrukci <E,F,G>, která značí, že zde může být řetězec E, nebo řetězec F, anebo řetězec G.

Dále mohou řetězce A, C a D obsahovat znaky zastupující skupiny znaků, čehož je využíváno především při spodobě znělosti, ke které dochází ve skupinách souhlásek, která obsahuje znělé i neznělé a je určena poslední souhláskou ve skupině, nebo na konci slov, kde dochází ke ztrátě znělosti. Pro tyto skupiny jsou použity znaky N pro neznělé, Z pro znělé, J pro znaky nepodléhající znělosti a S pro samohlásky. Všechny tyto skupiny musí být předem definovány pro každý jazyk.

Pokud řetězec A obsahuje znak pro skupinu znaků podléhající spodobě znělosti, řetězec B poté musí obsahovat buď znak A pro nahrazení jeho znělostním protějškem, nebo znak K pro ponechání původního znaku, pouze jeho převedení do fonetické abecedy. Na začátku je také potřeba definovat všechny znělostní páry.

Konkrétní pravidla pak vypadají následovně:

```
d => ɟ / _<i,í>
ą => ɔn / _<c, t, d>
szcz => ʃČ / _
N => A / _<Z,N>Z
```

Všechna pravidla jsou uložena v textových souborech, které si aplikace načte před výpočtem. Soubory jsou v následujícím formátu:

```
Z: výpis znělých souhlásek
N: výpis neznělých souhlásek
J: výpis jedinečných souhlásek
S: výpis samohlásek
A: výpis všech znělostních párů
Seznam pravidel
```

Pro ukázku soubor s pravidly pro polštinu vypadá následovně:

```
Z:b,d,g,h,w,z,ź,ż
N:c,ć,f,k,p,s,ś,t,ch
J:j,l,ł,m,n,ń,r
S:a,ą,e,ę,i,o,ó,u,y
A:d-t,t-d,f-v,w-f,k-g,g-k,h-x,s-z,z-s,p-b,b-p,ć-Ź,c-
D,ś-ʑ,ź-ɕ,ż-ʂ,ch-h
```

```
ę => ɛ / _<l, ł>
ę => ɛm / _<b, p>
ę => ɛɲ / _<j, dź>
...
```

Všechna pravidla pro všechny jazyky, která jsem vytvořil, jsou obsažena v příloze.

Zde je uveden příklad fonetické transkripce české věty:

```
Členské státy používají datový formát Gesmes

[Členskɛ statɪ pouʒivaji datovi format gɛsmɛs]
```

3.3 Podobnost znaků

Dalším mým úkolem bylo vytvořit systém mapování jazykově specifických znaků, ať už to jsou znaky s diakritikou nebo jiné znaky, které jsou si graficky podobné. Déle namapovat mezi sebou jiné abecedy, které byli použity, konkrétně tedy latinku, cyrilici a řeckou alfabetu. A nakonec také vytvořit systém pro mapování podobnosti fonémů.

3.3.1 Grafemická podobnost

Z grafemického hlediska, kde nejde o výslovnost, ale pouze o grafickou stránku znaků, se vyskytuje mezi znaky podobnost a to ať už se jedná o diakritiku jako v případě *a* a *á* nebo znaky jako například *d* a *đ*. Zde se samozřejmě při čtení textu stává, že pokud například člověk neznalý českého jazyka uvidí české *á*, nebude vědět, že ho má číst jako dlouhou samohlásku *a*, ale přečte ho nejspíš jako *a* ve svém jazyce. Je tedy potřeba pokusit se tuto podobnost zavést do měření a zjistit jaký to má vliv na výsledek měření.

Zde jsem tedy musel prostudovat všechny abecedy, projít všechny korpusy, vypsat všechny znaky v nich použité a tyto znaky sestavit do skupin a přiřadit jim nějakou podobnost. Tyto skupiny jsem sestavil na základě grafické podobnosti a například pro znak *a* vypadá takto:

a, á, à, â, ä, ǎ, ă, ā, ã, å, ǻ, ą, ą, æ, ǽ, ǣ

Některé znaky s diakritikou jsem musel vyřadit, jelikož byly sice systémem zobrazovány jako jeden, ale binárně byly reprezentovány jako dva znaky, což by narušovalo měření. Jednalo se naštěstí pouze o několik málo znaků.

Takto vzniklým skupinám znaků jsem poté přiřadil podobnostní skóre 1, 0,5 a 0 a vyzkoušel, jak bude ovlivněn výsledek měření. Do měřících metod tedy bylo nutno zanést podobnost jejich úpravou.

Dále bylo zapotřebí projít všechny znaky cyrilice a alfabety a namapovat je na znaky latinky.

Celá převodová tabulka všech znaků a skupiny znaků je obsažena v příloze.

3.3.2 Fonetická podobnost

Pro vytvoření podobnosti mezi znaky fonetické abecedy jsem zvolil jiný způsob výpočtu, a to pomocí fonetického rysu (distinctive feature), který popsali Noam Chomsky a Morris Halle v knize Sound Pattern of English (1968) [23]. Jde o přiřazení určitých vlastností, které jsou čistě binární, pro každý foném. Foném tedy má, nebo nemá danou vlastnost. Vlastnosti určují způsob a místo tvorby konkrétního fonému, postavení jazyka a tak dále. Vlastnosti jsou rozděleny do čtyř tříd, které dále pouze vypíši i se všemi vlastnostmi. Nebudu je více popisovat, jelikož jsou k tomu zapotřebí hlubší fonetické znalosti a je to nad rámec zadání práce. Hodnoty pro všechny fonémy jsem pouze převzal z [24][25]. Názvy vlastností ponechávám v anglickém jazyce, z důvodu využití pouze anglických zdrojů a složitosti překladu. Kompletní tabulka vlastností je obsažena v příloze a na přiloženém médiu.

Základní fonetická třída (Major feature class):

Syllabic, Consonantal, Approximant, Sonorant

Hlasivkový rys (Laryngeal feature):

Voiced, Spread glottis, Constricted glottis

Způsobový rys (Manner feature):

Continuant acoustic, Continuant articulation, Nasal, Lateral, Labial, Delayed release

Poziční rys (Place feature):

Labial, Round, Coronal, Anterior, Distributed, Dorsal, High, Low, Front, Back, Tense, Radical, Laryngeal

Skupina vlastní na příklad pro foném *m* vypadá následovně:

m - Consonantal, Sonorant, Voiced, Nasal, Labial, Anterior

Výpočet podobnosti fonémů je uskutečněn pomocí Jaccardova indexu, jedné z metod měření podobnosti, která je podrobně popsána v další kapitole. Metodě se předají množiny vlastností obou porovnávaných fonémů a na základě vztahu pro Jaccardův index (4.16) je vypočtena podobnost.

Například pro fonémy *m* a *ɱ*, které mají 7 vlastností společných a druhý foném má navíc jednu vlastnost, kterou první nemá, je podobnost vypočtena následovně:

$$PhoneticSim(\mathrm{m}, \mathrm{ɱ}) = \frac{7}{0 + 1 + 7} = 0{,}875$$

Podobnost pro nepodobné fonémy jako například *r* a *g* je následující (2 společné vlastnosti, 2 vlastnosti, které má první foném a druhý ne a 6 vlastností, které má druhý foném a první ne):

$$PhoneticSim(\mathrm{r}, \mathrm{g}) = \frac{2}{2 + 6 + 2} = 0{,}2$$

Samozřejmě z principu vyplývá, že mezi fonémy bude vždy nějaká podobnost, už například na základě vlastnosti *Consonantal*, kterou mají všechny souhlásky.

4. Metodika měření podobnosti

Ke způsobu měření podobnosti jsem došel na základě různých zdrojů o komparativní lingvistice a o práci s textovými řetězci. Následně jsem navrhl a otestoval několik metod. Z nich jsem nakonec vybral metody, které byly použitelné a dávaly smysluplné výsledky. Jedná se o metody pracující s textovými řetězci a metody pracující s množinami (v tomto případě množinami jednotlivých znaků a vyšších n-gramů). Některé z použitých metod jsem musel upravit a zanést do nich systém pro mapování nepodobných znaků, tedy grafemickou a fonetickou podobnost rozdílných znaků a fonémů. [1][2][3][6][26][27]

Výsledkem měřících metod je vždy reálné číslo v intervalu <0,1>, které se dá také interpretovat jako procentuelní vyjádření podobnosti. Při výsledku 0 je tedy podobnost nulová a při 1 je podobnost sto procentní, ale pouze v rámci použité metody. Zde je ale potřeba zdůraznit, že všechny výsledky jsou pouze relativní vůči použité měřící metodě. Nelze tedy říci, že pokud je naměřena podobnost mezi češtinou a slovenštinou 60 %, že by to znamenalo, že tyto jazyky jsou si absolutně podobné z 60 %, pouze pomocí použité metody a na vybraných datech byla naměřena tato podobnost. Výsledek měření je tedy potřeba interpretovat relativně mezi všemi měřenými jazyky a stanovit nějakou minimální mez podobnosti, proto bylo zapotřebí získat co nejvíc dat v co nejvíce jazycích.

Další podstatnou věcí při měření je, že zde vždy vzniká jistá podobnost i mezi nepodobnými jazyky. Ta může být způsobena jak stejnými jmény a názvy v textech, homonymií slov, slovy přejatými, anebo náhodnou znakovou shodou, která vyplývá z principu dále popsaných metod, jenž se snaží nelézt co nejvíce společného na porovnávaných textech, proto vždy dojde k naměření nějaké podobnosti, která vzniká už jen tím, že jazyky používají stejné písmo a navíc jsou slova složena vždy ze souhlásek a především samohlásek, kterých není mnoho, takže i v naprosto nepodobných slovech měřící algoritmy naleznou nějaký společný znak. Více se tomuto problému budu věnovat v následující kapitole společně s interpretací všech výsledků.

Dále podrobně popíši použité metody s ukázkou výsledků měření pro český jazyk na korpusu ASPAC, který nejvíce zastupuje přirozený jazyk, a v některých případech na slovníku.

4.1 Levenshteinova vzdálenost

Levenshteinova vzdálenost je metoda dynamického programování, která se snaží nelézt co nejmenší počet operací (inzerce, delece nebo substituce znaků) potřebných k převodu jednoho řetězce na druhý. Tato metoda se asi nejvíce ze všech použitých metod využívá v komparativní lingvistice pro měření podobnosti. Je vhodná pro porovnání jak slov, tak i vět. [28] Definována je následovně:

Mějme dva řetězce *A* a *B* o délce |*A*| a |*B*|, potom se jejich vzdálenost vypočte jako:

$$LevDist_{A,B}(|A|, |B|) \tag{4.1}$$

a platí, že:

$$LevDist_{A,B}(i,j) = \begin{cases} max(i,j), & pokud\ min(i,j) = 0 \\ min\begin{cases} LevDist_{A,B}(i-1,j) + 1 \\ LevDist_{A,B}(i,j-1) + 1 \\ LevDist_{A,B}(i-1,j-1) + [A_i \neq B_i] \end{cases} \end{cases} \tag{4.2}$$

Výpočet se provádí pomocí dynamického programování, kdy se vytvoří tabulka, jejíž řádky odpovídají znakům prvního řetězce a sloupce odpovídají znakům druhého řetězce, přičemž první sloupec a řádek neodpovídají žádnému znaku. Hodnoty v tabulce se vypočítají podle následujících výrazů. [11][12]

Mějme matici *M* a porovnávané řetězce *A* a *B*, potom hodnoty v matici vypočteme:

$$M_{i,1} = i - 1$$
$$M_{1,j} = j - 1$$
$$M_{i,j} = \begin{cases} M_{i-1,j-1} & pokud\ A_i = B_j \\ 1 + \min\left(M_{i-1,j-1}, M_{i-1,j}, M_{i,j-1}\right) \end{cases} \tag{4.3}$$

Výsledná vzdálenost je v posledním řádku a sloupci matice *M*.

Pro ukázku výpočtu uvádím příklad porovnání řetězců DOBRÝ a DOBŘE, kde výsledná vzdálenost slov je 2.

Tabulka 3. Příklad Levenshteinovy vzdálenosti

		D	O	B	R	Ý
	0	1	2	3	4	5
D	1	0	1	2	3	4
O	2	1	0	1	2	3
B	3	2	1	0	1	2
Ř	4	3	2	1	1	2
E	5	4	3	2	2	2

Výsledná podobnost se vypočítá jako výsledná vzdálenost řetězců *A* a *B* dělená délkou delšího ze slov a to celé odečteno od jedné.

$$SimD_{Lev}(A,B) = 1 - \frac{LevDist_{A,B}(|A|,|B|)}{\max(|A|,|B|)} \tag{4.4}$$

Děleno délkou delšího řetězce, protože je to maximální možná dosažitelná hodnota v případě změny prázdného řetězce na tento řetězec. A odečteno od jedné, jelikož pro dva identické řetězce je vzdálenost 0, takže výsledná podobnost musí být 1. A naopak pro dva naprosto odlišné řetězce je vzdálenost rovna maximu, což po dělení maximem je rovno 1 a následná podobnost je tedy 0.

Dále ukážu příklad měření podobnosti na dvou odstavcích z korpusu EUParl:

Český text:

> členské státy používají datový formát Gesmes v souladu s normami pro vzájemnou výměnu stanovenými Komisí (Eurostatem). Komise (Eurostat) zpřístupní podrobnou dokumentaci v souvislosti s těmito normami a poskytne pokyny k tomu, jak provádět tyto normy v souladu s požadavky tohoto nařízení.

Slovenský text:

> členské štáty používajú formát údajov Gesmes v súlade s normami pre výmenu, ktoré špecifikovala Komisia (Eurostat). Komisia (Eurostat) poskytuje podrobnú

dokumentáciu týkajúcu sa týchto noriem a tiež poskytuje usmernenia o spôsobe, ako implementovať tieto normy v súlade s požiadavkami tohto nariadenia.

Podobnost těchto dvou textů je vypočítána podle předchozího vztahu:

$$SimD_{Lev}(CS, SK) = 1 - \frac{131}{295} = 0{,}556$$

Texty mají tedy podobnost 55,6 % podle Levenshteinovy vzdálenosti.

V případě porovnání vět či odstavců tato metoda zároveň zohledňuje slovosled. To znamená, že už při záměně dvou slov dochází ke změně naměřené podobnosti. Pro příklad předvedu vliv na části z předchozího příkladu:

Členské státy používají datový formát Gesmes

Členské štáty používajú formát údajov Gesmes

Podobnost těchto dvou řetězců je 68 %. V případě záměny slov formát a údajov:

Členské státy používají datový formát Gesmes

Členské štáty používajú údajov formát Gesmes

Po této záměně se podobnost těchto řetězců zvýšila na 88 %. Je tedy jasně vidět, jak pořadí slov ve větě výrazně ovlivní výslednou podobnost. Zde pouhou záměnou dvou slov se výsledek změnil o 20 %.

4.1.1 Levenshteinova vzdálenost s mapováním znaků

Abych zavedl do měření systém pro mapování nepodobných znaků, musel jsem upravit Levenshteinovu vzdálenost a operaci substituce nahradit funkcí, která pouze neporovná, zda jsou znaky shodné, ale vrátí předem definovanou hodnotu podobnosti pro konkrétní dva rozdílné znaky.

Nechť je tato funkce definována jako:

$$CharSim(a,b) = x; x \in \mathbb{R} \wedge x \in <0,1> \tag{4.5}$$

Kde a a b jsou porovnávané znaky a x je předem definovaná tabulková hodnota podobnosti. Funkce je definována jak pro grafemickou, tak pro fonetickou podobnost.

Levenshteinova vzdálenost je následně upravena takto:

$$LevSDist_{A,B}(i,j) = \begin{cases} max(i,j), & pokud\ min(i,j) = 0 \\ min\begin{cases} LevSDist_{A,B}(i-1,j)+1 \\ LevSDist_{A,B}(i,j-1)+1 \\ LevSDist_{A,B}(i-1,j-1)+1-CharSim(a,b) \end{cases} \end{cases} \tag{4.6}$$

Podobnost je vypočtena jako v předchozím případě:

$$SimD_{LevS}(A,B) = 1 - \frac{LevSDist_{A,B}(|A|,|B|)}{\max(|A|,|B|)} \tag{4.7}$$

Ukázka výpočtu na předchozím příkladě pro řetězce DOBRÝ a DOBŘE:

Tabulka 4. Příklad Levenshteinovy vzdálenosti s mapováním znaků

		D	**O**	**B**	**R**	**Ý**
	0	1	2	3	4	5
D	1	0	1	2	3	4
O	2	1	0	1	2	3
B	3	2	1	0	1	2
Ř	4	3	2	1	0,5	1,5
E	5	4	3	2	1,5	1,5

Výsledná vzdálenost je v tomto případě 1,5 za využití podobnosti 0,5 mezi znaky *R* a *Ř*.

4.1.2 Damerau-Levenshteinova vzdálenost

Jedná se o rozšíření Levenshteinovy vzdálenosti. Přidává další operaci k základním třem, a to transpozici. Ta umožňuje, že pokud se dva sousední znaky z jednoho řetězce objeví v druhém, ale v opačném pořadí, jejich vzdálenost je počítána jako 1 (transpozice), namísto 2 (smazání, vložení). Do výpočtu jsem opět zavedl systém mapování grafemické a fonetické podobnosti. [28]

Vzdálenost je definována:

$$DamLevDist_{A,B}(|A|,|B|) \tag{4.8}$$

Pro výpočet znovu mějme matici M a porovnávané řetězce A a B, potom hodnoty v matici vypočteme podle následujícího vztahu:

$$M_{i,j} = \begin{cases} M_{i-1,j-1} & pokud\ A_i = B_j \\ 1 + \min\left(M_{i,j}, M_{i-2,j-2}\right) & pokud\ A_i = B_{j-1} \wedge A_{i-1} = B_j \\ 1 + \min\left(M_{i-1,j-1}, M_{i-1,j}, M_{i,j-1}\right) & \end{cases} \tag{4.9}$$

A podobnost je opět vypočtena vztahem:

$$SimD_{DamLev}(A,B) = 1 - \frac{DamLevDist_{A,B}(|A|,|B|)}{\max(|A|,|B|)} \tag{4.10}$$

Ukázku využití transpozice předvedu na slovech FORMAGGIO a FROMAGE (sýr v italštině a francouzštině). Kde výsledné hodnoty jsou:

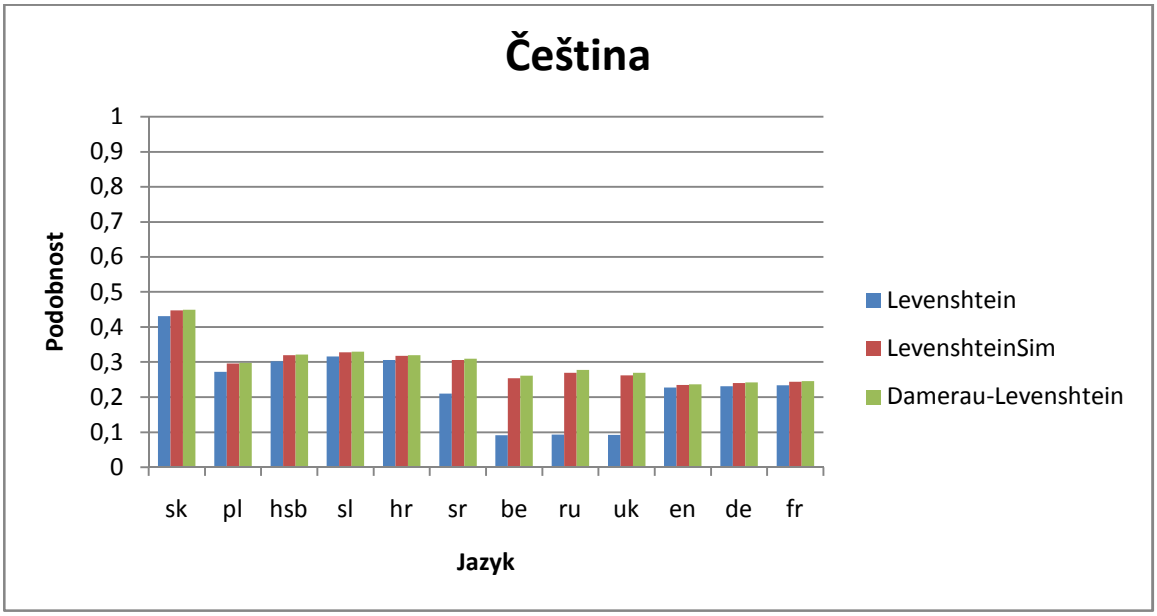
$$SimD_{Lev}(A,B) = 0{,}44$$

$$SimD_{DamLev}(A,B) = 0{,}55$$

Je jasně vidět, že díky transpozici *O* a *R* se podobnost slov zvýšila o 11 %. V důsledku toto vylepšení pomůže k lepšímu rozpoznání v případě menších jazykových variabilit. V ostatních případech se algoritmus chová stále stejně a poskytuje stejné výsledky.

4.1.3 Naměřené hodnoty

Ukázka naměřených dat pro všechny tři varianty Levenshteinovy vzdálenosti na korpusu ASPAC. Využito je porovnání češtiny s 9 slovanskými jazyky, z toho 4 psané cyrilicí, a 3 neslovanské jazyky. Konkrétně se jedná o slovenštinu (sk), polštinu (pl), hornolužickou srbštinu (hsb), slovinštinu (sl), chorvatštinu (hr), srbštinu (sr), běloruštinu (be), ruštinu (ru), ukrajinštinu (uk), angličtinu (en), němčinu (de) a francouzštinu (fr).



Graf 1. Levenshteinova vzdálenost

Z grafu je vidět několik věcí. Zaprvé naměřená podobnost je nejvyšší se slovenštinou, kolem 45 %, a dále menší se západoslovanskými a jihoslovanskými jazyky, mírně přes 30 %. S východoslovanskými jazyky je podobnost pod 30 % a s neslovanskými pod 25 %. Hranice 25 % se tak dá považovat za mez minimální podobnosti.

Zadruhé rozdíl mezi jednotlivými variantami Levenshteinovy vzdálenosti. U jazyků psaných latinkou je vidět malý přírůstek při využití mapování znakové podobnosti a další malý přírůstek při započítání transpozicí Damerau-Levenshteinovy vzdálenosti. U jazyků píšících cyrilicí je podobnost základní Levenshteinovy vzdálenosti bez podobnosti znaků mnohem nižší, prakticky na úrovni shody mezer a názvů psaných latinkou.

4.2 Jaro-Winklerova vzdálenost

Informatická a statistická metoda pro měření podobnosti mezi řetězci. Skládá se z Jarovy vzdálenosti a Winklerovy vzdálenosti, která pouze přidává další skóre k Jarově vzdálenosti. Vzdálenost je reálné číslo z intervalu od 0 do 1. Čím vyšší vzdálenost je, tím jsou si řetězce podobnější. Vzdálenost 1 tedy znamená stoprocentní shodu. Tato metoda je z principu svého algoritmu vhodná pro porovnání především kratších řetězců a nejlépe jednotlivých slov. [29]

Do metod bylo implementováno mapování podobnosti znaků a fonémů.

4.2.1 Jarova vzdálenost

Jarova vzdálenost, na rozdíl od Levenshteinovy vzdálenosti, nevyužívá dynamického programování a vyplňování matice, ale pouze prochází oba porovnávané řetězce a vyhledává shody znaků a transpozice znaků.

Definice pro výpočet Jarovy vzdálenosti pro dva řetězce A a B je definována následovně:

$$JaroDist(A,B) = \begin{cases} 0 & pokud\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|A|} + \frac{m}{|B|} + \frac{m-t}{m}\right) \end{cases} \tag{4.11}$$

Kde *m* je počet shodných znaků mezi řetězci a *t* je počet transpozic. Shoda znaků je hledána pouze do určité vzdálenosti *MaxD(A,B)*:

Vztahy pro výpočet jsou definovány následovně:

$$MaxD(A,B) = \left\lfloor \frac{\max(|A|,|B|)}{2} \right\rfloor - 1 \tag{4.12}$$

$$m = \sum_{i=1}^{|A|} A_i = B_j; i - MaxD(A,B) \leq j \leq i + MaxD(A,B) \tag{4.13}$$

$$t = \sum_{i=1}^{\min(|A|,|B|)-1} A_i = B_{i+1} \wedge A_{i+1} = B_i \tag{4.14}$$

Pro příklad uvedu výpočet podobnosti opět na slovech DOBRÝ a DOBŘE.

$$MaxD(A,B) = \left\lfloor \frac{5}{2} \right\rfloor - 1 = 1$$

$$m = 3$$

$$t = 0$$

Maximální vzdálenost pro hledání shodných znaků je tedy 1 a počet shodných znaků v této vzdálenosti je 3. Transpozice zde není žádná. Výsledná podobnost tedy je vypočtena následovně:

$$JaroDist(A,B) = \frac{1}{3}\left(\frac{3}{5} + \frac{3}{5} + \frac{3-0}{3}\right) = 0{,}73$$

V případě, že započteme i grafemickou podobnost 0,5 pro znaky *R* a *Ř*. Změní se pouze hodnota *m* a výsledná podobnost. Výsledný výpočet vypadá takto:

$$m = 3{,}5$$

$$JaroDist(A,B) = \frac{1}{3}\left(\frac{3{,}5}{5} + \frac{3{,}5}{5} + \frac{3{,}5-0}{3{,}5}\right) = 0{,}8$$

Podobnost slov vypočtená pomocí Jarrovy vzdálenosti je tedy 73 %, respektive 80 % při využití grafemické podobnosti znaků.

Při měření podobnosti delších řetězců se zvyšuje vzdálenost pro hledání shody znaků *MaxD(A,B)*, a tím pádem dochází ke shodě znaků mezi různými slovy a výsledná podobnost je tak velmi vysoká.

Pro následující dvě věty je podobnost s využitím grafemické podobnosti znaků 96 % a to i v případě záměny slov *formát* a *údajov*.

```
Členské státy používají datový formát Gesmes

Členské štáty používajú formát údajov Gesmes
```

4.2.2 Winklerova vzdálenost

Winklerova vzdálenost pouze přidává k Jarově vzdálenosti další skóre za shodu v začátcích slov (prefixu) a to pouze pokud Jarova vzdálenost přesáhne určitý práh *tzv. BoostThreshold* – b_t, který je nejčastěji nastaven na 0,7.

Tuto metodu je možné využít pouze na porovnávání slov, u delších řetězců postrádá smysl. Má smysl především u flektivních jazyků, kde dochází k ohýbání konců slov, přičemž základ slova zůstává stejný, a z části u aglutinačních jazyků, kde dochází k přidávání přípon.

Vztah pro Winklerovu vzdálenost je následující:

$$WinklerDist(A,B) = \begin{cases} JaroDist(A,B) & pokud\ JaroDist(A,B) < b_t \\ JaroDist(A,B) + & (l * p * (1 - JaroDist(A,B))) \end{cases} \quad (4.15)$$

Kde p je konstanta zajišťující, aby výsledná vzdálenost nepřesáhla hodnotu 1. Nejčastěji je p = 0,1 a maximálně může nabývat hodnoty 0,25, aby nebyla celkově přesažena vzdálenost 1. Hodnota l je délka společného prefixu, maximálně však pro 4 znaky. A b_t je již zmíněný práh *BoostThreshold*.

Příklad pro DOBŘE a DOBRÝ:

$$JaroDist(A,B) = \frac{1}{3}\left(\frac{3}{5} + \frac{3}{5} + \frac{3-0}{3}\right) = 0{,}73$$

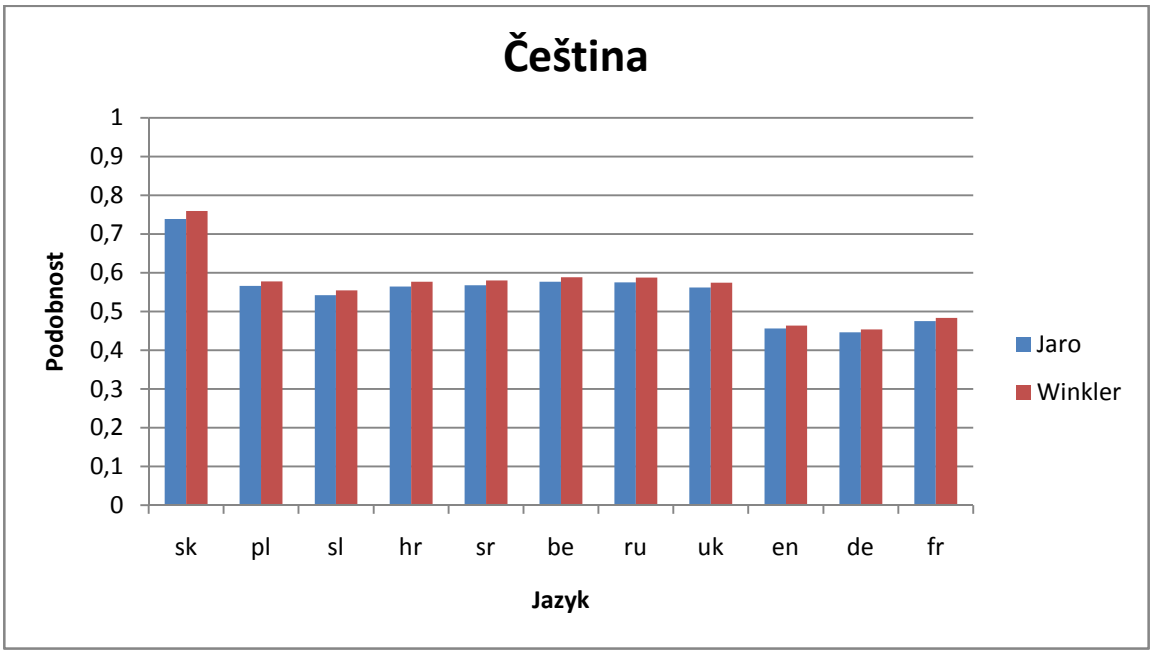
$$p = 0{,}1$$

$$l = 3$$

$$WinklerDist(A,B) = 0{,}73 + \left(3 * 0{,}1 * (1 - 0{,}73)\right) = 0{,}811$$

4.2.3 Naměřené hodnoty

Ukázka naměřených hodnot pro obě vzdálenosti. Měřeno na slovníku, protože obě vzdálenosti jsou vhodnější pro slova.



Graf 2. Jaro-Winklerova vzdálenost

Z grafu je vidět podobnost s hodnotami naměřenými Levenshteinovou vzdáleností, avšak s tím rozdílem, že mez minimální podobnosti je zde kolem 45 %.

Rozdíl mezi Jarovou vzdálenosti a Winklerovou vzdáleností je v podstatě minimální, což je ale z definice Winklerovy vzdálenosti očekávatelné.

4.3 Metody pracující s množinami

Obecně se jedná o statistické metody pro porovnávání podobnosti dvou množin. V tomto případě se pracuje s množinami jednotlivých využitých znaků a vyšších n-gramů. Metody k výpočtu podobnosti využívají množinové operace průniku a rozdílu porovnávaných množin. V případě měření podobnosti pomocí těchto metod není zohledněno pořadí znaků v textu, jedná se o neuspořádané množiny. Může tedy při měření dojít k 100% podobnosti i na zcela odlišných textech, které pouze obsahují stejné znaky. Metody jsou vhodné spíše pro slova a kratší texty. [30]

V případě těchto metod není využito mapování grafemické a fonetické podobnost znaků a fonémů.

4.3.1 Jaccardův index

Mějme množiny n-gramů A a B, které obsahují unikátní n-gramy dvou porovnávaných řetězců. Vztah pro výpočet je následující:

$$Sim_{Jaccard}(A,B) = \frac{|A \cap B|}{|A - B| + |B - A| + |A \cap B|} \tag{4.16}$$

Pro příklad výpočtu s jednotlivými znaky (unigramy) použiji věty využité dříve:

`Členské státy používají datový formát Gesmes`

`Členské štáty používajú formát údajov Gesmes`

Množiny A a B jsou následující:

$$A = \{\text{č}, l, e, n, s, k, \text{é}, t, \text{á}, y, p, o, u, \text{ž}, \text{í}, v, a, j, d, \text{ý}, f, r, m, g, \ \}$$

$$B = \{\text{č}, l, e, n, s, k, \text{é}, \text{š}, t, \text{á}, y, p, o, u, \text{ž}, \text{í}, v, a, j, \text{ú}, f, r, m, d, g, \ \}$$

A tedy výpočet celkové podobnosti je následující:

$$Sim_{Jaccard}(A,B) = \frac{24}{1 + 2 + 24} = 0{,}888$$

4.3.2 Sørensen-Diceův koeficient

Opět mějme množiny n-gramů A a B, jako v předchozím případě. Vztah pro výpočet je následující:

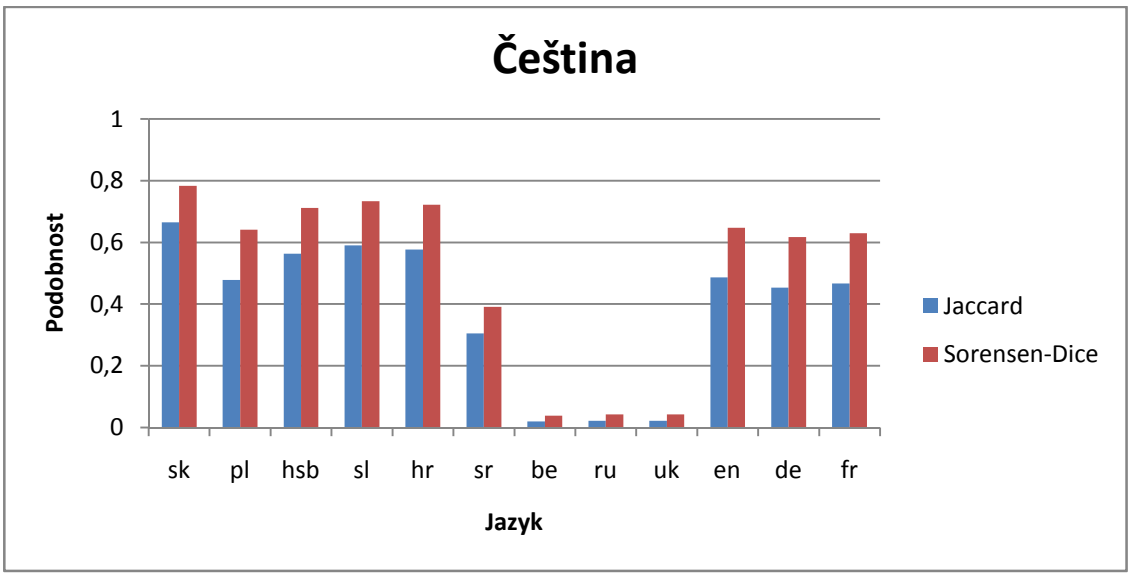
$$Sim_{Sørensen-Dice}(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{4.17}$$

Příklad výpočtu podobný, jako v případě Jaccardova indexu:

$$Sim_{Sørensen-Dice}(A,B) = \frac{2*24}{25+26} = 0{,}94$$

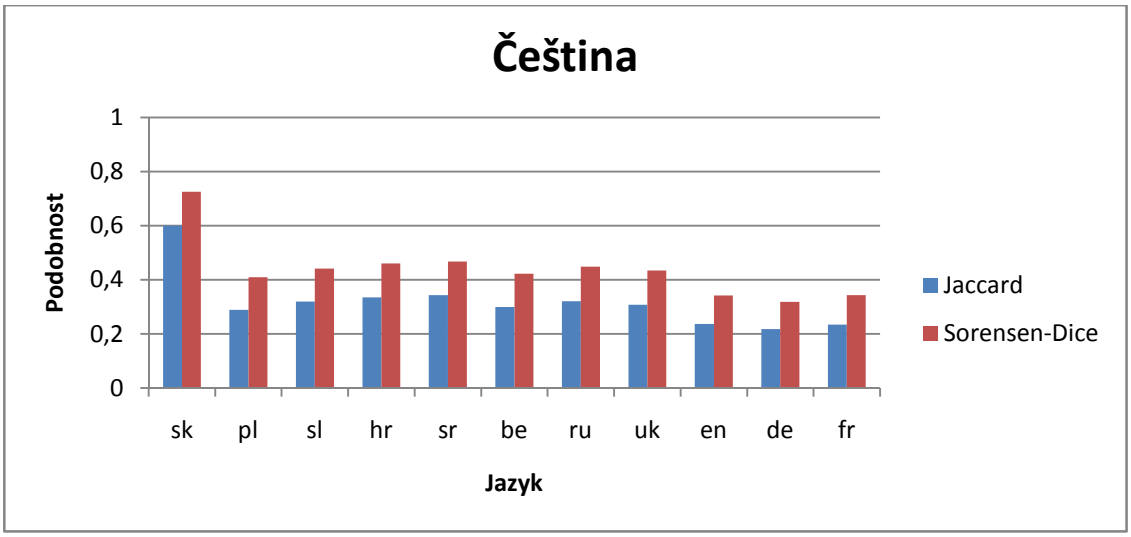
4.3.3 Naměřená podobnost

Hodnoty naměřené pro obě metody na korpusu ASPAC.



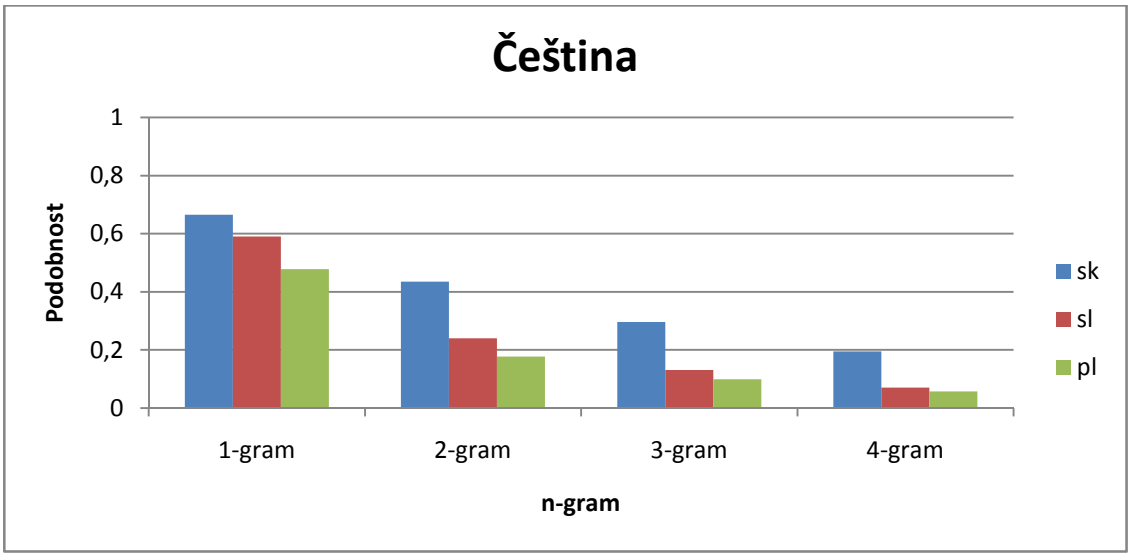
Graf 3. Množinové metody – ASPAC

Z grafu je patrná naprosto minimální podobnost pro jazyky psané cyrilicí, s výjimkou srbštiny, která používá jak latinku, tak cyrilici. U ostatních jazyků je podobnost vysoká. Je to z důvodu dlouhých textů korpusu ASPAC. Lepší příklad bude na kratších řetězcích ze slovníku.



Graf 4. Množinové metody – slovník

Na slovech ze slovníku je mnohem lépe vidět vysoká podobnost se slovenštinou a dále menší podobnost se slovanskými jazyky oproti třem neslovanským. Jazyky psané cyrilicí jsou zde přepsány do latinky, proto mají vyšší podobnost než v předchozím případě.



Graf 5. Jaccardův index – n-gramy

V tomto grafu je ukázán vliv na podobnost při použití různých stupňů n-gramu. Je vidět, že se stoupajícím stupněm, klesá podobnost, ale především vyniká rozdíl mezi jazyky.

4.4 Daitch-Mokotoff Soundex

Daitch-Mokotoff Soundex je fonetický algoritmus, který vytvořili Randy Daitch a Gary Mokotoff na základně původního algoritmu Soundex Roberta Russella, který byl vytvořen pro indexaci jmen podle anglické výslovnosti. Tuto metodu jsem vybral z toho důvodu, že výstupem je přesně definovaná sekvence spojující podobnosti mezi slovy a především se snaží jistým způsobem pracovat na fonetické úrovni. Dále popíši princip tohoto algoritmu a způsob měření podobnosti. [31]

Algoritmus původního Soundexu se snaží zakódovat homofonní souhlásky do společných skupin. Samohlásky jsou ignorovány, kromě samohlásky na začátku slova. Výsledkem je kód začínající prvním písmenem kódovaného slova a skupinou číslic, reprezentující souhlásky v kódovaném slově. Pro převod je dáno několik konkrétních pravidel.

Převod vypadá například následovně:

```
Robert => R163
Rubin => R150
Ashcroft => A261
```

Problémem tohoto algoritmu je, že reflektuje pouze anglickou výslovnost, a proto byl vytvořen Daitch-Mokotoff Soundex (dále D-M Soundex), který se snaží rozšířit výslovnost i na další evropské jazyky. Výsledný kód je zde tvořen šesticiferným číslem, kde každé číslo opět reprezentuje nějakou skupinu souhlásek.

Převod probíhá za využití 69 pravidel, převzatých z [31], které určují na jaké číslo se znak, nebo skupina znaků převede, a to ještě podle toho, zda se nachází na začátku řetězce, před samohláskou, a nebo jinde. Vstupní řetězec je procházen od počátku znak po znaku a porovnáván s pravidly. Pokud je řetězec dlouhý tak, že by byl převeden na více než šest číslic, jsou další znaky ignorovány. Z tohoto důvodu je metoda vhodná pouze na použití u slov. V případě, že je kratší, výsledný kód je doplněn nulami do celkové délky šesti cifer.

Pravidla jsou ve tvaru:

Tabulka 5. D-M Soundex - ukázka pravidel

Znaky vstupního řetězce	Převod		
	Začátek řetězce	Před samohláskou	Jinde
ds, dsh, dsz	4	4	4
dz, dzh, dzs	4	4	4
d, dt	3	3	3
chs	5	54	54
mn	-	66	66

Převod vypadá následovně:

```
Robert => 979300
Rubin => 976000
Ashcroft => 044973
```

Porovnání dvou řetězců probíhá tak, že jsou nejdříve převedeny na kód D-M Soundexu a poté tyto dva kódy porovnány Levenshteinovým algoritmem.

Funkce pro převod vstupního řetězce na Soundexový kód je definována jako:

$$DM\_Soundex(A) \tag{4.18}$$

A podobnost je poté definována:

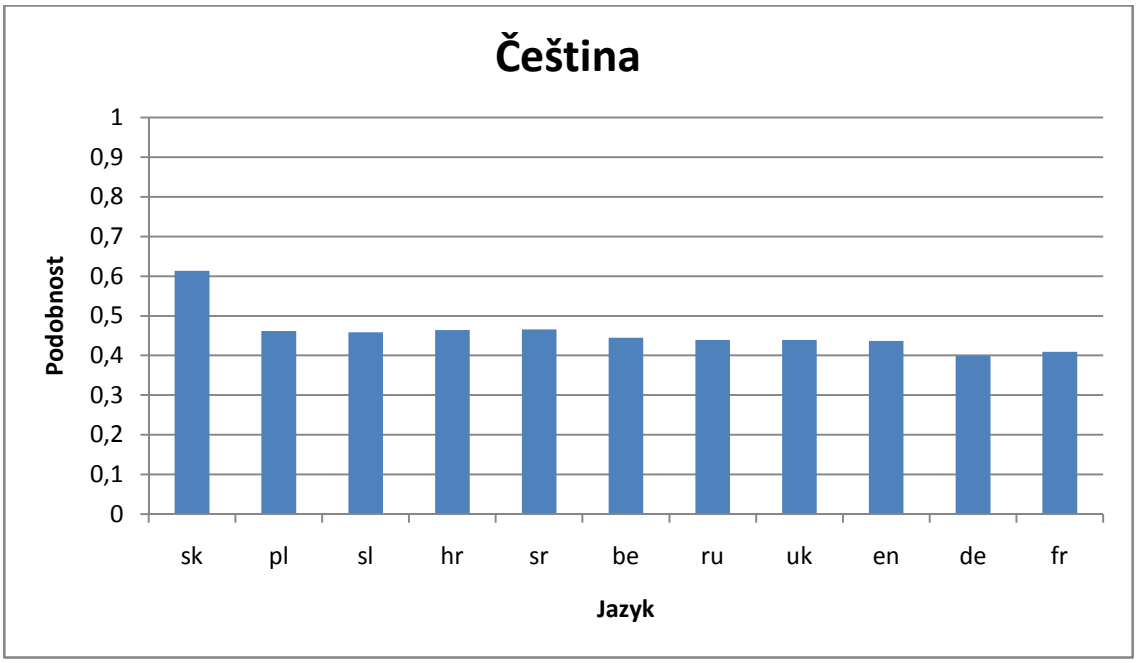
$$Sim_{D-M\ Soundex}(A, B) = Sim_{Lev}(DM\_Soundex(A), DM\_Soundex(B)) \tag{4.19}$$

Příklad podobnosti slov *Robert* a *Rubin*:

$$Sim_{D-M\ Soundex}(Robert, Rubin) = 0{,}666$$

Samozřejmě nelze jednou skupinou pravidel obsáhnout výslovnost všech jazyků, proto tuto metodu nelze považovat za úplné fonetické porovnání. Vybral jsem jí pro měření podobnosti hlavně z důvodu seskupování znaků do skupin, což ve výsledku může pomoci při měření podobnosti.

Protože je tato metoda vhodná jen pro porovnání slov, tak ukážu hodnoty naměřené na slovníku.



Graf 6. D-M Soundex

Z grafu je opět jasně vidět zvýšená podobnost se slovenštinou. Pro ostatní jazyky je výsledek těžko interpretovatelný. Celkem je vidět že pro podrobnější porovnání je tato metoda nevhodná, podobnost nalezne jen u velmi podobných jazyků jako je čeština a slovenština.

4.5 Měření podobnosti slov ve větách

Jelikož jsou všechny paralelní korpusy zarovnány na věty či odstavce, bylo vhodné vytvořit metodu, která se pokusí o porovnání na úrovni slov v těchto větách. Problémem je, že nelze bez dalších informací přesně říci, která slova k sobě patří. Proto jsem vytvořil metodu, která prochází obě porovnávané věty a za využití předchozích metod se snaží najít dvojice slov s podobností přesahující zadaný práh, tyto dvojice jsou uznány za podobné a vyřazeny z dalšího porovnávání. Celková podobnost je vypočtena jako poměr počtu rozpoznaných dvojic k celkovému počtu slov.

Pro začátek je definována funkce *w(A)*, která rozdělí řetězec *A* podle mezer na množinu slov a množina *W*, která obsahuje rozpoznaná slova. Výpočet podobnosti pro řetězce *A* a *B* je poté definován následovně:

$$W = \{a \in w(A);\ \forall\ b\ \in w(B) \colon Sim_f(a,b) > threshold\ \}$$

$$Sim_{word}(A,B) = \frac{|W|}{\min(|w(A)|,|w(B)|)} \tag{4.20}$$

Pro porovnání slov jsou použity metody Damerau-Levenshteinova vzdálenost, Jaro-Winklerova vzdálenost a D-M Soundex, které jsou při výpočtu dosazeny za funkci Sim_f. Práh *threshold* je nastaven pro každou metodu tak, aby byl vyšší než náhodná znaková shoda dané metody a odpovídal minimální podobnosti.

Pro ukázku výpočtu opět použiji věty:

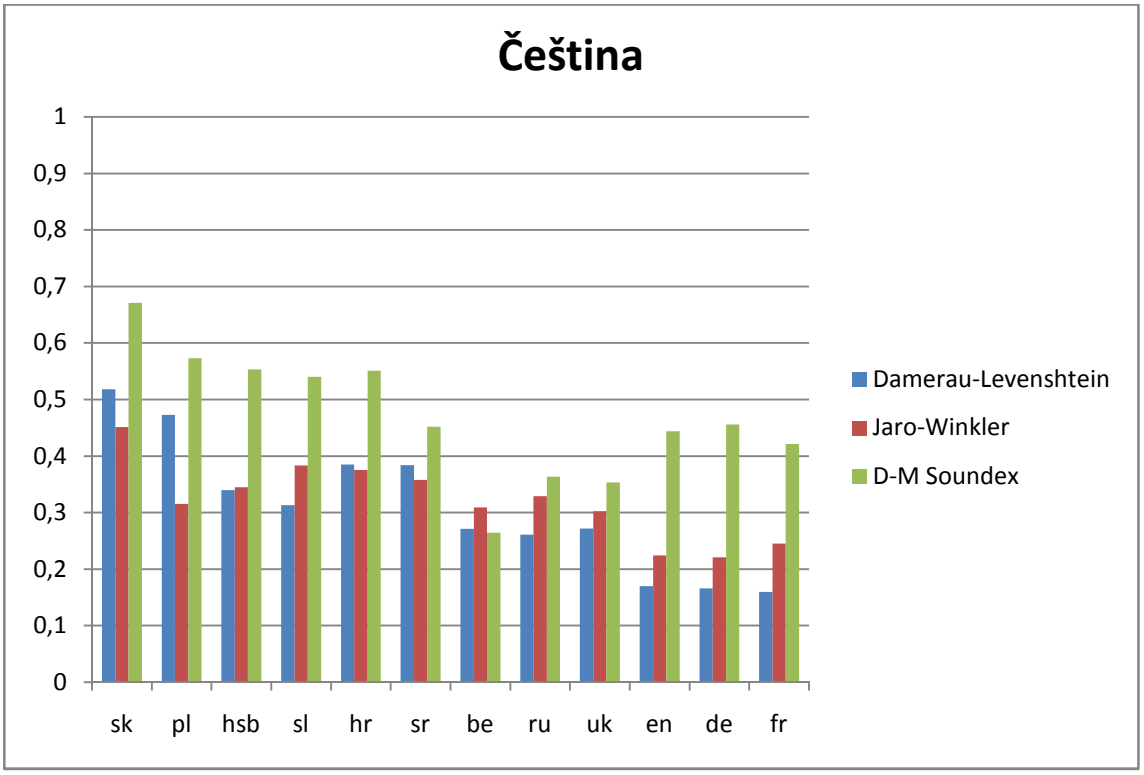
```
členské státy používají datový formát Gesmes
```

```
členské štáty používajú formát údajov Gesmes
```

Při výpočtu za využití Damerau-Levenshteinovy vzdálenosti a prahu 0,5 je rozpoznáno 5 z 6 slov. Výsledná podobnost je tedy 83,3 %. Nebyla přiřazena pouze slova *údajov* a *datový*, ostatní slova byla správně přiřazena.

Jako nejvhodnější hodnoty pro práh se po testování osvědčili hodnoty 0,5 pro Damerau-Levenshteinovu vzdálenost a 0,7 pro Jaro-Winklerovu vzdálenost a D-M Soundex.

Je zřejmé, že kratší nepodobná slova o několika znacích, například *ale* a *hle*, budou vyhodnoceny jako podobná a započítány do celkového výsledku. Statisticky se ale tento problém ztrácí a tvoří pouze malou homonymní podobnost.



Graf 7. Podobnost slov ve větách

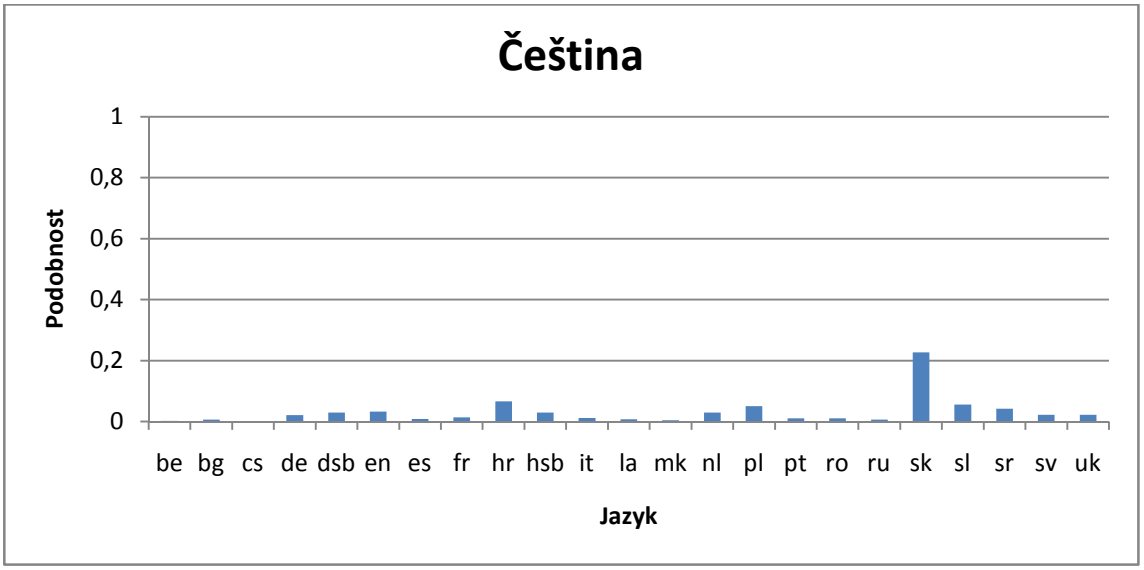
Z tohoto grafu lze vyčíst, že kromě slovenštiny se i zvýšila podobnost s polštinou za využití Damerau-Levenshteinovy vzdálenosti. Předchozí metody takovouto podobnost neodhalili, což může být způsobeno odlišným slovosledem v polštině, který tato metoda nezohledňuje.

4.6 Textově nezávislé porovnání

Pro textově nezávislé porovnání (nejsou porovnávány texty o stejném významu) jsem si nechal pro každý jazyk vypsat všechna nalezená slova s jejich četností, slova jsem seřadil podle četnosti a vybral prvních 20 tisíc slov. Poté jsem s těmito množinami slov provedl průnik mezi jazyky, tedy pro každé slovo z jednoho jazyka se pokusil vyhledat stejné slovo v jiném jazyce. Výsledná podobnost je vypočtena jako poměr stejných slov k celkovému počtu slov, tedy 20 tisícům. Pro porovnání slov mezi jazyky s jiným písmem je uplatněno mapování znaků na latinku.

U této metody samozřejmě dochází k vysoké možnosti rozpoznání homonymních slov, jelikož není ničím zaručena sémantická vazba. Také je problém v různých tvarech slov především u flektivních a aglutinačních jazyků. Porovnávaná slova nejsou v základním tvaru, ale v různých tvarech, tak jak se objevují v textu.

U některých jazyků, kde nebylo k dispozici dostatek textů, jako například dolnolužická srbština, nemusí být k dispozici ani porovnávaných 20 tisíc slov, proto se při porovnání počítá s nižší velikostí slovníku a je potřeba počítat s tím, že výsledek může být zkreslený oproti porovnání ostatních jazyků.



Graf 8. Textově nezávislé porovnání

Hodnoty podobnosti této metody jsou velmi nízké, takže více vynikne případná podobnost, jako v případě slovenštiny přes 20 %. Dále je vidět lehce vyšší podobnost u dalších slovanských jazyků, jinak se podobnost blíží nule.

5. Výsledky měření podobnosti

Pro velké množství výsledků měření podobnosti zde zmíním pouze některé, na kterých je dobře vidět fungování metod. Všechny naměřené výsledky jsou k dispozici na přiloženém médiu a to jak ve formě souborů s výsledky z měření, tak vyhodnocené výsledky v Excelu ve formě tabulek n krát n pro podobnost mezi všemi měřenými jazyky daného korpusu. Dvě tabulky z grafemického a fonetického porovnání korpusu ASPAC jsou pro ukázku v příloze.

Výsledná hodnota podobnosti pro každou metodu je aritmetickým průměrem ze všech naměřených hodnot. Měření podobnosti bylo potřeba provézt na dostatečném množství dat, jelikož se v některých korpusech občas objevovaly podobné texty obsahující jména, výpisy produktů, cizí názvy a další řetězce, které zvyšovaly celkovou podobnost. Při dostatečném množství porovnání se ale tyto odchylky ztratily.

Při měření nebylo z časových důvodů možné nechat změřit podobnost na všech datech (jedná se cca o 160 GB dat). Navíc většina korpusů je značně asymetrická a například v korpusu EUBookshop jsou 2 GB dat pro angličtinu a francouzštinu, zatímco pro angličtinu a islandštinu pouze 3 MB dat. U těch nejméně obvyklých jazyků jako velština, skotská gaelština a podobně se dokonce jedná pouze o několik kilobytů dat. Zvolil jsem proto krok omezit měření mezi každou dvojicí jazyků v každém korpusu na maximálně 5000 porovnání. Jak jsem při pozorování zjistil, už po několika stovkách porovnání výsledná hodnota konverguje na určitou hodnotu. Celkem bylo provedeno okolo 22 miliónů porovnání.

Dále popíši jakým způsobem interpretovat naměřené výsledky, některé výsledky měření podobnosti a několik informací k vytvořeným aplikacím a výpočetní náročnosti měření.

5.1 Interpretace výsledků

Jak již bylo zmíněno na začátku předchozí kapitoly, výsledky měření je potřeba interpretovat relativně vůči použité metodě. To znamená, že je potřeba se podívat jaké výsledky vrací metoda u všech porovnávaných jazyků a odhadnout, kde se nachází mez náhodné podobnosti (náhodná znaková shoda, homonymie slov, shodné názvy atd.) a kde už je tato mez překročena a jedná se o skutečnou podobnost. Například při porovnání češtiny je vhodné se podívat na naměřenou podobnost například se španělštinou, francouzštinou, švédštinou a podobnými jazyky, u kterých se nepředpokládá žádná podobnost, a na základě naměřených hodnot pro tyto jazyky dále pohlížet na další jazyky u kterých se očekává nějaká podobnost, pro češtinu tedy u slovanských jazyků.

Dále v této kapitole popíšu jaká je náhodná znaková shoda u jednotlivých metod a dále výsledky z grafemického a fonetického měření podobnosti. Vždy vyberu jeden jazyk a k němu naměřené hodnoty podobnosti s několika dalšími jazyky. Výsledky jsou zobrazeny pro přehlednost v grafech, kde mnohem lépe vyniknou rozdíly mezi jednotlivými jazyky. Na ose x jsou vždy zobrazeny porovnávané jazyky a na ose y podobnost s hlavním porovnávaným jazykem.

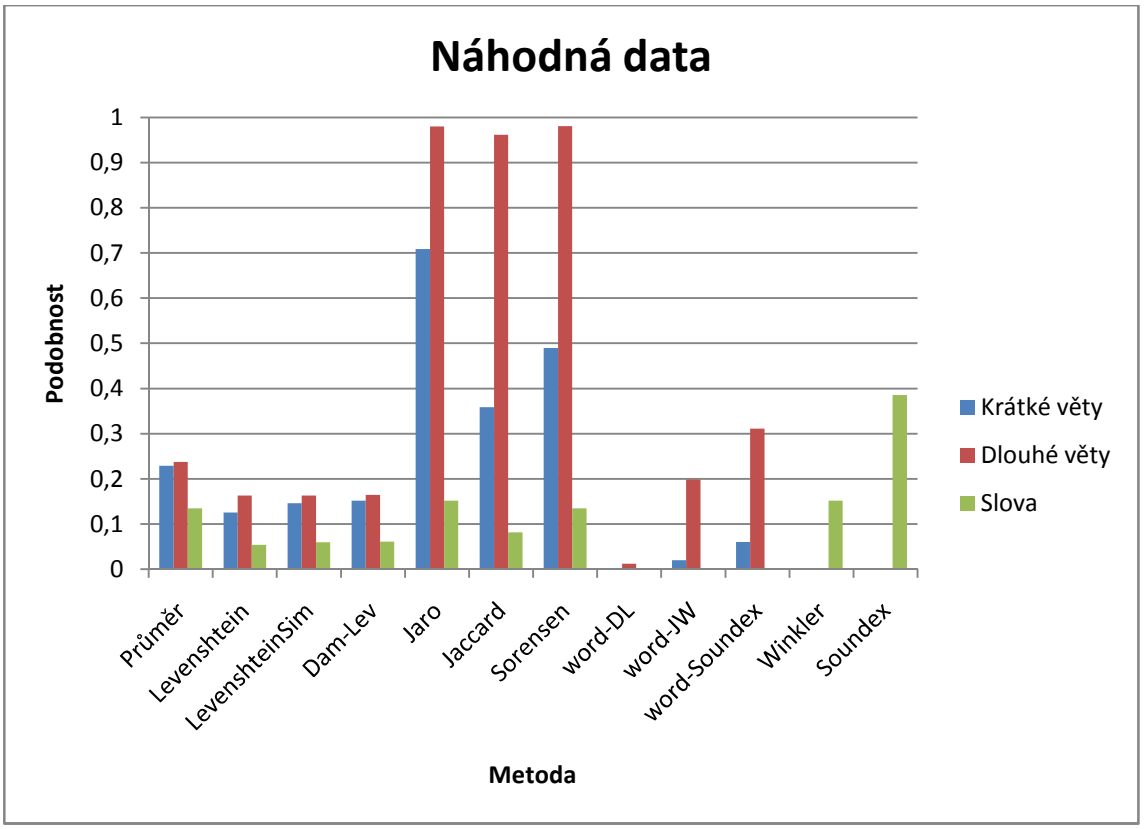
Pro označení jazyků v práci jsou použity mezinárodní kódy dle standardů ISO 639-1 a ISO 639-2 [32]. Seznam všech jazyků a jejich kódů je uveden v příloze.

5.1.1 Náhodná znaková shoda

Jelikož metody pro porovnání pracují s omezenou množinou znaků, vždy dojde k naměření nějaké podobnosti. V některých případech se jedná o stejná slova, jako jsou názvy, jména a podobně, jindy může jít o slova přejatá, anebo o čistou homonymii slov.

V největší míře se ale jedná o náhodnou znakovou shodu, jelikož metody jsou navrženy tak, aby hledaly co nejvíce podobností. Proto je vždy nalezen nějaký společný znak, i když se nemusí jednat o podobnost.

Z tohoto důvodu jsem vygeneroval náhodná data a otestoval na nich všechny metody, aby bylo jasné, jak velká je míra náhodné znakové shody u každé z metod. Vygeneroval jsem tři druhy textů. Krátké věty (3 až 10 slov), dlouhé věty (20 až 40 slov) a jednotlivá slova. Slova tvoří 2 až 12 znaků, které jsou náhodně vybrány z 26 znaků základní latinky a z 10 znaků s diakritikou.



Graf 9. Podobnost náhodných dat

Z grafu je vidět, že nejmenší znaková shoda vzniká při porovnání slov, zatímco nejvyšší u dlouhých vět.

Co se týče metod, tak nejmenší znaková shoda je u metod porovnávající slova ve větách a to především na krátkých větách, je to způsobeno nastaveným prahem pro rozpoznání slov. Další metodou s nejmenší znakovou shodou je Levenshteinova vzdálenost, kde je pro věty podobnost kolem 15 % a pro slova kolem 5 %.

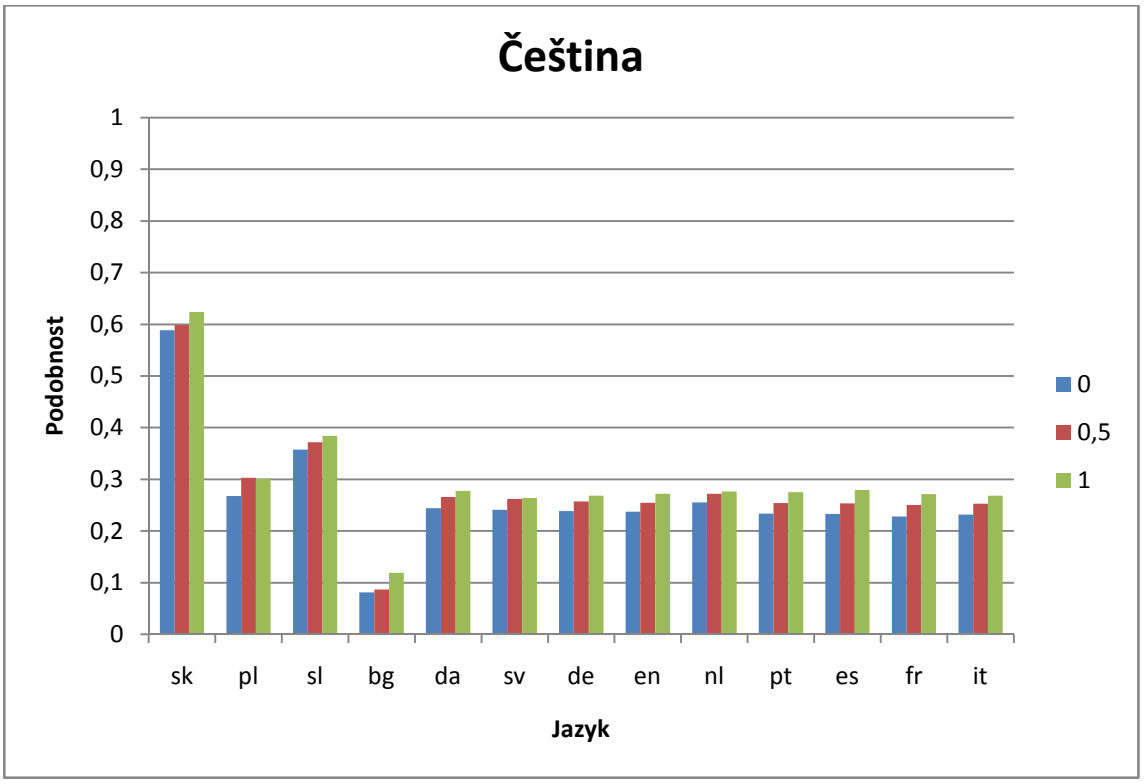
Dále je z grafu vidět jak nevhodné jsou množinové metody a Jarova vzdálenost pro věty. Při dlouhých větách je podobnost náhodných vět téměř 100 %, zatímco pro slova v rozmezí 9 % až 15 %.

D-M Soundex, který je využit jen na slova, má shodu na náhodných datech ke 40 %. Z toho se dá usoudit, že tato metoda není příliš vhodná pro měření podobnosti.

5.1.2 Grafemická podobnost

Naměřené výsledky pro češtinu byly už ukázány pro každou z použitých metod, proto dále ukážu pouze pár dalších výsledků s některými vlivy a naměřené hodnoty pro některé další jazyky. Jako nejvhodnější se pro měření ukázala Levenshteinova vzdálenost, která je také v komparativní lingvistice nejvíce používána, proto budu prezentovat výsledky na této metodě.

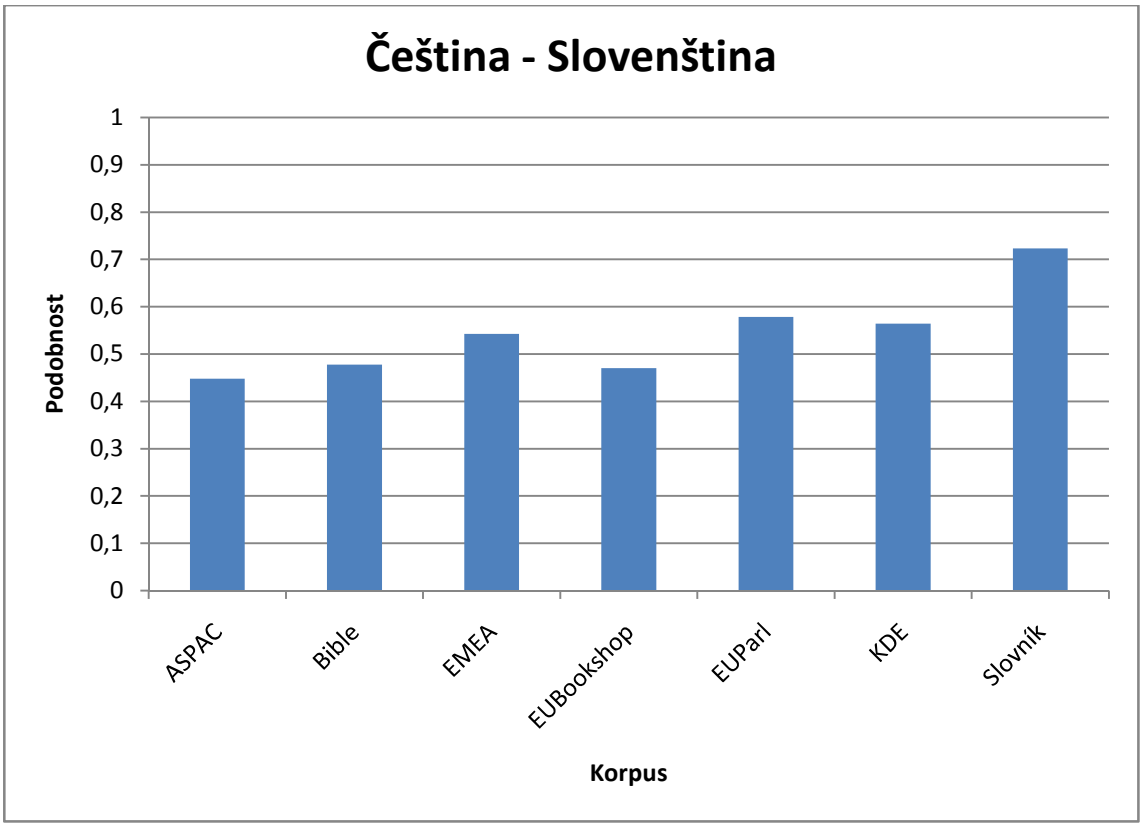
Jako první je vhodné ukázat vliv grafemické podobnosti znaků. Vybrán je paralelní korpus EUParl a Levenshteinova vzdálenost.



Graf 10. Vliv mapování grafemické podobnosti – Levenshteinova vzdálenost

Jak je z grafu patrné, celkem je vliv zanesené podobnosti do Levenshteinovy vzdálenosti v řádu několika procent. Na výsledek to tedy nemá až takový vliv, jaký byl očekáván.

Dále ukážu, jak obsah korpusu může ovlivnit naměřený výsledek. Pro ukázku jsem vybral podobnost mezi češtinou a slovenštinou opět za využití Levenshteinovy vzdálenosti.

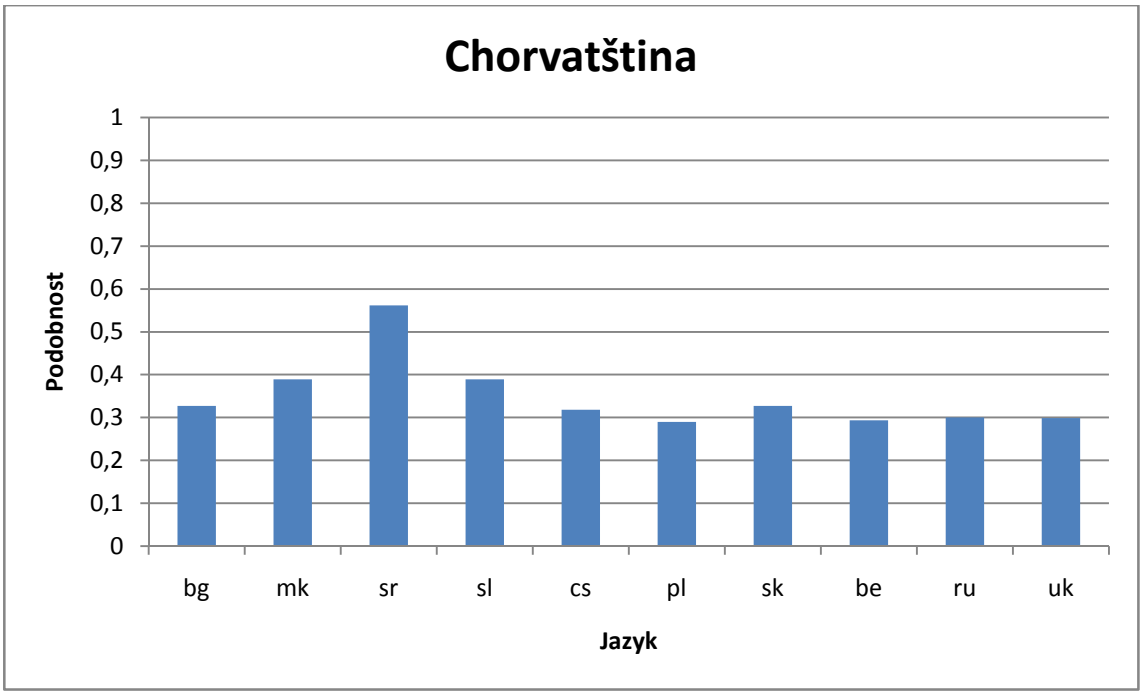


Graf 11. Vliv obsahu textu na výsledek měření – Levenshteinova vzdálenost

Výsledek jasně ukazuje, že korpusy, kde je použit nejpřirozenější jazyk, vykazují menší podobnost. Zatímco korpusy, kde je spíše více názvů a jmen, jako jsou korpusy EMEA, EUParl a KDE, vykazují o něco větší podobnost.

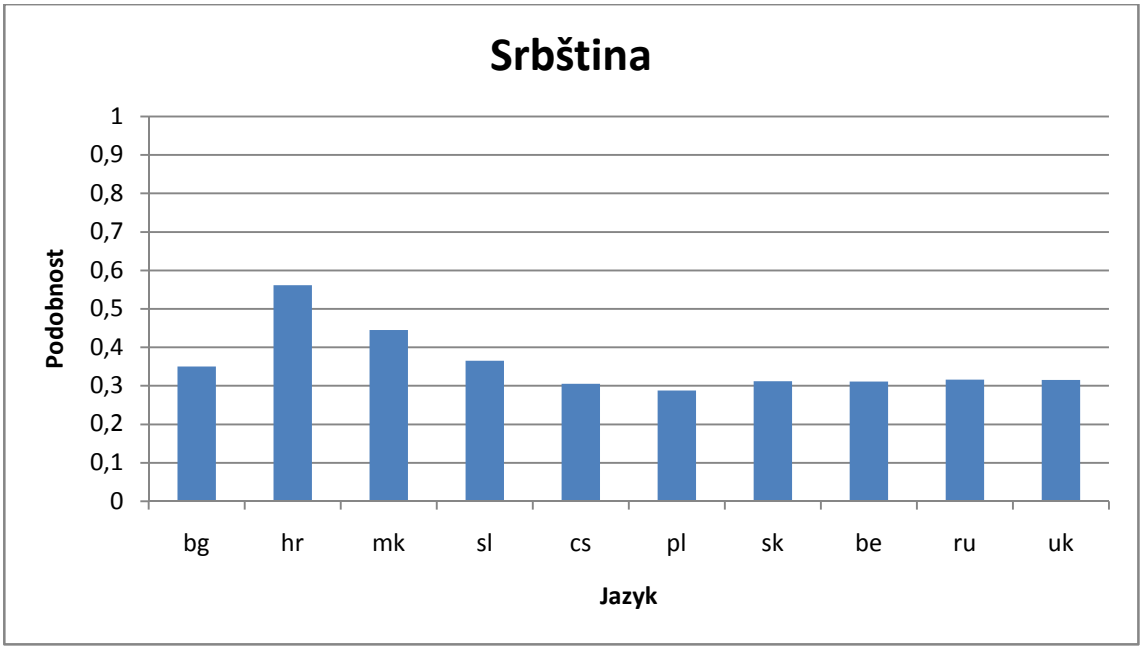
Nejvyšší podobnost má porovnání slov ze slovníku, kde je z výsledku vyloučen vliv slovosledu ve větách. Dá se tedy říci, že co se týče porovnání jednotlivých slov, jsou si velmi podobné, zatímco věty už méně.

Pro ukázku jsem dále vybral porovnání pro další tři slovanské jazyky s dalšími slovanskými jazyky, a to chorvatštinu, srbštinu a ruštinu. Výsledky jsou porovnáním textů z korpusu ASPAC, měřené pomocí Levenshteinovy vzdálenosti s mapováním cyrilice na latinku.



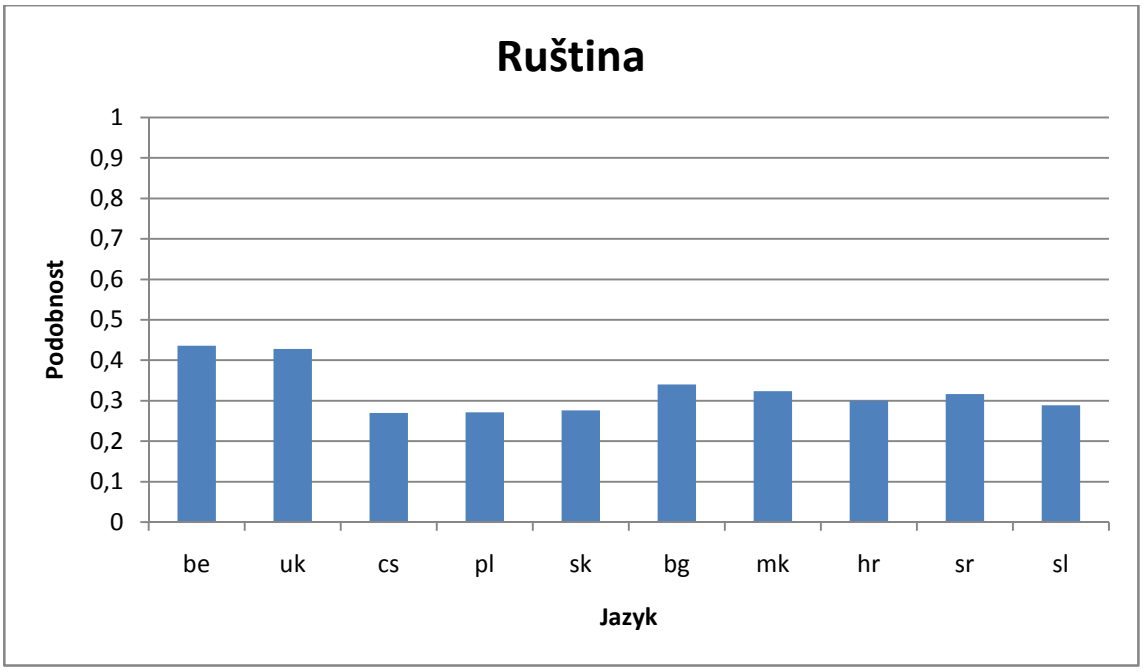
Graf 12. Chorvatšina – Levenshteinova vzdálenost

U chorvatštiny je vidět celkem vysoká podobnost se srbštinou, jelikož to jsou jazyky si blízké. Dále zvýšená podobnost s dalšími jihoslovanskými jazyky (slovinština, makedonština) oproti ostatním.



Graf 13. Srbština – Levenshteinova vzdálenost

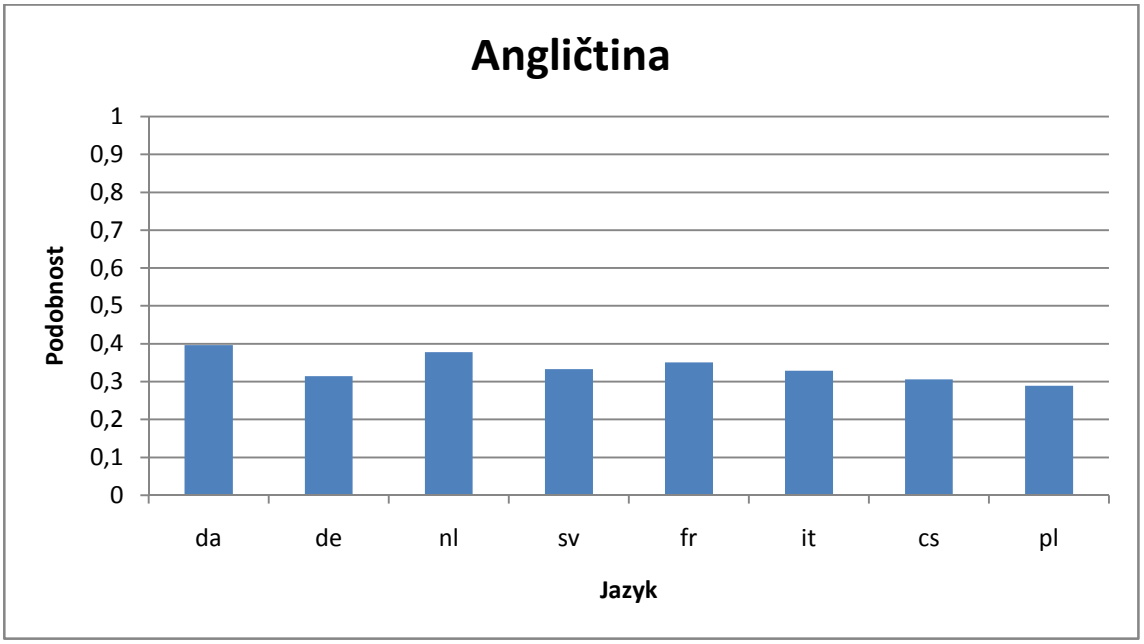
U srbštiny je vidět zase z druhé strany stejná podobnost s chorvatštinou a dále s ostatními jihoslovanskými jazyky, především makedonštinou.



Graf 14. Ruština – Levenshteinova vzdálenost

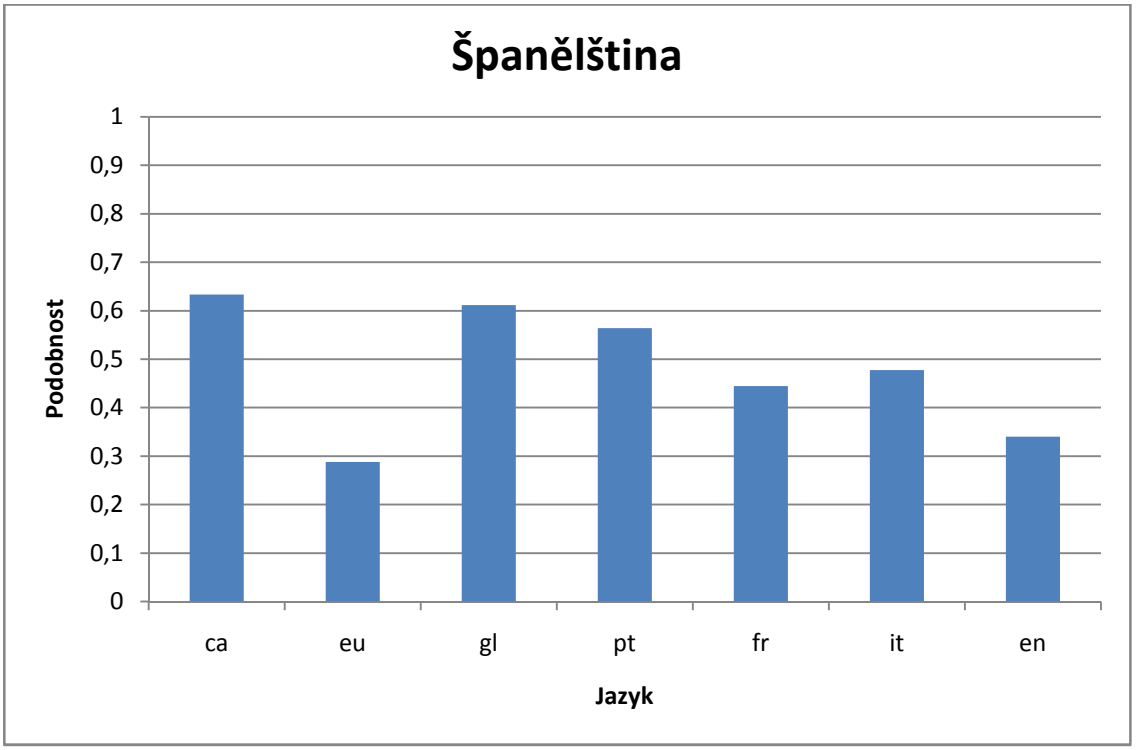
Z grafu pro ruštinu je vidět očekávaná podobnost, i když ne příliš vysoká, s dalšími východoslovanskými jazyky (běloruština a ukrajinština), déle trochu s bulharštinou a ostatními jihoslovanskými jazyky.

Dále jsem pro ukázku vybral výsledky měření na korpusu KDE pro angličtinu a španělštinu, ve kterém je zajímavé srovnání s dalšími jazyky používanými ve Španělsku.



Graf 15. Angličtina – Levenshteinova vzdálenost

Z grafu pro angličtinu lze vyčíst lehká podobnost s dánštinou, nizozemštinou a částečně by se dalo také říct, že i s francouzštinou. Ale rozdíl mezi podobností s češtinou či polštinou je velmi malý, maximálně do 10 %.



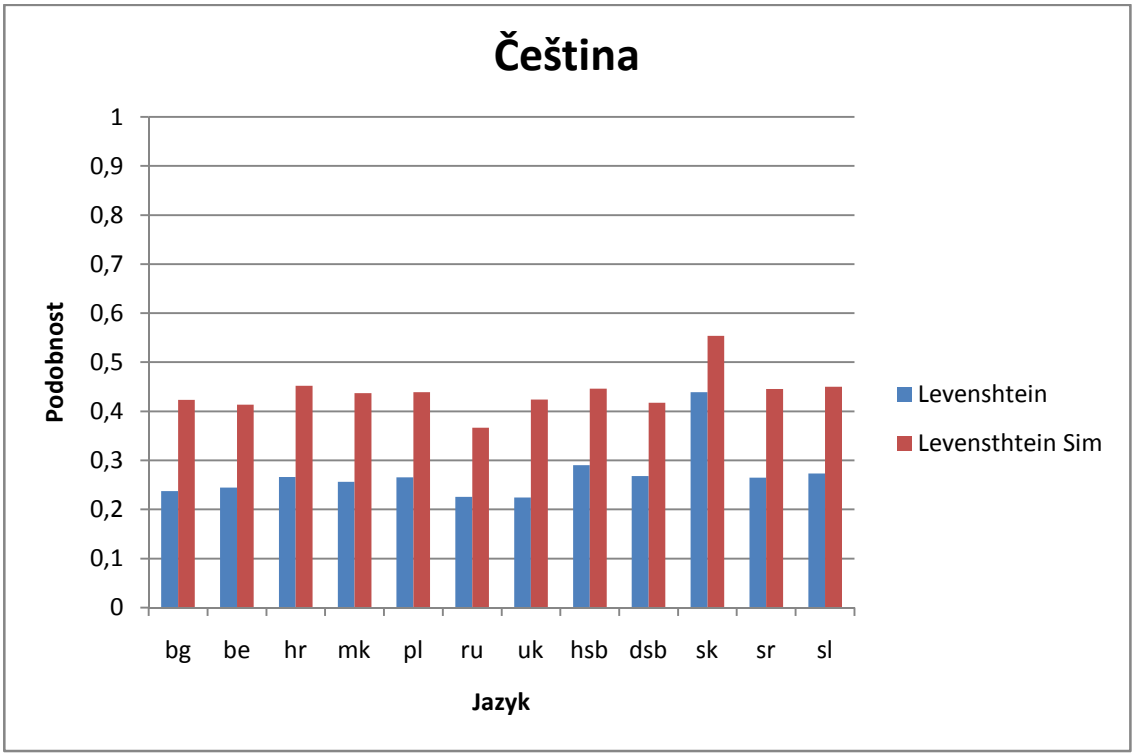
Graf 16. Španělština – Levenshteinova vzdálenost

Pro španělštinu jsem vybral porovnání s jazyky používanými ve Španělsku (katalánštinou, baskičtinou, galicijštinou), dalšími románskými jazyky a angličtinou.

Co se týče katalánštiny a galicijštiny, tak je zde vysoká podobnost, přes 60 %. U baskičtiny je podobnost pod 30 %, tedy menší podobnost než s angličtinou. Dále je vidět i vysoká podobnost s portugalštinou a o něco menší i s dalšími románskými jazyky.

5.1.3 Fonetická podobnost

Pro ukázku fonetického porovnání, které bylo provedeno pouze pro slovanské jazyky na základně fonetických transkripcí, jsem vybral hodnoty pro češtinu, srbštinu a ruštinu naměřené na korpusu ASPAC.

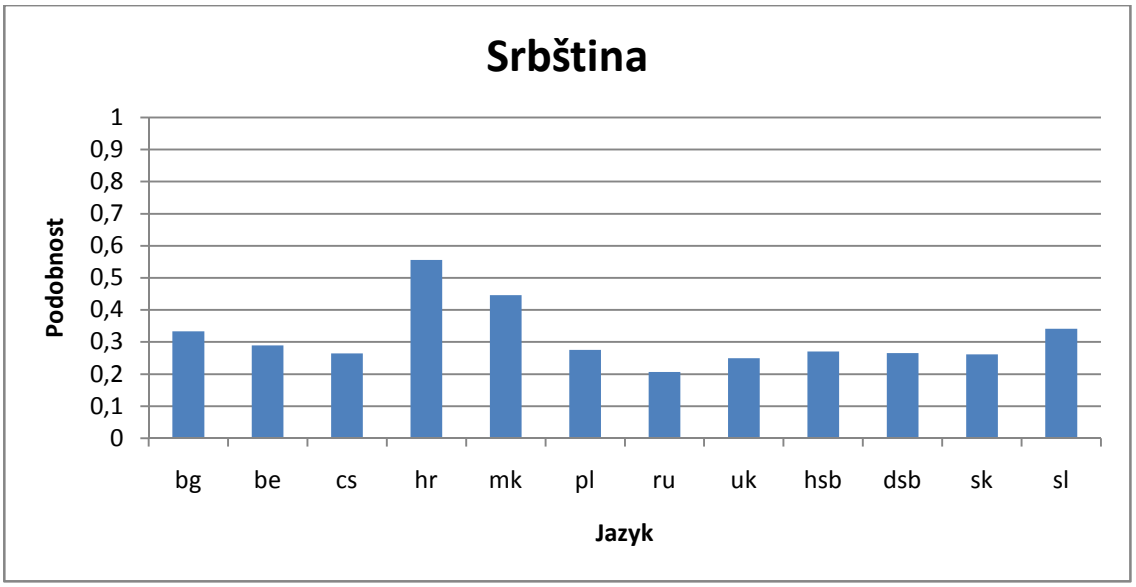


Graf 17. Čeština – Fonetické porovnání

Pro ukázku porovnání češtiny jsem vybral jak standardní Levenshteinovu vzdálenost, tak i s mapováním fonetické podobnosti. Jak je z grafu vidět, podobnost je opět nejvyšší se slovenštinou, podobnost s ostatními jazyky je podobná, jako u grafemického porovnání.

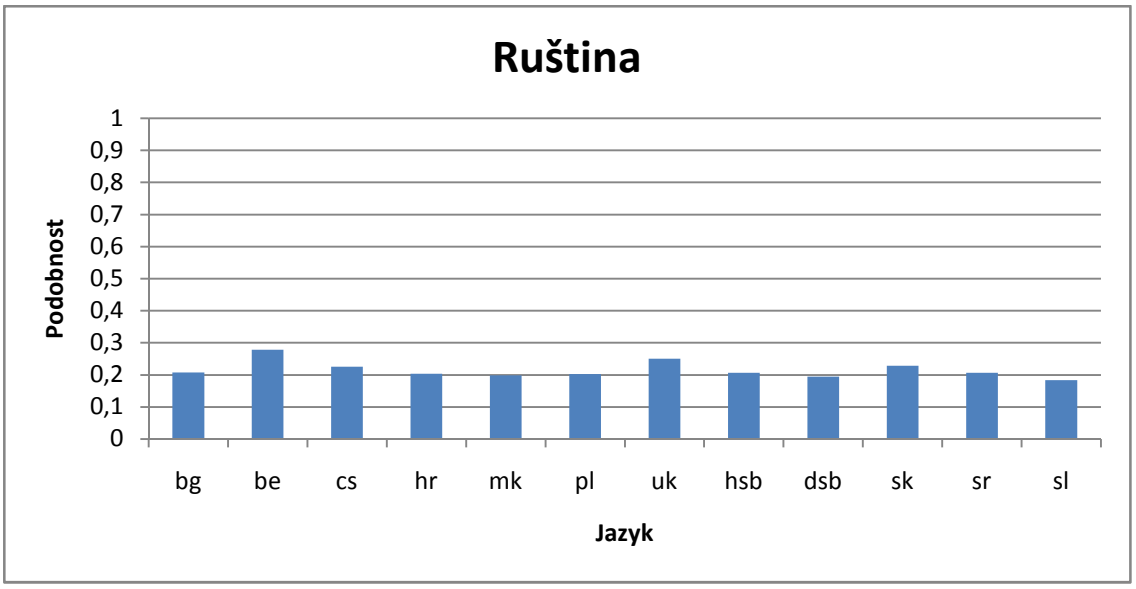
Vliv mapování fonetické podobnosti fonémů je sice viditelný, ale relativně jsou výsledky podobné, jako bez podobnosti. Pouze se zvýšilo podobnostní skóre o několik procent u všech porovnávaných jazyků.

Dále uvedu příklady pro další dva jazyky už bez mapování fonetické podobnosti.



Graf 18. Srbština – Fonetické porovnání

Z grafu pro podobnost srbštiny je vidět vysoká podobnost s chorvatštinou a makedonštinou a lehce vyšší se slovinštinou a bulharštinou. Nejmenší podobnost je s ruštinou.



Graf 19. Ruština – Fonetické porovnání

Ruština má, nejspíš díky svému dosti odlišnému fonetickému systému, celkem malou podobnost s ostatními slovanskými jazyky. Nejvyšší podobnost je zde s běloruštinou a ukrajinštinou, ale i tak není příliš vysoká.

5.2 Aplikace pro měření podobnosti a náročnost výpočtu

Ke všem úlohám bylo zapotřebí naprogramovat aplikace. Zdrojové kódy aplikací jsou přiloženy na CD. Aplikace jsem vytvořil v jazyce C#. Jelikož ale tvorba aplikací nebyla přímo zadaná, nevytvořil jsem uživatelsky použitelné aplikace a nebudu je zde popisovat. Je ale vhodné zde zmínit výpočetní náročnost použitých metod. Jelikož bylo nutné zpracovat velké množství dat, bylo vhodné optimalizovat algoritmy pro co nejrychlejší výpočet. Veškeré výpočty jsem prováděl na svém osobním počítači (AMD Phenom II 3,2 GHz, 4 GB RAM). Nebudu rozepisovat časovou a paměťovou náročnost jednotlivých metod, pouze pro každou z nich vypíši průměrný čas výpočtu. Výpočetní doba je změřena při porovnávání řetězců o průměrné délce 300 znaků.

Tabulka 6. Výpočetní doba metod pro porovnání

Metoda	Průměrný čas výpočtu
Levenshtein	20 ms
LevenshteinSim	50 ms
Damerau-Levenshtein	52 ms
Jaro	3 ms
Jaro-Winkler	3 ms
Jaccard	1 ms
Sørensen-Dice	1 ms
Výskyt slov - Dam-Lev	65 ms
Výskyt slov - Jaro-Winkler	17 ms
Výskyt slov - Soundex	130 ms

Všechny naměřené výsledky jsou uloženy ve formátu:

```
jazyk1;jazyk2;délka1;délka2;korpus;podobnost;komentář;
metoda1:hodnota:(parametr); ...metodaN:hodnota:(parametr);
```

Celkem bylo provedeno přes 22 miliónů porovnání a výsledky byly ukládány do souboru po sto tisících porovnání. Názvy souborů jsou pojmenovány podle data a času uložení, lze tedy jednoduše zjistit, jak dlouho výpočty trvaly. Průměrně výpočet těchto sto tisíc porovnání zabral dvě hodiny, celkem tedy veškeré výpočty trvaly přibližně 440 hodin.

Všechny výsledky jsou na přiloženém CD.

Závěr

Cílem této diplomové práce bylo seznámit se s problematikou hledání a měření podobnosti mezi jazyky a prakticky implementovat metody pro měření podobnosti se zaměřením na evropské jazyky. Jelikož tato problematika spadá do oboru komparativní lingvistiky, což je humanitní obor, bylo zapotřebí se na problematiku podívat z informatického hlediska a pokusit se navrhnout automatizovatelné metody pro měření podobnosti.

Práce se zabývala měřením podobnosti na textových datech. Bylo tedy nutné získat dostatečné množství dat s různým obsahem. Podařilo se mi získat šest paralelních korpusů, celkem pro 57 jazyků, a slovník pro 41 jazyků.

Metody pro měření podobnosti byly navrženy na základě algoritmů pro práci s textovými řetězci a množinami (využívající množiny jednotlivých znaků a vyšších n-gramů). Tyto metody byly využity pro porovnání dat ze všech korpusů a ze slovníku. Jako nejvhodnější a nejuniverzálnější metoda pro měření podobnosti se ukázala Levenshteinova vzdálenost a to jak pro měření podobnosti celých vět, tak slov. Ostatní metody jsou více specifické pro konkrétní využití. Pro textově nezávislé porovnání jsem použil data z korpusů a získal statistiku četnosti slov, podle které se dále porovnávala slovní zásoba jazyků.

Dalším mým úkolem bylo navrhnout mapování různých použitých abeced a specifických znaků. To znamenalo vytvořit převod pro cyrilici a řeckou alfabetu do latinky, aby mohly být porovnány jazyky využívající různé písmo. A dále vytvořit mapování různých národně specifických znaků a znaků s diakritikou. Pro to bylo potřeba sestavit skupiny podobných znaků a přiřadit jim podobnostní skóre. Jak se ale ukázalo, toto mapování mělo na výsledek minimální vliv.

Posledním úkolem bylo provést měření podobnosti výslovnosti slovanských jazyků. Musel jsem proto vytvořit systém pro fonetickou transkripci, nastudovat základní výslovnost 16 slovanských jazyků a sestavit pravidla pro transkripční systém. Poté všechny texty ve slovanských jazycích převést tímto systémem do výslovnostní podoby a změřit podobnost. Transkripci jsem provedl do mezinárodní fonetické abecedy IPA. Jelikož i zde se vyskytuje určitá podobnost mezi různými fonémy, vytvořil jsem systém pro mapování podobnosti různých fonémů na základě fonetického

rysu (výslovnostních vlastností každého fonému). Výsledný vliv na měření ale také nebyl příliš značný, jako v případě grafemického porovnání.

V závěru práce jsem popsal způsob jak interpretovat naměřená data a uvedl některé výsledky. Jelikož bylo ale výsledků z měření podobnosti velké množství, vybral jsem pro ukázku jen několik příkladů. Veškeré naměřené výsledky jsou přiloženy na CD.

Pro všechny měřící úlohy bylo potřeba naprogramovat jednoduché aplikace, které ale nebyly přímo zadány, proto jsem je nepopisoval, pouze zmínil výpočetní náročnost a přidal zdrojové kódy na přiložené CD.

Ve výsledku je z naměřených dat viditelné, že navržené metody fungují a výsledky jistým způsobem odpovídají očekávání, které se dalo odhadnout na základě obecných znalostí vztahů mezi jazyky. Hlavním cílem bylo především zjistit, v jaké relativní míře jsou si jazyky podobné mezi sebou a využít tyto informace v dalších úlohách.

Výsledky diplomové práce budou využity v Laboratoři počítačového zpracování řeči při efektivní modifikaci systémů rozpoznávání řeči (které již fungují v několika jazycích) pro další příbuzné jazyky. Pokud je totiž objektivně známo, které z existujících jazykových verzí je cílová aplikace nejpodobnější (zejména foneticky), lze uspořit velké množství lidské práce při sběru mluvených nahrávek, na nichž se cílový systém učí. Podobnost na úrovni ortografické a lexikální může zase usnadnit tvorbu výslovnostního slovníku a do určité míry i statistického jazykového modelu.

Seznam použité literatury

[1] ANTTILA, Raimo. Historical and comparative linguistics. Philadelphia: John Benjamins Pub. Co., 2009c1989, xv, 462 p. v. 6. ISBN 9789027286086.

[2] STAROSTIN, Sergei. Comparative-historical linguistics and lexicostatistics: Time depth in historical linguistics. 2000. Dostupné z: <http://newstar.cust.rinet.ru/Texts/Starostin_Glotto.pdf>

[3] MATTILA, Heikki E. Comparative legal linguistics. Burlington, VT: Ashgate, 2006, xv, 347 p. ISBN 978-075-4648-741.

[4] AUGST, Gerhard. New trends in graphemics and orthography. New York: W. de Gruyter, 1986, xi, 464 p. ISBN 9783110867329.

[5] KRČMOVÁ, Marie. Disciplíny studující zvukovou rovinu jazyka; od zvuku tvořeného mluvidly k jevům relevantním. [online]. 2003 [cit. 2014-04-10]. Dostupné z: <http://www.phil.muni.cz/jazyk/krcmova/fon/ucebnitext/>

[6] STROSSA, Petr. Vybrané kapitoly z počítačového zpracování řeči. Slezská univerzita v Opavě, 1999.

[7] PALA, Karel. Počítačové zpracování přirozeného jazyka. Fakulta informatiky Masarykovy univerzity, 2000.

[8] Co je korpus?. Český národní korpus [online]. [cit. 2013-05-02]. Dostupné z: <http://www.korpus.cz/>

[9] End-of-Sentence Detection and Text Segment Classification [online]. 2013-02-26 [cit. 2013 05-03]. Dostupné z: <http://www.cs.jhu.edu/~yarowsky/600.466.hw1.pdf>

[10] GALE, William A. a Kenneth W. CHURCH. A Program for Aligning Sentences in Bilingual Corpora [online]. 1991 [cit. 2013-05-03]. Dostupné z: <http://www.cs.jhu.edu/~yarowsky/600.466.hw1.pdf>

[11] Hunalign – sentence aligner. MOKK Centre for Media Research and Education [online]. 2013 [cit. 2013-05-4]. Dostupné z: <http://mokk.bme.hu/resources/hunalign/>

[12] JRC-Acquis. Joint research centre [online]. 2013-05-01 [cit. 2013-05-02]. Dostupné z: <http://ipsc.jrc.ec.europa.eu/index.php?id=198>

[13] OPUS. The open parallel corpus [online]. 2009 [cit. 2014-04-28]. Dostupné z: <http://opus.lingfil.uu.se/>

[14] BARENTSEN, Adrie. UNIVERSITY OF AMSTERDAM. Amsterdam Slavic Parallel Aligned Corpus [online]. [cit. 2014-04-28]. Dostupné z:

<http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/b/a/a.a.barentsen/a.a.barentsen.html>
[15] UNIVERSITY OF EDINBURG. Bible corpus [online]. [cit. 2014-04-28]. Dostupné z: <http://homepages.inf.ed.ac.uk/s0787820/bible/>
[16] ENGLISHCLUB. 5000 Most Common Words [online]. 2014 [cit. 2014-04-30]. Dostupné z: <http://www.englishclub.com/vocabulary/common-words-5000.htm>
[17] GOOGLE. Inside Google translate [online]. [cit. 2014-04-30]. Dostupné z: <http://translate.google.com/about/intl/en_ALL/>
[18] NEJEDLOVÁ, Dana a Marek VOLEJNÍK. Transkripce psaného českého textu do fonetické podoby. Laboratoř počítačového zpracování řeči, Technická univerzita v Liberci, 2009.
[19] OMNIGLOT. The online encyclopedia of writing systems and languages [online]. [cit. 2014-05-02]. Dostupné z: <http://www.omniglot.com/writing/languages.htm>
[20] WIKIPEDIA CONTRIBUTORS. Wikipedia, The Free Encyclopedia.: Language phonologies [online]. 2013 [cit. 2014-05-02]. Dostupné z: <http://en.wikipedia.org/wiki/Category:Language_phonologies>
[21] OMNIGLOT. International Phonetic Alphabet (IPA) [online]. [cit. 2014-05-02]. Dostupné z: <http://www.omniglot.com/writing/ipa.htm>
[22] Přispěvatelé Wikipedie, Ruština [online], Wikipedie: Otevřená encyklopedie, 2014 [cit. 2014-05-02] Dostupné z: <http://cs.wikipedia.org/w/index.php?title=Ru%C5%A1tina&oldid=11429127>
[23] CHOMSKY, Noam a Morris HALLE. The sound pattern of english. Cambridge, Massachusetts: MIT Press, 1968. ISBN 978-026-2530-972.
[24] DEPARTMENT OF LINGUISTICS. Introduction to Segmental Phonology [online]. [cit. 2014-05-04]. Dostupné z: <http://www.linguistics.ucsb.edu/projects/featuresoftware/index.php>
[25] KIPARSKY, Paul. The phonological basis of sound change. The handbook of phonological theory, 1995
[26] RISTAD, Eric Sven a Peter N. YIANILOS. Learning String Edit Distance. 1996. Dostupné z: <http://arxiv.org/pdf/cmp-lg/9610005.pdf>
[27] JOY OF DATA. Comparison of String Distance Algorithms [online]. 2013 [cit. 2014-05-05]. Dostupné z: <http://www.joyofdata.de/blog/comparison-of-string-distance-algorithms/>

[28] DUGGAN, Bryan. Edit (Levenshtein) distance [online]. 2008 [cit. 2013-05-05]. Dostupné z: <http://www.comp.dit.ie/bduggan/Courses/OOP/EditDistance.pdf>
[29] COHEN, William W., Pradeep RAVIKUMAR a Stephen E. FIENBERG. A Comparison of String Distance Metrics for Name-Matching Tasks. 2003. Dostupné z: <https://www.cs.cmu.edu/~pradeepr/papers/ijcai03.pdf>
[30] THE UNIVERSITY OF ARIZONA. Similarity measures [online]. [cit. 2014-05-05]. Dostupné z: <http://ag.arizona.edu/classes/rnr555/lecnotes/10.html>
[31] JEWISHGEN. Soundex Coding [online]. 2013 [cit. 2014-05-07]. Dostupné z: <http://www.jewishgen.org/InfoFiles/soundex.html>
[32] W3C. ISO 639 Language Codes [online]. 1999 [cit. 2014-05-10]. Dostupné z: <http://www.w3.org/WAI/ER/IG/ert/iso639.htm>

Obsah přiloženého CD

- Text diplomové práce
    - diplomova_prace_2014_Radek_Safarik.pdf
- Výsledky měření podobnosti
    - Tabulky s vyhodnocenými výsledky měření podobnosti
    - Výstupní data z měřící aplikace
- Fonetická transkripce – pravidla
    - Soubory s pravidly pro fonetický přepis pro každý jazyk
- Mapování podobnosti
    - Soubory pro mapování grafemické podobnosti
    - Tabulka pro fonetický rys
- Zdrojové kódy (v jazyce C#)
    - Aplikace pro grafemické měření podobnosti
    - Aplikace pro fonetické měření podobnosti
    - Aplikace pro vyhodnocení výsledků měření

Příloha 1. Tabulky převodů abeced na latinku

Cyrilice	
А	a
Б	b
в	v
г, Ґ, Ѓ	g
д, Ђ	d
Е, Є, Э	e
Ё	o
Ж, З, З́	z
S, Ц	dz
И, І, Ї, Й	i
J	j
К	k
Л	l
Љ	lj
М	m
Н	n
Њ	nj
О	o
П	p
Р	r
С, С́	s
Т, Ћ	t
Ќ	kj
У, Ў	u
Ф	f
Х	h
Ц, Ч	c
Ш	s
Щ	sc
Ы	y
Ю	ju
Я	ja

Alfabeta	
α, ά	a
β, ϐ	b
γ	g
δ	d
ε, έ, η, ή	e
ζ	z
θ, ϑ	th
ι, ΐ, ϊ, ί	i
κ	k
λ	l
μ	m
ν	n
ξ	x
ο, ό, ω, ώ	o
π, ϖ	p
ρ	r
σ, ς	s
τ	th
υ, ΰ, ϋ, ύ	y
φ	f
χ	ch
ψ	ps

Příloha 2. Seznam nežádoucích znaků

Čísla

1234567890

Interpunkce

. , ; : / \ " ' ¶ ? ! ( ) [ ] { } < > –

Další znaky

_ @ & # * $ € ° ~ ^ % § | + - =

Příloha 3. Seznam jazykových kódů podle ISO-639

af	afrikánština
ast	asturština
be	běloruština
bg	bulharština
br	bretonština
bs	bosenština
ca	katalánština
cs	čeština
csb	kašubština
cy	velština
da	dánština
de	němčina
dsb	dolnolužická srbština
el	řečtina
en	angličtina
eo	esperanto
es	španělština
et	estonština
eu	baskičtina
fi	finština
fr	francouzština
fy	západofrýština
ga	irština
gd	skotská gaelština
gl	galicijština
hr	chorvatšina
hsb	hornolužická srbština
hu	maďarština

is	islandština
it	italština
la	latina
lb	lucemburština
lv	lotyština
lt	litevština
mk	makedonština
mt	maltština
nb	bokmål
nds	dolní němčina
nl	nizozemština
nn	nynorsk
no	norština
oc	okcitánština
pl	polština
pt	portugalština
ro	rumunština
ru	ruština
se	sámština
sh	srbochorvatština
sk	slovenština
sl	slovinština
sq	albánština
sr	srbština
sv	švédština
svm	moliská chorvatšina
uk	ukrajinština
tr	turečtina

Příloha 4. Tabulka naměřených hodnot – Grafemické porovnání

	bg	be	cs	nl	en	de	hr	mk	pl	ru	uk	hsb	dsb	svm	sk	sr	sl
bg	0	0,3044	0,3731	0,3298	0,3303	0,3306	0,383	0,519	0,3738	0,4452	0,406	0,3562	0,336	0,3405	0,3756	0,4198	0,3768
be	0,3044	0	0,2642	0,2242	0,2194	0,2236	0,2675	0,2861	0,2765	0,4129	0,4155	0,2594	0,2819	0,3002	0,268	0,3034	0,2604
cs	0,3731	0,2642	0	0,4369	0,4437	0,4561	0,5507	0,3233	0,5733	0,3637	0,3535	0,5535	0,4524	0,4596	0,671	0,4521	0,5402
nl	0,3298	0,2242	0,4369	0	0,619	0,6088	0,5297	0,2672	0,5466	0,4141	0,3792	0,4311	0,5692	0,5717	0,5689	0,5135	0,5598
en	0,3303	0,2194	0,4437	0,619	0	0,5617	0,5699	0,4	0,5548	0,4127	0,3837	0,5284	0,5316	0,5576	0,5633	0,4922	0,5621
de	0,3306	0,2236	0,4561	0,6088	0,5617	0	0,4963	0,3	0,4844	0,3279	0,2898	0,4328	0,422	0,4176	0,4916	0,4086	0,481
hr	0,383	0,2675	0,5507	0,5297	0,5699	0,4963	0	0,4339	0,6248	0,4467	0,4538	0,6104	0,613	0,5836	0,6489	0,6308	0,6524
mk	0,519	0,2861	0,3233	0,2672	0,4	0,3	0,4339	0	0,2903	0,3771	0,3477	0,3333	0,4268	0,3205	0,2916	0,3999	0,2958
pl	0,3738	0,2765	0,5733	0,5466	0,5548	0,4844	0,6248	0,2903	0	0,4073	0,4059	0,5743	0,6166	0,4939	0,6013	0,4716	0,5562
ru	0,4452	0,4129	0,3637	0,4141	0,4127	0,3279	0,4467	0,3771	0,4073	0	0,436	0,3594	0,4463	0	0,2953	0,3234	0,2875
uk	0,406	0,4155	0,3535	0,3792	0,3837	0,2898	0,4538	0,3477	0,4059	0,436	0	0,3584	0,4495	0	0,3743	0,3985	0,4173
hsb	0,3562	0,2594	0,5535	0,4311	0,5284	0,4328	0,6104	0,3333	0,5743	0,3594	0,3584	0	0,6676	0,4685	0,5861	0,4323	0,5656
dsb	0,336	0,2819	0,4524	0,5692	0,5316	0,422	0,613	0,4268	0,6166	0,4463	0,4495	0,6676	0	0,6043	0,623	0,5356	0,6368
svm	0,3405	0,3002	0,4596	0,5717	0,5576	0,4176	0,5836	0,3205	0,4939	0	0	0,4685	0,6043	0	0	0	0
sk	0,3756	0,268	0,671	0,5689	0,5633	0,4916	0,6489	0,2916	0,6013	0,2953	0,3743	0,5861	0,623	0	0	0,4847	0,5606
sr	0,4198	0,3034	0,4521	0,5135	0,4922	0,4086	0,6308	0,3999	0,4716	0,3234	0,3985	0,4323	0,5356	0	0,4847	0	0,5365
sl	0,3768	0,2604	0,5402	0,5598	0,5621	0,481	0,6524	0,2958	0,5562	0,2875	0,4173	0,5656	0,6368	0	0,5606	0,5365	0

Příloha 5. Tabulka naměřených hodnot – Fonetické porovnání

	bg	be	cs	hr	mk	pl	ru	uk	hsb	dsb	sk	sr	sl
bg	0	0,2583	0,2376	0,3189	0,4149	0,2417	0,2071	0,2352	0,2345	0,2305	0,2363	0,3335	0,2878
be	0,2583	0	0,2448	0,2841	0,2796	0,3024	0,2785	0,3181	0,2531	0,2351	0,2602	0,2891	0,2731
cs	0,2376	0,2448	0	0,2659	0,2564	0,2655	0,2258	0,2242	0,2905	0,2684	0,4389	0,2649	0,2736
hr	0,3189	0,2841	0,2659	0	0,4031	0,2757	0,2039	0,2504	0,2695	0,2598	0,272	0,5563	0,3705
mk	0,4149	0,2796	0,2564	0,4031	0	0,2616	0,1991	0,2351	0,2563	0,2433	0,258	0,446	0,3253
pl	0,2417	0,3024	0,2655	0,2757	0,2616	0	0,2027	0,2337	0,2638	0,2454	0,256	0,2756	0,2664
ru	0,2071	0,2785	0,2258	0,2039	0,1991	0,2027	0	0,2509	0,2068	0,1945	0,229	0,2062	0,1839
uk	0,2352	0,3181	0,2242	0,2504	0,2351	0,2337	0,2509	0	0,2074	0,1988	0,2161	0,2491	0,2364
hsb	0,2345	0,2531	0,2905	0,2695	0,2563	0,2638	0,2068	0,2074	0	0,3588	0,286	0,2703	0,2748
dsb	0,2305	0,2351	0,2684	0,2598	0,2433	0,2454	0,1945	0,1988	0,3588	0	0,274	0,2659	0,279
sk	0,2363	0,2602	0,4389	0,272	0,258	0,256	0,229	0,2161	0,286	0,274	0	0,2617	0,2726
sr	0,3335	0,2891	0,2649	0,5563	0,446	0,2756	0,2062	0,2491	0,2703	0,2659	0,2617	0	0,3417
sl	0,2878	0,2731	0,2736	0,3705	0,3253	0,2664	0,1839	0,2364	0,2748	0,279	0,2726	0,3417	0

Příloha 6. Skupiny znaků pro grafemickou podobnost

a, á, à, â, ä, ă, ā, ã, å, ǻ, ą, ạ, æ, ǽ, ǣ

e, æ, ǽ, ǣ, é, è, ė, ê, ë, ĕ, ē, ę, ẹ, ə̧, ẹ, œ, ə, ə,

ɓ, b, ß

c, ć, ċ, ĉ, č, ç

d, ď, ḍ, đ, ɗ, ð

1.0, f, *ƒ*, Φ, φ

g, ġ, ĝ, ğ, ǧ, ģ

h, ĥ, ḥ, ħ, ɦ, χ

i, ij, ι, į, ɨ, ı, í, ì, î, ï, ǐ, ĭ, ī, ĩ, ị

j, ij, ĵ

k, ķ, ƙ, κ

l, Í, ļ, ł, ľ, l·

n, ȵ, ń, ï, ň, ñ, ņ, ŋ, ɲ

o, ó, ò, ô, ö, ǒ, ŏ, ō, õ, ő, ǫ, ọ, ø, ǿ, ơ, œ

r, ŕ, ř, ŗ, ɾ

s, ſ, ś, ŝ, š, ş, ș, ṣ

t, ť, ţ, ț, ŧ

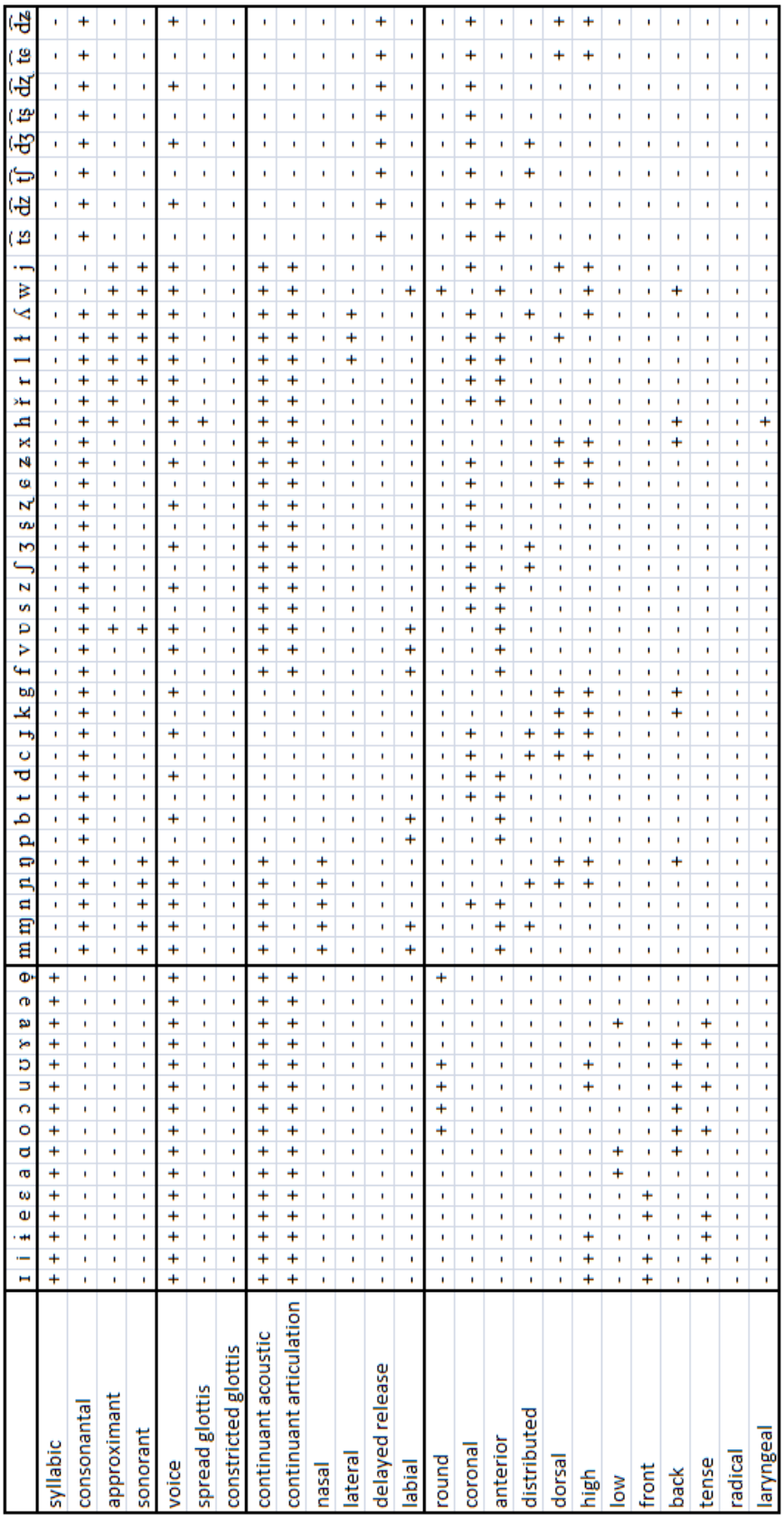
u, ú, ù, û, ü, ǔ, ŭ, ū, ũ, ű, ů, ų, ụ, ư, ʉ

w, ẃ, ẁ, ŵ, ẅ

y, ý, ỳ, ŷ, ÿ, ȳ, ỹ, ƴ, ẏ

z, ź, ż, ž, ẓ

Příloha 7. Fonetický rys - vlastnosti



	ɪ	i	ɨ	e	ɛ	a	ɑ	o	ɔ	u	ʊ	ɤ	ə	ɐ	ɵ	m	ɱ	n	ɲ	ŋ	p	b	t	d	c	ɟ	k	g	f	v	ʊ	s	z	ʃ	ʒ	ʂ	ʐ	ɕ	ʑ	x	h	ř	r	l	ɬ	ʎ	w	j	t͡s	d͡z	t͡ʃ	d͡ʒ	t͡ʂ	d͡ʐ	t͡ɕ	d͡ʑ
syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
consonantal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
approximant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-
voice	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	-	+	+	-	+	-	+	-	+	-	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
spread glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
constricted glottis	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
continuant acoustic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
continuant articulation	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-
delayed release	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+
labial	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-
round	-	-	-	-	-	-	-	+	+	+	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-
coronal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	+	+	+	+	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	+
anterior	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	+	+	+	+	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	+	-	+	+	-	-	-	-	-	-	-
distributed	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-
dorsal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	-	-	+	+	-	-	-	-	-	-	-	+	+
high	+	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	+	+	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	+	+	+	+	-	-	-	-	-	-	-	+	+
low	-	-	-	-	-	+	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-
front	+	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
back	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	-	+	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-
tense	-	+	+	+	-	-	-	+	-	-	+	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
radical	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
laryngeal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Příloha 8. Pravidla pro fonetickou transkripci

Běloruština

Z:б,в,г,д,дж,дз,ж,з

N:к,п,с,т,ф,ч,ш,ц,х

J:й,м,н,р,ў,л

S:а,е,ё,і,о,у,ы,э,ю,я

A:б-p,п-b,в-f,ф-v,ґ-x,к-g,д-t,т-d,ж-ʂ,ш-ʐ,з-s,с-z,г-k,дж-Č,дз-C,ч-Ď,ц-D,х-h

а => a / _

е => jɛ / _

ё => jɔ / _

і => i / _

о => ɔ / _

у => u / _

ў => w / _

ы => ɨ / _

ь => j / _

э => ɛ / _

ю => ju / _

я => ja / _

дзь => Ž / _

зь => ʐ / _

ль => ʎ / _

лю => ʎu / _

ля => ʎa / _

ле => ʎɛ / _

лё => ʎɔ / _

нь => ɲ / _

сь => ʂ / _

ць => Š / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

б => b / _

в => v / _

г => x / _

ґ => g / _

д => d / _

дж => Ď / _

з => z / _

й => j / _

к => k / _

л => l / _

м => m / _

н => n / _

п => p / _

р => r / _

с => s / _

т => t / _

ф => f / _

х => x / _

ц => c / _

ч => Č / _

ш => ʃ / _

Bulharština

Z:б,в,г,д,ж,з

N:к,п,с,т,ф,х,ц,ч,ш

J:й,л,м,н,р,ь,ю,я

S:а,е,и,о,у,ъ

A:б-p,в-f,г-k,д-t,ж-ʃ,з-s,к-g,п-b,с-z,т-d,ф-v,х-h,ц-D,ч-Ď,ш-Č

ъ => ɤ / _

й => j / _

и => i / _

м => ɱ / _<ф,в,>

м => m / _

н => ɱ / _<ф,в>

н => ŋ / _<к,г>

н => n / _

х => x / _

дз => C / _N

дз => D / _

дж => Č / _N

дж => Ď / _

л => l / _<и,е>

л => ɬ / _

о => o / -_

о => ɔ / _

а => ɐ / _-

а => a / _

у => u / _

ьо => jɔ / _

ц => C / _

ч => Č / _

б => b / _

в => v / _

г => g / _

д => d / _

е => ɛ / _

ж => ʒ / _

з => z / _

к => k / _

п => p / _

р => r / _

с => s / _

т => t / _

ф => f / _

ш => ʃ / _

щ => ʃt / _

ю => ju / _

я => jɐ / _-

я => ja / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

Bosenština

Z:b,d,đ,dž,dz,g,h,v,z,ž

N:c,č,ć,f,k,p,s,š,t

J:j,l,m,n,r

S:a,e,i,o,u

A:v-f,f-v,c-D,č-Ď,ć-Ź,đ-Č,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,dz-C,s-z,dž-Č,z-s

e => ɛ / _

o => ɔ / _

lj => ʎ / _

l => ł / _

ie => je / _

x => ks / _

q => kʊ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

ć => Ś / _

dž => Ď /_

dz => D / _

đ => Ź / _

h => x / _

nj => ɲ / _

š => ʃ / _

v => ʊ / _

ž => ʒ / _

Čeština

Z:b,d,ď,g,h,v,z,ž,dz,dž

N:c,č,f,ch,k,p,s,š,t,ť

J:j,l,m,n,ň,r,ř

S:a,á,e,i,í,o,ó,u,ú,ů,y,ý

A:v-f,f-v,c-D,č-Ď,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,ch-h,s-z,z-s,ď-c,ť-ɟ,dz-C,dž-Č

dě => ɟɛ / _

tě => cɛ / _

ně => ɲɛ / _

ě => ně / m_

d => ɟ / _<i,í>

t => c / _<i,í>

n => ɲ / _<i,í>

á => a / _

e => ɛ / _

é => ɛ / _

ě => jɛ / _

i => ij / _S

i => ɪ / _

í => i / _

y => ɪ / _

ý => i / _

ó => o / _

ú => u / _

ů => u / _

m => ɱ / _<f,v>

n => ŋ / _<k,g>

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

dz => D / _

dž => Ď / _

ď => ɟ / _

ch => x / _

ň => ɲ / _

š => ʃ / _

ž => ʒ / _

ť => c / _

q => kv / _

w => v / _

x => ks / _

Kašubština

Z:b,d,g,h,z,ż,dz,rz,dż,w

N:c,f,k,p,s,t,cz,sz,ch

J:j,l,ł,m,n,ń,r

S:a,ą,ã,e,é,ë,i,o,ò,ó,ô,u,ù,y

A:d-t,t-d,f-v,w-f,k-g,g-k,s-z,z-s,p-b,b-p,c-D,ś-ʑ,ź-ɕ,ż-ʂ,dz-C,dż-Č,ch-h,h-x,sz-ʒ,cz-Ď,rz-ʃ

ch => x / _

ą => ɔn / _

ã => an / _

e => ɛ / _

é => e / _

ë => ə / _

ł => w / _

ń => ɲ / _

o => ɔ / _

ò => wɛ / _

ó => o / _

ô => ɛ / _

ù => wu / _

y => i / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

ż => ʒ / _

c => C / _

cz => Č / _

dz => D / _

dż => Ď / _

rz => ʐ / _

sz => ʃ / _

w => v / _

Dolnolužická srbština

Z:b,d,g,h,v,z,ž,dz,dź

N:c,č,ć,f,ch,k,p,s,š,t

J:j,l,ł,m,n,ń,r,ř

S:a,e,i,o,ó,u,y

A:p-b,b-p,k-g,g-k,t-d,d-t,c-D,č-Ď,š-ʒ,ž-ʃ,h-x,s-z,z-s,dz-C,dź-Č,ć-Ź,f-v,v-f,ch-h

a => aj / _ž

e => ɛj / _ž

e => ɛ / _

e => ɛj / _ž

x => ks / _

q => kv / _

ch => kh / -_

ně => ɲie / _

u => ɔ / _

h => h / -_

h => - / _<Z,N,J>

ě => ie / _

y => ɨ / _

ł => / <Z,N,J>_-

ł => w / _

r => - / <Z,N,J>_-

n => ɲ / _i

ń => ɲ / _

ó => uj / _ž

ó => u / _

ř => ʃ / <p,k>_

ř => s / t_

r => - / <Z,N,J>_-

w => - / _<N,Z,J>

w => u / _<N,Z,J,->

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

š => ʃ / _

ž => ʒ / _

ć => Ś / _

dz => D / _

dź => Ď / _

ź => ʐ / _

Chorvatština

Z:b,d,đ,dž,g,h,v,z,ž,dz

N:c,č,ć,f,k,p,s,š,t

J:j,l,m,n,r

S:a,e,i,o,u

A:v-f,f-v,c-D,č-Ď,ć-Ź,đ-Č,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,s-z,z-s,dž-Č,dz-C

e => ɛ / _

o => ɔ / _

lj => ʎ / _

l => ɬ / _

ie => je / _

x => ks / _

q => kʊ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

ć => Ś / _

dž => Ď /_

dz => D / _

đ => Ź / _

h => x / _

nj => ɲ / _

š => ʃ / _

v => ʊ / _

ž => ʒ / _

Hornolužická srbština

Z:b,d,g,h,v,z,ž,dz,dž,dź

N:c,č,ć,f,ch,k,p,s,š,t

J:j,l,ł,m,n,ń,r,ř

S:a,e,i,o,ó,u,y

A:p-b,b-p,k-g,g-k,t-d,d-t,c-D,č-Ď,š-ʒ,ž-ʃ,h-x,s-z,z-s,dz-C,dž-Ś,dź-Č,ć-Ź,f-v,v-f,ch-h

a => aj / _ž

e => ɛj / _ž

e => ɛ / _

e => ɛj / _ž

x => ks / _

q => kv / _

ch => kh / -_

ně => ɲie / _

u => ɔ / _

h => h / -_

h => - / _<Z,N,J>

ě => ie / _

y => ɨ / _

ł => / <Z,N,J>_-

ł => w / _

r => - / <Z,N,J>_-

n => ɲ / _i

ń => ɲ / _

ó => uj / _ž

ó => u / _

ř => ʃ / <p,k>_

ř => s / t_

r => - / <Z,N,J>_-

w => - / _<N,Z,J>

w => u / _<N,Z,J,->

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

š => ʃ / _

ž => ʒ / _

ć => Ś / _

dz => D / _

dž => Ź / _

dź => Ď / _

ź => ʐ / _

Makedonština

Z:б,в,г,ѓ,д,ж,s,з,ц

N:к,ќ,п,с,т,ф,х,ц,ч,ш

J:j,л,љ,м,н,њ,р

S:а,е,и,о,у

A:б-p,в-f,г-k,ѓ-c,д-t,ж-ʃ,з-s,к-g,ќ-ɟ,п-b,с-z,т-d,ф-v,ш-ʒ,х-h,ц-D,ч-Ď,s-C,џ-Č

а => a / _

е => ɛ / _

и => i / _

j => j / _

л => ɬ / _

љ => l / _<a,у,о>

љ => ʎ / _

м => ɱ / _<ф,в>

м => m / _

н => ŋ / _<к,г>

н => n / _

њ => ɲ / _

о => ɔ / _

р => r / _

у => u / _

х => x / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

ц => C / _

џ => Ď / _

ч => Č / _

s => D / _

б => b / _

в => v / _

г => g / _

ѓ => ɟ / _

д => d / _

ж => ʒ / _

з => z / _

к => k / _

ќ => c / _

п => p / _

с => s / _

т => t / _

ф => f / _

ш => ʃ / _

Polština

Z:b,d,g,h,w,z,ź,ż

N:c,ć,f,k,p,s,ś,t,ch

J:j,l,ł,m,n,ń,r

S:a,ą,e,ę,i,o,ó,u,y

A:d-t,t-d,f-v,w-f,k-g,g-k,h-x,s-z,z-s,p-b,b-p,ć-Ź,c-D,ś-ʑ,ź-ɕ,ż-ʂ,ch-h

ę => ɛ / _<l, ł>

ę => ɛm / _<b, p>

ę => ɛɲ / _<j, dź>

ę => ɛŋ / _<k, g>

ę => ɛn / _<c, t, d>

ę => ɛw / _

y => ɨ / _

o => ɔ / _

e => ɛ / _

ą => ɔ / _<l, ł>

ą => ɔm / _<p, b>

ą => ɔn / _<c, t, d>

ą => ɔŋ / _<k, g>

ą => ɔɲ / _<dź>

ą => ɔw / _

i => j / _S

i => j / S_

ó => u / _

szcz => ʃČ / _

cz => Ď / _<Z,N>Z

cz => Ď / _Z

cz => Č /_

dz => C / _<Z,N>N

dz => C / _N

dz => D / _

dź => Ś / _<Z,N>N

dź => Ś / _N

dź => Ź / _

dzi => Ź / _

dż => Č / _<Z,N>N

dż => Č / _N

dż => Ď / _

rz => ʂ / < k, ch, p, t>_

rz => ʐ / _

sz => ʒ / _<Z,N>N

sz => ʒ / _N

sz => ʃ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => Š / _i

n => ɲ / _i

x => ks / _

q => kv / _

c => ts / _

w => v / _

ć => Ś / _

ł => w / _

ḿ => m / _

ń => ɲ / _

h => x / _

ś => ɕ / _

ź => ʑ / _

ż => ʐ / _

z => ʑ / _i

ch => x / _

Ruština

Z:б,в,г,д,ж,з

N:к,п,с,т,ф,х,ц,ч,ш

J:й,л,м,н,р

S:а,е,ё,и,о,у,ы,э,ю,я

A:б-p,п-b,в-f,ф-v,г-k,к-g,д-t,т-d,ж-ʂ,з-s,с-z,х-h,ц-D,ч-Ď,ш-ʐ

дя => ʝä / _

дю => ʝu / _

де => ʝɛ / _

дё => ʝo̯ / _

ди => ʝɪ / _

дь => ʝ / _

тя => cä / _

тю => cu / _

те => cɛ / _

тё => co̯ / _

ти => cɪ / _

ть => c / _

ь => j / _

я => jä / _

ю => ju / _

е => jɛ / _

ё => jo̯ / _

а => ä / _

о => o̯ / _

э => ɛ / _

у => u / _

ы => i̯ / _

и => i̯ / -_

и => jɪ / <г,в,м,н,п,р>_

и => ɪ / _

й => j / _

сч => ɕŚ / _

зч => ɕŚ / _

шь => ʂ / _

л => ʎ / _<я,е,ё,ю,и,ь>

г => x / _-

сь => zj / _Z

сь => sj / _

ч => Ź / _Z

ч => Ź / _<Z,N>Z

ц => D / _Z

ц => D / _<Z,N>Z

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

б => b / _

в => v / _

г => g / _

д => d / _

ж => ʐ / _

з => z / _

к => k / _

л => ł / _

м => m / _

н => n / _

п => p / _

р => r / _

с => s / _

т => t / _

ф => f / _

х => x / _

ц => C / _

ч => Ś / _

ш => ʂ / _

щ => ɕŚ / _

Srbochorvatština

Z:b,d,đ,dž,dz,g,h,v,z,ž

N:c,č,ć,f,k,p,s,š,t

J:j,l,m,n,r

S:a,e,i,o,u

A:v-f,f-v,c-D,č-Ď,ć-Ź,đ-Č,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,s-z,dž-Č,dz-C,z-s

e => ɛ / _

o => ɔ / _

lj => ʎ / _

l => ɬ / _

ie => je / _

x => ks / _

q => kʊ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

ć => Ś / _

dž => Ď /_

dz => D / _

đ => Ź / _

h => x / _

nj => ɲ / _

š => ʃ / _

v => ʊ / _

ž => ʒ / _

Slovenština

Z:b,d,ď,dz,dž,g,h,v,z,ž

N:c,č,f,ch,k,p,s,š,t,ť

J:j,l,Ĺ,ľ,m,n,ň,r,ŕ

S:a,á,ä,e,é,i,í,o,ó,ô,u,ú,y,ý

A:v-f,f-v,c-D,č-Ď,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,ch-h,s-z,z-s,ď-c,ť-ɟ,dz-C,dž-Č

ia = > ja / _

ie => jɛ / _

iu => ju / _

tí => tɪ / -_-

ne => nɛ / _-

te => tɛ / _-

de => dɛ / _-

le => lɛ / _-

l‘ => ʎ / _

Ĺ => l / _

ŕ => r / _

á => a / _

ä => ɛ / _

e => ɛ / _

ó => o / _

ô => uɔ / _

q => kv / _

w => v / _

x => ks / _

i => ɪ / _

í => i / _

y => ɪ / _

ý => i / _

ú = > u / _

v = > u / _-

m => ɱ / _<f,v>

n => ŋ / _<k,g>

ň => ɲ / _

s => z / _<m>

dz => C / _<N,Z>N

dz => C / _<N,->

dž => ʃ / _<N,Z>N

dž => ʃ / _<N,->

dž => ʒ / _

ch => x / _

tě => cɛ / _

dě => ɟɛ / _

mě => mɛ / _

ně => nɛ / _

ne => ɲɛ / _

de => ɟɛ / _

te => cɛ / _

le => ʎɛ / _

ni => ɲɪ / _

di => ɟɪ / _

ti => cɪ / _

li => ʎɪ / _

ní => ɲi / _

dí => ɟi / _

tí => ci / _

lí => ʎi / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _-Z

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

š => ʃ / _

ž => ʒ / _

č => Č / _

c => C / _

ď => ɟ / _

Slovinština

Z:b,d,g,h,v,z,ž,dz,dž

N:c,č,f,k,p,s,š,t

J:j,l,m,n,r

S:a,e,i,o,u

A:v-f,f-v,c-D,č-Ď,p-b,b-p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-h,s-z,z-s,dz-C,dž-Č

aj => aɪ / _

ej => ɛɪ / _

oj => ɔɪ / _

uj => uɪ / _

el => ew / _

ev => ew / _

iv => ɪw / _

e => ɛ / _

o => ɔ / _

h => x / _

v => w / S_

v => ʊ / _<Z,N,J,->

v => ʊ / _

lj => l / _<Z,N,J,->

lj => ʎ / _

nj => n / _<Z,N,J,->

nj => ɲ / _

rj => r / _<Z,N,J,->

l => w / _<Z,N,J,->

m => ɱ / _<f,v>

n=> ŋ / _ <k,g>

x => ks / _

q => kʊ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

dž => Ď / _

dz => D / _

c => C / _

č => Č / _

š = > ʃ / _

ž = > ʒ / _

Srbština – latinka

Z:b,d,đ,dž,dz,g,h,v,z,ž

N:c,č,ć,f,k,p,s,š,t

J:j,l,m,n,r

S:a,e,i,o,u

A:v-f,f-v,c-D,č-Ď,ć-Ź,đ-Č,dz-C,p-b,b-

p,k-g,g-k,d-t,t-d,š-ʒ,ž-ʃ,h-x,s-z,dž-Č,z-s

e => ɛ / _

o => ɔ / _

lj => ʎ / _

l => ɬ / _

ie => je / _

x => ks / _

q => kʊ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

c => C / _

č => Č / _

ć => Ś / _

dž => Ď /_

đ => Ź / _

h => x / _

nj => ɲ / _

š => ʃ / _

v => ʊ / _

ž => ʒ / _

Srbština – cyrilice

Z:б,в,г,д,ђ,ж,з,џ

N:к,п,с,т,ћ,ф,х,ц,ч,ш

J:j,л,љ,м,н,њ,р

S:а,е,и,о,у

A:б-p,п-b,в-f,ф-v,г-k,к-g,д-t,т-d,ж-ʃ,ш-ʒ,з-s,с-z,ц-D,џ-Č,ч-Ď,ђ-Ś,ћ-Ź,х-h

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

а => a / _

еи => jɛ / _

е => ɛ / _

и => i / _

о => ɔ / _

у => u / _

р => r / _

в => ʊ / _

j => j / _

л => ɬ / _

љ => ʎ / _

м => m / _

н => n / _

њ => ɲ / _

ф => f / _

с => s / _

з => z / _

ш => ʃ / _

ж => ʒ / _

х => x / _

ц => C / _

џ => Ď / _

ч => Č / _

ђ => Ź / _

ћ => Ś / _

б => b / _

п => p / _

д => d / _

т => t / _

г => g / _

к => k / _

Ukrajinština

Z:б,в,г,ґ,д,ж,з

N:к,п,с,т,ф,х,ц,ч,ш

J:й,л,м,н,р

S:а,е,є,и,і,ї,о,у,ю,я

A:б-p,п-b,в-f,ф-v,ґ-k,к-g,д-t,т-d,ж-ʃ,ш-ʒ,з-s,с-z,г-x,х-h,ц-D,ч-Ď

а => ɑ / _

е => ɛ / _

є => jɛ / _

и => ɪ / _

і => i / _

ї => ji / _

о => ɔ / _

у => u / _

ь => j / _

ю => ju / _

я => jɑ / _

дї => ʝi / _

дю => ʝu / _

дя => ʝɑ / _

дь => ʝ / _

дє => ʝɛ / _

тї => ci / _

тю => cu / _

тя => cɑ / _

ть => c / _

тє => cɛ / _

лю => ʎu / _

ля => ʎɑ / _

лє => ʎɛ / _

лї => ʎi / _

ль => ʎ / _

ню => ɲu / _

ня => ɲɑ / _

нє => ɲɛ / _

нї => ɲi / _

нь => ɲ / _

N => A / _<Z,N>Z

N => A / _Z

Z => A / _Z-

Z => A / _<Z,N>N

Z => A / _N

Z => A / _-

б => b / _

в => ʊ / _S

в => w / <-,S>_<Z,N,J>

в => u / S_-

в => v / _

г => h / _

ґ => g / _

д => d / _

ж => ʒ / _

з => z / _

й => j / _

к => k / _

л => l / _

м => m / _

н => n / _

п => p / _

р => r / _

с => s / _

т => t / _

ф => f / _

х => x / _

ц => C / _

ч => Č / _

ш => ʃ / _

щ => ʃČ / _