TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky, informatiky a mezioborových studií

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

Identifikace jazyka textového dokumentu

Language identification of text documents

Diplomová práce

Autor: Bc. Jan Valta
Vedoucí práce: prof. Ing. Jan Nouza, CSc.

V Liberci 16.5.2012

Abstrakt

Diplomová práce se zabývá problematikou identifikace jazyka textového dokumentu pomocí statistických n-gramových modelů.

Teoretická část popisuje statistický n-gramový model, jeho vytváření a vyhodnocování. Dále popisuje základní vyhlazovací techniky a typy n-gramových modelů.

Praktická část porovnává výsledky identifikace jazyka pro různé n-gramové modely, které se liší ve vyhlazovací technice, stupni a typu modelu. Dále pak zjišťuje vliv diakritiky při identifikaci jazyka.

Klíčová slova

identifikace jazyka, n-gramový model, vyhlazovací technika

Abstract

This diploma thesis addresses issues about language identification of text documents with statistical n-gram models.

Theoretical section describes statistical n-gram models, its creates and evaluations. Next part describes basic smoothing technique and types n-gram models.

Practical section describes results of language identification for different n-gram models, which differs in smoothing technique, order and type n-gram models. Further determines the influence diacritics in language identification.

Keywords

language identification, n-gram model, smoothing technique

Obsah

Přílohy

Úvod

V mnoha oblastech zpracování dat se využívá modelování pomocí n-gramových modelů. Tato práce se zaměřuje na jednu konkrétní oblast, a to zpracování textových dat s cílem identifikovat jazyk v dokumentu. Znalost jazyka textového dokumentu pak může pomoct v různých aplikacích, např. v překladačích jazyků, v aplikacích rozpoznávající text z obrázků apod.

V teoretické části této práce jsou nejdříve popsány základní pojmy teorie pravděpodobnosti, ze kterých vychází n-gramový model, následně je popsán samotný n-gramový model, jeho vytváření (trénování) a vyhodnocování. Další část pak popisuje nutnost použití vyhlazování n-gramového modelu a popisuje základní vyhlazovací techniky.

V praktické části této práce je nejdříve popsán způsob, jakým byla získána textová data různých jazyků, a které jazyky se takto podařilo získat. Následně jsou zobrazeny a porovnány výsledky identifikace jazyka různých n-gramových modelů. Modely jsou porovnány podle typu, podle použité vyhlazovací techniky a podle stupně modelu. V další části je posouzen vliv diakritiky na úspěšnost identifikace jazyka.

V poslední části jsou popsány vytvořené aplikace v rámci této práce, a to hlavně aplikace pro třídění textů podle jazyka, která byla jedním z cílů samotné práce.

1 Teorie pravděpodobnosti

V této části uvedu základní pojmy teorie pravděpodobnosti, které slouží pro odvození jazykového modelu.

1.1 Náhodný pokus

Je děj, který lze libovolněkrát opakovat, přičemž výsledek tohoto děje není jednoznačně určen vstupními podmínkami.

1.2 Náhodný jev

Náhodný jev je tvrzení o výsledku náhodného pokusu, o kterém lze po provedení náhodného pokusu prohlásit, zda je či není pravdivý.

1.3 Pravděpodobnost náhodného jevu

Pravděpodobnost náhodného jevu A lze vypočítat podle následujícího vztahu.

$$P(A) = \frac{c(A)}{N} \tag{1.1}$$

kde $c(A)$ je počet výskytů jevu A, N je počet pokusů

1.4 Podmíněná pravděpodobnost

Pravděpodobnost, že nastane jev A za podmínky jevu B, lze vypočítat podle následujícího vztahu.

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{1.2}$$

kde $P(A,B)$ je pravděpodobnost, že nastaly jevy A a B,
$P(B)$ je pravděpodobnost jevu B

Tento vztah lze vyjádřit i pomocí počtu výskytů daných jevů.

$$P(A|B) = \frac{\frac{c(A,B)}{N}}{\frac{c(B)}{N}} = \frac{c(A,B)}{c(B)} \tag{1.3}$$

kde $c(A,B)$ počet výskytů, kdy nastaly jevy A a B, $c(B)$ počet výskytů jevu B,
N je počet pokusů

2 Jazykový model

Jazykový model pro identifikaci jazyka slouží k výpočtu pravděpodobnosti sekvence znaků $P(z_1, z_2, z_3, \dots z_n)$. Jinými slovy slouží k výpočtu podmíněné pravděpodobnosti $P(z_n|z_1, z_2, \dots z_{n-1})$, že přijde znak z_n po sekvenci znaků $z_1, z_2, \dots z_{n-1}$. Jazykový model tedy není nic jiného, než souhrn informací, které slouží pro výpočet pravděpodobnosti jakékoliv sekvence znaků, která se objeví v neznámém vyhodnocovaném textu. V nejjednodušším případě to je přímo souhrn pravděpodobností všech možných sekvencí znaků.

2.1 Ideální jazykový model

S přihlédnutím na vztah podmíněné pravděpodobnosti (1.2) lze napsat následující vztah.

$$P(A,B) = P(A|B) \cdot P(B) \tag{2.1}$$

$P(A|B)$ je podmíněná pravděpodobnost, že nastal jev A po jevu B,
kde $P(A,B)$ je pravděpodobnost, že nastaly jevy A a B,
$P(B)$ je pravděpodobnost jevu B

Obecně pak lze napsat následující vztah pro ideální jazykový model.

$$P(z_1, z_2, z_3, \dots z_n) = P(z_1) \cdot P(z_2|z_1) \cdot P(z_3|z_1, z_2) \cdot \dots \cdot P(z_n|z_1 \dots z_{n-1}) \tag{2.2}$$

V praxi nelze tento ideální jazykový model použít, protože vyžaduje příliš velký trénovací korpus, v podstatě nekonečný. Zjednodušení přináší N-gramový model.

2.2 N-gramový model

Předpoklad je, že platí následující zjednodušení. Místo toho, abychom zjišťovali pravděpodobnost znaku z_n, že přijde po sekvenci znaků $z_1, \dots, z_{n-1}$ (tedy po celé historii), použijeme pouze $N-1$ předchozích znaků.

$$P(z_n | z_1^{n-1}) \approx P(z_n | z_{n-N+1}^{n-1}) \tag{2.3}$$

kde N představuje stupeň N-gramu,
značení z_1^{n-1} představuje sekvenci znaků $z_1, z_2, \dots, z_{n-1}$
značení z_{n-N+1}^{n-1} představuje sekvenci znaků $z_{n-N+1}, \dots, z_{n-1}$

U unigramového modelu neuvažujeme žádnou předchozí historii znaků, ale pouze násobíme pravděpodobnosti jednotlivých znaků. U vyšších stupňů se však už využívá předchozí vztah.

Pro unigramový model tedy získáme následující vztah.

$$P(z_1, z_2, z_3, \dots z_n) = P(z_1) \cdot P(z_2) \cdot \dots \cdot P(z_n) \tag{2.4}$$

Pro bigramový model tak získáme následující vztah.

$$P(z_1, z_2, z_3, \dots z_n) = P(z_2|z_1) \cdot P(z_3|z_2) \cdot \dots \cdot P(z_n|z_{n-1}) \tag{2.5}$$

Pro trigramový model získáme následující vztah.

$$P(z_1, z_2, z_3, \dots z_n) = P(z_3|z_1, z_2) \cdot P(z_4|z_2, z_3) \cdot \dots \cdot P(z_n|z_{n-2}, z_{n-1}) \tag{2.6}$$

Čím vetší stupeň N použijeme, tím více se blížíme ideálnímu jazykovému modelu, protože používáme větší historii předchozích znaků, ale taktéž potřebujeme větší trénovací korpus pro natrénování modelu.

2.3 Trénování modelu

Trénování n-gramového modelu se provádí na větším množství textových dat, na trénovacím korpusu. Tyto textová data se prochází po n-ticích a zaznamenávají se jejich četnosti. Pro bigramový model tak procházíme trénovací korpus po dvojicích znaků, pro trigramový model procházíme trénovací korpus po trojicích znaků, apod. Z těchto četností pak podle zvoleného modelu vypočítáme pravděpodobnosti jednotlivých n-gramů, které uložíme do souboru, což je výsledný n-gramový model. Při identifikaci jazyka tedy natrénujeme model pro každý jazyk, které chceme rozpoznávat, následně pak vyhodnocujeme neznámý text v každém z těchto modelů.

2.4 Vyhodnocování modelu

Při vyhodnocování modelu máme neznámý text, který označíme jako data D. Zavedeme si množinu tříd T_r, kde každá třída odpovídá jednomu jazyku. Snažíme se určit pravděpodobnost $P(T_r|D)$, že daná data D odpovídají třídě T_r jazyka r.

Pravděpodobnost $P(T_r|D)$ počítáme pomocí Bayesova klasifikátoru podle následujícího vztahu.

$$P(T_r|D) = \frac{P(D|T_r) \cdot P(T_r)}{P(D)} \tag{2.7}$$

kde $P(D|T_r)$ je pravděpodobnost, že třída T_r vygenerovala data D
$P(T_r)$ je apriorní pravděpodobnost jazyka
$P(D)$ je pravděpodobnost dat D

Protože pravděpodobnost $P(T_r|D)$ počítáme pro každý model jazyka, a poté je porovnáváme mezi sebou, lze některé části předchozího vztahu zanedbat. Všechny jazyky považujeme za stejně pravděpodobné, proto apriorní pravděpodobnost $P(T_r)$ zanedbáváme. Pravděpodobnost dat $P(D)$ je konstantní vzhledem ke všem jazykům, proto jí také zanedbáváme. Pravděpodobnost $P(D|T_r)$, že třída T_r vygenerovala data D, reprezentuje náš n-gramový model a počítáme jí tak, že postupně projíždíme n-gramy neznámého textu a násobíme jejich pravděpodobnosti v daném modelu, viz kapitola N-gramový model. Tuto pravděpodobnost počítáme pro každý model jazyka a porovnáváme mezi sebou. Model s nejvyšší pravděpodobností pak považujeme za identifikovaný jazyk.

2.5 Nevyhlazený model

Ze vztahu podmíněné pravděpodobnosti (1.3) lze odvodit obecný vztah pro výpočet pravděpodobnosti n-gramu.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1} z_n)}{c(z_{n-N+1}^{n-1})} \tag{2.8}$$

kde $c(z_{n-N+1}^{n-1} z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$

Pro unigramový model získáme následující vztah.

$$P(z) = \frac{c(z)}{N} \tag{2.9}$$

kde $c(z)$ je četnost znaku z, N je celkový počet znaků

Pro bigramový model získáme následující vztah.

$$P(z_n|z_{n-1}) = \frac{c(z_{n-1}, z_n)}{c(z_{n-1})} \tag{2.10}$$

kde $c(z_{n-1}, z_n)$ je četnost dvojice znaků z_{n-1}, z_n
$c(z_{n-1})$ je četnost znaku z_{n-1}

Pro trigramový model získáme následující vztah.

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{c(z_{n-2}, z_{n-1}, z_n)}{c(z_{n-2}, z_{n-1})} \tag{2.11}$$

kde $c(z_{n-2}, z_{n-1}, z_n)$ je četnost trojice znaků z_{n-2}, z_{n-1}, z_n
$c(z_{n-2}, z_{n-1})$ je četnost dvojice znaků z_{n-2}, z_{n-1}

Obdobně pak získáme vztah pro jakýkoliv n-gramový model. Hlavní nevýhodou nevyhlazeného modelu je, že n-gramy, které se nevyskytly v trénovacím korpusu, budou mít nulovou pravděpodobnost a to i ty n-gramy, které jsou možné, ale nevyskytly se z důvodu malého trénovacího korpusu. Tyto nulové pravděpodobnosti pak ovlivňuji vyhodnocení neznámého textu, kde se vyskytne neviděný n-gram. Tento problém řeší vyhlazování modelu.

2.6 Vyhlazování modelu

Vyhlazování řeší problém nulových pravděpodobností n-gramů, které se nevyskytly v trénovacím korpusu. Viděným n-gramům se odebere část pravděpodobnosti a ta se přidělí všem neviděným. Jak velká část se odebere viděným n-gramům závisí na použité vyhlazovací technice. Jednotlivé vyhlazovací techniky jsou popsané v kapitole Vyhlazovací techniky.

Spolu s vyhlazováním nastává otázka, jak pravděpodobnost, kterou jsme odebrali viděným n-gramům, rozdělit mezi neviděné n-gramy. Tento problém popíšu v následující kapitole.

2.7 Typy modelů

V této kapitole popíšu tři typy modelů, jak lze pravděpodobnost, kterou jsme odebrali viděným n-gramům, rozdělit mezi neviděné n-gramy. Následující tabulka obsahuje značení, které použiji ve vztazích pro popis těchto modelů.

Tabulka 2.1 Značení

Značení	Popis
$z_1\_z_n$	Představuje n-gram s prvním znakem z_1, posledním znakem z_n, podtržítko přestavuje žádný, jeden nebo více znaků.
$z_1\_$	Prefix n-gramu $z_1\_z_n$, respektive n-gram $z_1\_z_n$ bez posledního znaku.
$\_z_n$	Sufix n-gramu $z_1\_z_n$, respektive n-gram $z_1\_z_n$ bez prvního znaku.
$c(z_1\_z_n)$	Četnost n-gramu $z_1\_z_n$ v trénovacím korpusu.
$P(z_1\_z_n)$	Pravděpodobnost n-gramu $z_1\_z_n$ v daném modelu.
$P^*(z_1\_z_n)$	Pravděpodobnost n-gramu $z_1\_z_n$ vypočtená z trénovacího korpusu podle zvolené vyhlazovací techniky.
$BOW(z_1\_)$	(Backoff weight) Část pravděpodobnosti získaná zvolenou vyhlazovací technikou z viděných n-gramů s prefixem $z_1\_$.
Z	Počet neviděných n-gramů v trénovacím korpusu.

2.7.1 Model s rovnoměrným rozložením

Tento způsob rovnoměrně rozděluje pravděpodobnost, kterou jsme získali od viděných n-gramů, mezi všechny neviděné n-gramy tak, že každý neviděný n-gram dostane stejnou pravděpodobnost. Tento způsob je nejjednodušší, ale také méně komplexnější.

Tento model je možné popsat následujícími vztahy. Značení použité v těchto vztazích je popsáno v tabulce 2.1.

$$P(z_1\_z_n) = P^*(z_1\_z_n) \qquad když\ c(z_1\_z_n) > 0 \qquad (2.12)$$

$$P(z_1\_z_n) = \frac{1 - \sum_{\substack{n-gram \\ c(z_1\_z_n)>0}} P^*(z_1\_z_n)}{Z} \qquad když\ c(z_1\_z_n) = 0 \qquad (2.13)$$

Čitatel předchozího vztahu představuje pravděpodobnost získanou od všech viděných n-gramů $z_1\_z_n$ pomocí zvolené vyhlazovací techniky, jmenovatel pak počet neviděných n-gramů, mezi které se tato pravděpodobnost rozděluje.

2.7.2 Backoff model

Tento způsob rozděluje pravděpodobnost, kterou jsme získali od viděných n-gramů pomocí zvolené vyhlazovací techniky, mezi všechny neviděné n-gramy podle pravděpodobnosti sufixu n-gramu v modelu o jeden stupeň nižším. Sufixem n-gramu označuji n-gram bez prvního znaku. Takže v případě bigramového modelu se pravděpodobnost neviděným n-gramům bude rozdělovat podle pravděpodobnosti posledního znaku n-gramu v unigramovém modelu. V případě trigramového modelu se pak pravděpodobnost neviděným n-gramům bude rozdělovat podle pravděpodobnosti poslední dvojice znaků n-gramu v bigramovém modelu. Obdobně pak pro modely s vyšším stupněm N. Speciální případ je pak unigramový model, který rozděluje pravděpodobnost neviděným n-gramům podle 0-gramového modelu, což není nic jiného než rovnoměrné rozložení, viz model s rovnoměrným rozložením.

Tento model je možné popsat následujícími vztahy. Značení použité v těchto vztazích je popsáno v tabulce 2.1.

$$P(z_1\_z_n) = P^*(z_1\_z_n) \qquad když\ c(z_1\_z_n) > 0 \tag{2.14}$$

$$P(z_1\_z_n) = BOW(z_1\_) \cdot P(\_z_n) \qquad když\ c(z_1\_z_n) = 0 \tag{2.15}$$

Součet pravděpodobností všech n-gramů musí být roven 1.

$$\sum_{\substack{n-gram \\ c(z_1\_z_n)\geq 0}} P(z_1\_z_n) = 1 \tag{2.16}$$

Pak musí platit následující vztah.

$$\sum_{\substack{n-gram \\ c(z_1\_z_n)>0}} P^*(z_1\_z_n) + \sum_{\substack{n-gram \\ c(z_1\_z_n)=0}} BOW(z_1\_) \cdot P(\_z_n) = 1 \tag{2.17}$$

Pravděpodobnost $BOW(z_1\_)$, která se přiděluje neviděným n-gramům s prefixem $z_1\_$ lze vypočítat podle následujícího vztahu.

$$BOW(z_1\_) = \frac{1 - \sum_{\substack{n-gram \\ c(z_1\_z_n)>0}} P^*(z_1\_z_n)}{\sum_{\substack{n-gram \\ c(z_1\_z_n)=0}} P(\_z_n)} \tag{2.18}$$

Čitatel předchozího vztahu představuje pravděpodobnost získanou od všech viděných n-gramů $z_1\_z_n$ pomocí zvolené vyhlazovací techniky, jmenovatel pak představuje součet všech pravděpodobností sufixů neviděných n-gramů, tak aby při výpočtu pravděpodobnosti neviděného n-gramu podle vtahu (2.15) dostal každý n-gram správnou část z $BOW(z_1\_)$ podle své pravděpodobnosti $P(\_z_n)$.

2.7.3 Interpolovaný model

Tento model je podobný jako model backoff, rozdíl je v tom, že zatímco model backoff používá pro přidělení pravděpodobnosti model s nižším stupněm jen u neviděných n-gramů, interpolovaný model používá pro přidělení pravděpodobnosti model s nižším stupněm i pro viděné n-gramy. Pravděpodobnost získaná pomocí zvolené vyhlazovací techniky tak není rozdělena pouze mezi neviděné n-gramy, ale je rozdělena mezi všechny n-gramy. Viděným n-gramům se tak vrátí část pravděpodobnosti, které jim byla odebrána vyhlazovací technikou, neviděné n-gramy tak dostanou menší pravděpodobnost než v modelu backoff.

Tento model je možné popsat následujícími vztahy. Značení použité v těchto vztazích je popsáno v tabulce 2.1.

$$P(z_1\_z_n) = P^*(z_1\_z_n) + BOW(z_1\_) \cdot P(\_z_n) \tag{2.19}$$

Součet pravděpodobností všech n-gramů musí být roven 1.

$$\sum_{\substack{n-gram \\ c(z_1\_z_n)>0}} P^*(z_1\_z_n) + \sum_{\substack{n-gram \\ c(z_1\_z_n)\geq 0}} BOW(z_1\_) \cdot P(\_z_n) = 1 \tag{2.20}$$

Pravděpodobnost $BOW(z_1\_)$, která se přiděluje jak neviděným tak viděným n-gramům s prefixem $z_1\_$ lze vypočítat podle následujícího vztahu.

$$BOW(z_1\_) = 1 - \sum_{\substack{n-gram \\ c(z_1\_z_n)>0}} P^*(z_1\_z_n) \tag{2.21}$$

2.8 Praktické vytváření modelů

Vzhledem k tomu, že při vyhodnocování modelu násobíme pravděpodobnosti jednotlivých n-gramů neznámého textu a tyto pravděpodobnosti jsou velmi malé hodnoty, tak by při násobení těchto hodnot docházelo k podtečení proměnné v paměti počítače, a tak by výsledky těchto součinů kvůli zaokrouhlování byly nulové. V praxi se proto místo pravděpodobností n-gramů počítá s logaritmy pravděpodobností n-gramů a místo násobení pravděpodobností se sčítají logaritmy pravděpodobností. Protože je logaritmus funkce rostoucí a porovnáváme součty logaritmů pravděpodobností n-gramů, nedochází ke zkreslení výsledků, dostáváme stejné výsledky, jako kdybychom porovnávali součiny pravděpodobností n-gramů. Při sčítání logaritmů pravděpodobností n-gramů už nedochází k podtečení proměnné.

S logaritmy pravděpodobností n-gramů se počítá už při vytváření jazykových modelů, takže samotné hodnoty pravděpodobností n-gramů už jsou v jazykovém modelu zlogaritmované.

3 Vyhlazovací techniky

Postupně zde popíšu základní vyhlazovací techniky, které se používají. Většinu z nich jsem použil pro porovnání modelů při identifikace jazyka textového dokumentu.

3.1 Additive smoothing

Additive smoothing, někdy označované také jako Lidstone smoothing, je vyhlazovací technika založená na přičítání konstanty λ ke každé četnosti n-gramu. Obecný vztah pro výpočet pravděpodobnosti n-gramu při tomto vyhlazování a s rovnoměrným rozložením neviděným n-gramům vypadá takto.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1}, z_n) + \lambda}{c(z_{n-N+1}^{n-1}) + \lambda \cdot V} \tag{3.1}$$

kde $c(z_{n-N+1}^{n-1}, z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$
V je počet znaků slovníku

Pro unigramový model získáme následující vztah.

$$P(z) = \frac{c(z) + \lambda}{N + \lambda \cdot V} \tag{3.2}$$

kde $c(z)$ je četnost znaku z, N je celkový počet znaků, V je počet znaků slovníku

Pro bigramový model získáme následující vztah.

$$P(z_n|z_{n-1}) = \frac{c(z_{n-1}, z_n) + \lambda}{c(z_{n-1}) + \lambda \cdot \mathrm{V}} \tag{3.3}$$

kde $c(z_{n-1}, z_n)$ je četnost dvojice znaků z_{n-1}, z_n
$c(z_{n-1})$ je četnost znaku z_{n-1}
V je počet znaků slovníku

Pro 3-gramový model získáme následující vztah.

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{c(z_{n-2}, z_{n-1}, z_n) + \lambda}{c(z_{n-2}, z_{n-1}) + \lambda \cdot \mathrm{V}} \tag{3.4}$$

kde $c(z_{n-2}, z_{n-1}, z_n)$ je četnost trojice znaků z_{n-2}, z_{n-1}, z_n
$c(z_{n-2}, z_{n-1})$ je četnost dvojice znaků z_{n-2}, z_{n-1}
V je počet znaků slovníku

Tato vyhlazovací technika závisí na zvolené velikosti konstanty λ. Pokud zvolíme $\lambda < 1$, více důvěřujeme pravděpodobnosti viděných n-gramů a odebíráme jim jen malou pravděpodobnost, kterou pak rozdělujeme neviděným n-gramům. Naopak, pokud zvolíme

$\lambda > 1$, méně důvěřujeme pravděpodobnosti viděných n-gramů, odebíráme jim větší část pravděpodobnosti, kterou pak rozdělujeme neviděným n-gramům. Speciálním případem je pak $\lambda = 1$, které se označuje jako Add-One smoohthing nebo také Laplace smoothing. Toto vyhlazování jsem použil pro porovnání modelů.

3.2 Witten-Bell discounting

Tato vyhlazovací technika je založena na myšlence, že n-gramům s nulovou četností v trénovacím korpusu by měla být přidělena pravděpodobnost, že se objeví nový n-gram. Pravděpodobnost, že se objeví nový n-gram je dána počtem různých n-gramů, které se objevily v trénovacím korpusu, protože každý z těchto n-gramů se objevil poprvé právě jednou. Tuto pravděpodobnost můžeme vyjádřit následujícím vztahem.

$$P_{nový\ n-gram} = \frac{T}{N+T} \tag{3.5}$$

kde T je počet různých n-gramů s daným prefixem znaků
N je počet všech n-gramů s daným prefixem znaků

Prefixem n-gramu uvažuji n-gram bez posledního znaku. Počet všech n-gramů s daným prefixem není nic jiného než počet výskytu daného prefixu, proto se v následujících vztazích neobjevuje značení N, které je nahrazeno četnostmi c, s vyjímkou unigramového modelu.

Následující vztahy vyjadřují obecné vztahy pro pravděpodobnost n-gramu při vyhlazování Witten-Bell discounting a s rovnoměrným rozložením pravděpodobnosti neviděným n-gramům.

$$P(z_n | z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1}, z_n)}{c(z_{n-N+1}^{n-1}) + T(z_{n-N+1}^{n-1})} \qquad když\ c(z_{n-N+1}^{n-1}, z_n) > 0 \tag{3.6}$$

$$P(z_n | z_{n-N+1}^{n-1}) = \frac{T(z_{n-N+1}^{n-1})}{Z(z_{n-N+1}^{n-1}) \cdot (c(z_{n-N+1}^{n-1}) + T(z_{n-N+1}^{n-1}))} \qquad když\ c(z_{n-N+1}^{n-1}, z_n) = 0 \tag{3.7}$$

kde $T(z_{n-N+1}^{n-1})$ představuje počet všech různých n-gramů s prefixem znaků z_{n-N+1}^{n-1}
$Z(z_{n-N+1}^{n-1})$ představuje počet všech neviděných n-gramů s prefixem znaků z_{n-N+1}^{n-1}
$c(z_{n-N+1}^{n-1}, z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$

Pro unigramový model získáme následující vztahy.

$$P(z) = \frac{T}{Z \cdot (N + T)} \qquad když\ c(z) = 0 \tag{3.8}$$

$$P(z) = \frac{c(z)}{N + T} \qquad když\ c(z) > 0 \tag{3.9}$$

kde T je počet všech různých znaků, Z je počet neviděných znaků,
N je počet všech znaků, $c(z)$ je četnost znaku z

Pro bigramový model získáme následující vztahy.

$$P(z_n|z_{n-1}) = \frac{T(z_{n-1})}{Z(z_{n-1}) \cdot (c(z_{n-1}) + T(z_{n-1}))} \qquad když\ c(z_{n-1}, z_n) = 0 \tag{3.10}$$

$$P(z_n|z_{n-1}) = \frac{c(z_{n-1}, z_n)}{c(z_{n-1}) + T(z_{n-1})} \qquad když\ c(z_{n-1}, z_n) > 0 \tag{3.11}$$

kde $T(z_{n-1})$ představuje počet všech různých n-gramů s prefixem z_{n-1}
$Z(z_{n-1})$ představuje počet všech neviděných n-gramů s prefixem z_{n-1}
$c(z_{n-1}, z_n)$ je četnost dvojice znaků z_{n-1}, z_n
$c(z_{n-1})$ je četnost znaku z_{n-1}

Pro trigramový model získáme následující vztahy.

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{T(z_{n-2}, z_{n-1})}{Z(z_{n-2}, z_{n-1}) \cdot (c(z_{n-2}, z_{n-1}) + T(z_{n-2}, z_{n-1}))} \qquad když\ c(z_{n-2}, z_{n-1}, z_n) = 0 \tag{3.12}$$

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{c(z_{n-2}, z_{n-1}, z_n)}{c(z_{n-2}, z_{n-1}) + T(z_{n-2}, z_{n-1})} \qquad když\ c(z_{n-2}, z_{n-1}, z_n) > 0 \tag{3.13}$$

kde $T(z_{n-2}, z_{n-1})$ představuje počet všech různých n-gramů s prefixem znaků z_{n-2}, z_{n-1}
$Z(z_{n-2}, z_{n-1})$ představuje počet všech neviděných n-gramů s prefixem znaků z_{n-2}, z_{n-1}
$c(z_{n-2}, z_{n-1}, z_n)$ je četnost trojice znaků z_{n-2}, z_{n-1}, z_n
$c(z_{n-2}, z_{n-1})$ je četnost dvojic znaků z_{n-2}, z_{n-1}

3.3 Good-Turing discounting

Tato vyhlazovací technika je založena na tom, že n-gramům s nulovou nebo malou četností přiřazujeme pravděpodobnost podle počtu n-gramů s vyšší četností. Následující vztah vyjadřuje přepočet četností zjištěných z trénovacího korpusu na četnosti, které odpovídají této vyhlazovací technice.

$$c^* = (c+1) \cdot \frac{N_{c+1}}{N_c} \tag{3.14}$$

kde c^* je přepočtená četnost, c je četnost z trénovacího korpusu,
N_{c+1} je počet n-gramů s četností $c+1$, N_c je počet n-gramů s četností c

V praxi se nepřepočítávají všechny četnosti, které se vyskytly v trénovacím korpusu, ale jen četnosti do určité hodnoty prahu k. Od tohoto prahu k se pak četnosti považují za spolehlivé. Následující vztahy toto vyjadřují.

$$c^* = c \qquad když\ c > k \tag{3.15}$$

$$c^* = \frac{(c+1) \cdot \frac{N_{c+1}}{N_c} - c \cdot \frac{(k+1) \cdot N_{k+1}}{N_1}}{1 - \frac{(k+1) \cdot N_{k+1}}{N_1}} \qquad když\ 1 \leq c \leq k \tag{3.16}$$

kde k je práh, od kterého se četnosti nepřepočítávají,
c^* je přepočtená četnost, c je četnost z trénovacího korpusu,
N_{c+1} je počet n-gramů s četností $c+1$, N_c je počet n-gramů s četností c,
N_{k+1} je počet n-gramů s četností $k+1$, N_1 je počet n-gramů s četností 1

Pravděpodobnost n-gramů se pak vypočte podle vztahu

$$P(z_n | z_{n-N+1}^{n-1}) = \frac{c^*(z_{n-N+1}^{n-1} z_n)}{V} \tag{3.17}$$

kde $c^*(z_{n-N+1}^{n-1} z_n)$ je přepočtená četnost, V je počet znaků slovníku

3.4 Absolute discounting

Absolute discounting je jednoduchá vyhlazovací metoda založená na odečítání konstanty D od každé četnosti n-gramu, která je větší než nula. Hodnota konstanty D by se měla pohybovat v rozmezí 0 až 1. Následuje obecný vztah pro tuto vyhlazovací techniku.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1}z_n) - D}{c(z_{n-N+1}^{n-1})} \quad když\ c(z_{n-N+1}^{n-1}z_n) > 0 \tag{3.18}$$

kde $c(z_{n-N+1}^{n-1}z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$

Pro unigramový model získáme následující vztah.

$$P(z) = \frac{c(z) - D}{N} \quad když\ c(z) > 0 \tag{3.19}$$

kde $c(z)$ je četnost znaku z, N je celkový počet znaků

Pro bigramový model získáme následující vztah.

$$P(z_n|z_{n-1}) = \frac{c(z_{n-1}, z_n) - D}{c(z_{n-1})} \quad když\ c(z_{n-1}, z_n) > 0 \tag{3.20}$$

kde $c(z_{n-1}, z_n)$ je četnost dvojice znaků z_{n-1}, z_n
$c(z_{n-1})$ je četnost znaku z_{n-1}

Pro trigramový model získáme následující vztah.

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{c(z_{n-2}, z_{n-1}, z_n) - D}{c(z_{n-2}, z_{n-1})} \quad když\ c(z_{n-2}, z_{n-1}, z_n) > 0 \tag{3.21}$$

kde $c(z_{n-2}, z_{n-1}, z_n)$ je četnost trojice znaků z_{n-2}, z_{n-1}, z_n
$c(z_{n-2}, z_{n-1})$ je četnost dvojice znaků z_{n-2}, z_{n-1}

Doporučený vzorec pro výpočet konstanty D je následující.

$$D = \frac{N_1}{N_1 + 2 \cdot N_2} \tag{3.22}$$

kde N_1 je počet n-gramů s četností 1, N_2 je počet n-gramů s četností 2

3.5 Kneser-Ney discounting

Kneser-Ney Discounting je založeno na myšlence, že nejvyšší stupeň n-gramového modelu je počítán jinak než ostatní nižší stupně. Tyto nižší stupně se v případě backoff modelu používají pro výpočet pravděpodobnosti neviděných n-gramů a v případě interpolovaného modelu se používají pro výpočet viděných i neviděných n-gramů, viz kapitola Typy modelů.

Pro vyhlazování nejvyššího stupně se používá Absolute discounting, kde se od četností n-gramů větších než nula odečítá konstanta D.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1}z_n) - D}{c(z_{n-N+1}^{n-1})} \quad když\ c(z_{n-N+1}^{n-1}z_n) > 0 \tag{3.23}$$

kde $c(z_{n-N+1}^{n-1}z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$

Původní verze Kneser-Ney discounting používá pro výpočet konstanty D doporučený vztah (3.22). Chen and Goodman modifikace Kneser-Ney Discounting používá tři různé konstanty pro odečítání, D_1 pro n-gramy s četností jedna, D_2 pro n-gramy s četností dva, D_{3+} pro n-gramy s četností tři a více. Následující vztahy vyjadřují tyto konstanty.

$$Y = \frac{N_1}{N_1 + 2 \cdot N_2} \tag{3.24}$$

$$D_1 = 1 - 2 \cdot Y \cdot \frac{N_2}{N_1} \tag{3.25}$$

$$D_2 = 2 - 3 \cdot Y \cdot \frac{N_3}{N_2} \tag{3.26}$$

$$D_{3+} = 3 - 4 \cdot Y \cdot \frac{N_4}{N_3} \tag{3.27}$$

kde N_1 je počet n-gramů s četností 1, N_2 je počet n-gramů s četností 2
kde N_3 je počet n-gramů s četností 3, N_4 je počet n-gramů s četností 4

Pro výpočet ostatních nižších stupňů n-gramu se používá následující vztah.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{T(*, z_{n-N+1}^{n-1}, z_n) - D}{T(*, z_{n-N+1}^{n-1}, *)} \tag{3.28}$$

kde $T(*, z_{n-N+1}^{n-1}, z_n)$ představuje počet různých znaků, které se vyskytly před n-gramem, $T(*, z_{n-N+1}^{n-1}, *)$ představuje počet různých k-tic, složených z prefixu n-gramu a jakéhokoliv znaku před a za prefixem, které se vyskytly v trénovacím korpusu

Tento výpočet je založen na myšlence, že výpočet pravděpodobnosti nižších n-gramů není založen na četnostech, ale na počtu unikátních slov, které předcházejí n-gramu. Důvod, proč tímto způsobem počítat pravděpodobnost nižších stupňů n-gramu lze nastínit na následujícím příkladu. V případě, že máme dvojici slov, která se ve většině případů vyskytuje spolu,

např. dvojice „San Francisco“ a jejich četnost v korpusu je vysoká, mají obě slova „San“ i „Francisco“ vysokou pravděpodobnost v unigramovém modelu, i když slovo „Francisco“ se samostatně v korpusu nebude téměř vyskytovat, vždy pouze za slovem „San“. Pokud tedy použijeme výpočet pravděpodobnosti založený na počtu unikátních slov, které předcházejí před n-gramem, tak pravděpodobnost slova „Francisco“ bude nízká, což je správně. V případě identifikace jazyka, kdy nepoužíváme slova, ale znaky, tato myšlenka nemusí být opodstatněná, ale analogicky lze uvažovat zápis spřežek, ale to jen v případě, že by se druhý znak spřežky samostatně příliš nepoužíval. Zda je tato metoda vhodná pro identifikaci jazyka ukáží samotné výsledky těchto modelů s porovnáním s ostatními.

3.6 Ristad's natural discounting

Představme si, že máme řetězec X^N délky N, který je složen ze znaků konečné abecedy A. Abeceda A má V různých znaků, kde každý znak představuje pozorovaný jev, a řetězec X^N pak představuje historii pozorovaných jevů délky N. Našim cílem je určit pravděpodobnost $P(i|\{c_i\}, N)$, tedy pravděpodobnost, že bude následovat jev i za předpokladu, že známe historii N jevů, tedy množinu četností jevů $\{c_i\}$.

Předpoklad je, že jednodušší řetězce jsou více pravděpodobné než složité. Dále pak, že řetězec neobsahuje všechny znaky z abecedy A, ale pouze některé znaky, tedy znaky z podmnožiny abecedy A. Lze uvažovat dvě různé interpretace omezení, že řetězec je tvořen znaky z podmnožiny abecedy A. První interpretace je, že všechny neprázdné podmnožiny abecedy A jsou stejně pravděpodobné. Druhá interpretace je, že všechny nenulové kardinality podmnožin jsou stejně pravděpodobné, tzn. každá kardinalita dostane stejnou část pravděpodobnosti, takže např. podmnožiny s kardinalitou 1 budou mít v součtu stejnou pravděpodobnost jako podmnožiny s kardinalitou 2 apod. Ve SRILM je implementována pouze druhá interpretace tohoto omezení, proto se v následujícím popisu omezím jen na tuto interpretaci, protože tu jsem použil při identifikaci jazyka.

Pokud přiřadíme rovnoměrnou pravděpodobnost všem nenulovým kardinalitám podmnožin abecedy A, dostaneme následující vzorec.

$$P(X^N|N) = \left(\min(V, N) \cdot \binom{V}{T} \cdot \binom{N-1}{T-1} \cdot \binom{N}{\{c_i\}}\right)^{-1} \tag{3.29}$$

První část $1/\min(V, N)$ představuje rovnoměrnou pravděpodobnost přes kardinality, tedy $\min(V, N)$ představuje počet možných kardinalit podmnožin abecedy A. Druhá část $1/\binom{V}{T}$ představuje rovnoměrnou pravděpodobnost podmnožin abecedy A s určitou kardinalitou T. Třetí část $1/\binom{N-1}{T-1}$ představuje rovnoměrnou pravděpodobnost přes množinu četností $\{c_i\}$ pro

vybranou podmnožinu A. Čtvrtá část $1/\binom{N}{\{c_i\}}$ reprezentuje rovnoměrnou pravděpodobnost všech řetězců, které obsahují přesně c_i výskytů znaku i pro všechny znaky i.

Podmíněná pravděpodobnost $P(i|X^N, N)$, že přijde znak i po řetězci X^N, lze vypočítat jako pravděpodobnost, že přijde znak i po řetězci X^N vzhledem k pravděpodobnosti, že přijde jakýkoliv znak j z abecedy A.

$$P(i|X^N, N) = \frac{P(X^N i|N+1)}{\sum_{j=1}^{V} P(X^N j|N+1)} \tag{3.30}$$

Vzhledem k tomu, že tato pravděpodobnost závisí pouze na četnostech znaků $\{c_j\}$ v řetězci X^N, lze napsat:

$$P\left(i|\{c_j\}, N\right) = P(i|X^N, N) \tag{3.31}$$

Algebraickou úpravou (3.28) pak získáme následující vztah Natural law:

$$P(i|\{c_i\}, N) = \begin{cases} (c_i + 1)/(N+V) & když\ T = V \\ T(T+1)/(V-T)(N^2+N+2\cdot T) & když\ T < V \wedge c_i = 0 \\ (c_i+1)(N+1-T)/(N^2+N+2\cdot T) & když\ T < V \wedge c_i > 0 \end{cases} \tag{3.32}$$

V praxi se využívá následující upravený vztah, který se více blíží nevyhlazenému modelu $P(i|\{c_i\}, N) = c_i/N$ a zároveň zachovává hlavní vlastnost, že přiřazuje malou pravděpodobnost novým jevům.

$$P(i|\{c_i\}, N) = \begin{cases} \dfrac{c_i}{N} & když\ T = V \\ \dfrac{1}{V-T} \cdot \dfrac{T\cdot(T+1)}{N^2+N+2\cdot T} & když\ T < V \wedge c_i = 0 \\ \dfrac{c_i}{N} \cdot \dfrac{N(N+1)+T\cdot(1-T)}{N^2+N+2\cdot T} & když\ T < V \wedge c_i > 0 \end{cases} \tag{3.33}$$

Ve SRILM je pak implementován backoff model tohoto vyhlazování. Následuje obecný vztah pro výpočet pravděpodobnosti n-gramu.

$$P(z_n|z_{n-N+1}^{n-1}) = \frac{c(z_{n-N+1}^{n-1}, z_n)}{c(z_{n-N+1}^{n-1})} \frac{c(z_{n-N+1}^{n-1})\cdot(c(z_{n-N+1}^{n-1})+1) + T(z_{n-N+1}^{n-1})\cdot(1-T(z_{n-N+1}^{n-1}))}{c(z_{n-N+1}^{n-1})^2 + c(z_{n-N+1}^{n-1}) + 2\cdot T(z_{n-N+1}^{n-1})} \tag{3.34}$$
$$když\ c(z_{n-N+1}^{n-1}, z_n) > 0$$

kde $T(z_{n-N+1}^{n-1})$ představuje počet všech různých n-gramů s prefixem znaků $z_{n-N+1}, \dots, z_{n-1}$
$c(z_{n-N+1}^{n-1}, z_n)$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}, z_n$
$c(z_{n-N+1}^{n-1})$ představuje četnost sekvence znaků $z_{n-N+1}, \dots, z_{n-1}$

Pro unigramový model získáme následující vztah.

$$P(z) = \frac{c(z)}{N} \cdot \frac{N \cdot (N+1) + T \cdot (1-T)}{N^2 + N + 2 \cdot T} \tag{3.35}$$

kde T je počet různých znaků, N je počet všech znaků, $c(z)$ je četnost znaku z

Pro bigramový model získáme následující vztah.

$$P(z_n|z_{n-1}) = \frac{c(z_{n-1}, z_n)}{c(z_{n-1})} \cdot \frac{c(z_{n-1}) \cdot (c(z_{n-1}) + 1) + T(z_{n-1}) \cdot (1 - T(z_{n-1}))}{c(z_{n-1})^2 + c(z_{n-1}) + 2 \cdot T(z_{n-1})} \tag{3.36}$$

kde $c(z_{n-1})$ je četnost znaku z_{n-1}
$c(z_{n-1}, z_n)$ je četnost dvojice znaků z_{n-1}, z_n
$T(z_{n-1})$ představuje počet všech různých n-gramů s prefixem z_{n-1}

Pro trigramový model získáme následující vztah.

$$P(z_n|z_{n-2}, z_{n-1}) = \frac{c(z_{n-2}^{n})}{c(z_{n-2}^{n-1})} \cdot \frac{c(z_{n-2}^{n-1}) \cdot (c(z_{n-2}^{n-1}) + 1) + T(z_{n-2}^{n-1}) \cdot (1 - T(z_{n-2}^{n-1}))}{c(z_{n-2}^{n-1})^2 + c(z_{n-2}^{n-1}) + 2 \cdot T(z_{n-2}^{n-1})} \tag{3.37}$$

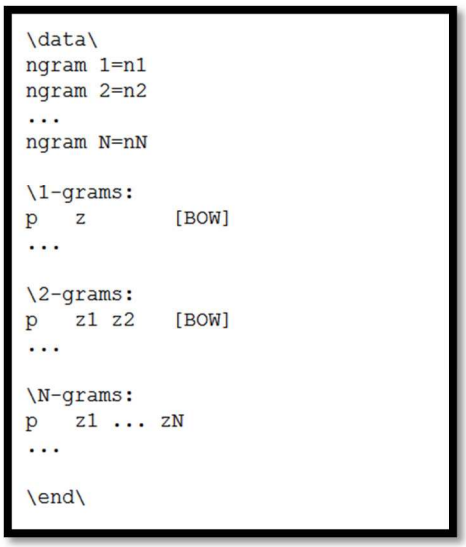
kde $c(z_{n-2}^{n-1})$ je četnost dvojice znaků z_{n-2}, z_{n-1}
$c(z_{n-2}^{n})$ je četnost trojice znaků z_{n-2}, z_{n-1}, z_n
$T(z_{n-2}^{n-1})$ představuje počet všech různých n-gramů s prefixem z_{n-2}, z_{n-1}

4 Projekt SRILM

SRILM (The SRI Language Modeling Toolkit) je sada nástrojů používána pro práci se statistickým jazykovým modelem. Je vyvíjena od roku 1995 v laboratoři Speech Technology and Research (STAR).

4.1 Vytváření jazykového modelu

Pro vytváření modelu se používá aplikace „ngram-count". Ta má v sobě implementován back-off a interpolovaný model, který je popsán v kapitole Typy modelů. Dále má v sobě implementované vyhlazovací techniky, které jsou popsané výše, viz kapitola Vyhlazovací techniky. Jazykový model se ukládá ve formátu ARPA backoff. Ten vypadá následovně.



Obrázek 4.1 Formát ARPA backoff

V horní části jsou obsaženy počty různých n-gramů daného stupně, které byly viděny v trénovacím korpusu. V dalších částech jsou pak samotné viděné n-gramy a jejich pravděpodobnosti v daném modelu a také jejich backoff váhy BOW (Backoff Weight), které slouží k výpočtu pravděpodobností n-gramů v backoff a interpolovaném modelu, viz kapitola Typy modelů.

4.2 Vyhodnocování jazykového modelu

Pro vyhodnocování jazykového modelu se používá aplikace „ngram“, které předáme jazykový model a rozpoznávaný text, u kterého chceme určit jeho pravděpodobnost v daném modelu.

5 Textová data

Textová data pro trénování i testování modelů jsem stahoval ze zpravodajských serverů daných zemí, tak abych získal příslušné texty většiny evropských jazyků.

Pomocí regulárních výrazů jsem získal odkazy na články daných serveru, které jsem pak projížděl. V případě nedostatku odkazů, jsem použil jiný server. Většinu textů daného jazyka jsem tak získal z více serverů. Některé jazyky nebylo možné stáhnout z důvodu nedostatku vhodných serverů, ze kterých by bylo možné stahovat texty.

Textová data jsem získal pomocí regulárního výrazu z článků serverů. Pro každý jazyk jsem si udělal množinu velkých a malých znaků, které se v daném jazyce mohou vyskytnout. Jako základ jsem použil základní latinku, tedy 26 znaků bez diakritiky, a k těmto znakům jsem připojil příslušné znaky daného jazyka, tedy znaky s diakritikou, případně jiné. Dále jsem měl množinu znaků, které se běžně vyskytují ve větách, jako je čárka, uvozovky, pomlčky apod. a množinu číslic. Pomocí těchto množin znaků jsem pak vygeneroval regulární výraz pro věty příslušného jazyka, který jsem použil pro získání textových dat.

V následující tabulce jsou jazyky, u kterých se mi podařilo získat dostatečné množství textových dat.

Tabulka 5.1 Získané jazyky

	Jazyk		Jazyk
1	albánština	18	maďarština
2	angličtina	19	makedonština
3	baskičtina	20	němčina
4	běloruština	21	norština
5	bulharština	22	polština
6	čeština	23	portugalština
7	dánština	24	rumunština
8	estonština	25	ruština
9	finština	26	řečtina
10	francouzština	27	slovenština
11	holandština	28	slovinština
12	chorvatština	29	srbština
13	islandština	30	španělština
14	italština	31	švédština
15	katalánština	32	turečtina
16	litevština	33	ukrajinština
17	lotyština	34	vietnamština

V následující tabulce jsou jazyky, které jsem se pokoušel získat, ale z důvodu nedostatečného zdroje dat se mi nepodařilo získat dostatečné množství textových dat.

Tabulka 5.2 Nezískané jazyky

	Jazyk
1.	galicijština
2.	irština
3.	maltština
4.	velština

V následující tabulce jsou pak zpravodajské servery, ze kterých jsem získal nejvíce textů příslušného jazyka.

Tabulka 5.3 Hlavní zdroje dat jednotlivých jazyků

	Jazyk	Server
1	albánština	http://infoalbania.org
2	angličtina	http://www.guardian.co.uk
3	baskičtina	http://sustatu.com
4	běloruština	http://www.svaboda.org
5	bulharština	http://www.btv.bg
6	čeština	http://www.novinky.cz
7	dánština	http://www.bt.dk
8	estonština	http://uudised.err.ee
9	finština	http://www.uusisuomi.fi
10	francouzština	http://www.lefigaro.fr
11	holandština	http://www.volkskrant.nl
12	chorvatština	http://dnevnik.hr
13	islandština	http://www.visir.is
14	italština	http://www.corriere.it
15	katalánština	http://www.europapress.cat
16	litevština	http://www.diena.lt
17	lotyština	http://www.tvnet.lv
18	maďarština	http://inforadio.hu
19	makedonština	http://daily.mk
20	němčina	http://www.welt.de
21	norština	http://www.adressa.no
22	polština	http://www.wprost.pl
23	portugalština	http://www.publico.pt
24	rumunština	http://www.antena3.ro
25	ruština	http://www.dni.ru
26	řečtina	http://www.sigmalive.com
27	slovenština	http://www.sme.sk
28	slovinština	http://www.rtvslo.si
29	srbština	http://serbian.ruvr.ru
30	španělština	http://www.abc.es
31	švédština	http://www.jnytt.se
32	turečtina	http://www.zaman.com.tr
33	ukrajinština	http://www.expres.ua
34	vietnamština	http://dantri.com.vn

Protože jsem textová data stahoval ze zpravodajských serverů v daných zemích, nebylo zaručeno, že zcela všechny získané texty jsou v očekávaném jazyce, ale dalo se předpokládat, že některé texty budou v jazyce jiném, hlavně v angličtině, případně v jazyce sousedních států. Proto bylo nutné tyto texty ještě profiltrovat, tak aby výsledné texty byly co nejpřesnější.

Protože stažené (nefiltrované) texty obsahovaly převážně texty v předpokládaném jazyce a jen několik málo vět v jiném jazyce, bylo možné je přímo použít pro odfiltrování. Ze stažených textů jsem si tedy vytvořil n-gramové modely pro příslušné jazyky a pomocí aplikace „Language Recognizer“, kterou jsem vytvořil právě pro filtrování textů podle jazyka (popsána v kapitole Vytvořené aplikace), jsem z těchto textů odfiltroval ostatní jazyky. Filtrování sice nefunguje úplně stoprocentně, je závislé na kvalitě natrénovaných modelů a také na délce filtrovaných textů, ale protože jsem použil dlouhé věty délky 100 a více znaků, tak lze předpokládat, že většina nesprávných dat se odfiltrovala. Vyfiltrovaná textová data jsem později použil pro testování modelů, viz následující kapitola.

6 Výsledky

6.1 Způsob testovaní

Pro každý jazyk jsem měl k dispozici přibližně 100 tisíc různých vět délky 100 a více znaků. Z těchto vět jsem náhodně vybral 10 tisíc vět pro testování modelů a ze zbylých vět jsem vzal 11 milionů znaků pro trénování modelů.

Testování probíhalo na testovacích řetězcích dané délky, které jsem zvolil v délkách 5, 10, 15, 20, 25, 30, 40, 50 a 100 znaků. Tyto testovací řetězce jsem generoval z testovacích vět tak, že jsem náhodně vybral část testovací věty dané délky, ale tak aby začínala na začátku slova.

Samotné testování probíhalo tak, že jsem vzal 10 tisíc testovacích řetězců určitého jazyka a dané délky a určil jsem úspěšnost identifikace daného jazyka, neboli kolik z těchto testovacích řetězců bylo identifikováno jako daný jazyk. Toto jsem udělal pro každý jazyk, který rozpoznávám, a úspěšnost identifikace těchto jazyků jsem zprůměroval. Takto jsem získal průměrnou úspěšnost identifikace jazyka, kterou porovnávám a zobrazuji v grafech.

6.2 Porovnání typů modelů

V této části porovnám modely podle jejich typu, viz kapitola Typy modelů. Budu zde tedy porovnávat model s rovnoměrným rozložením, backoff model a interpolovaný model. Jako vyhlazovací techniku jsem zvolil Witten-Bell discounting.

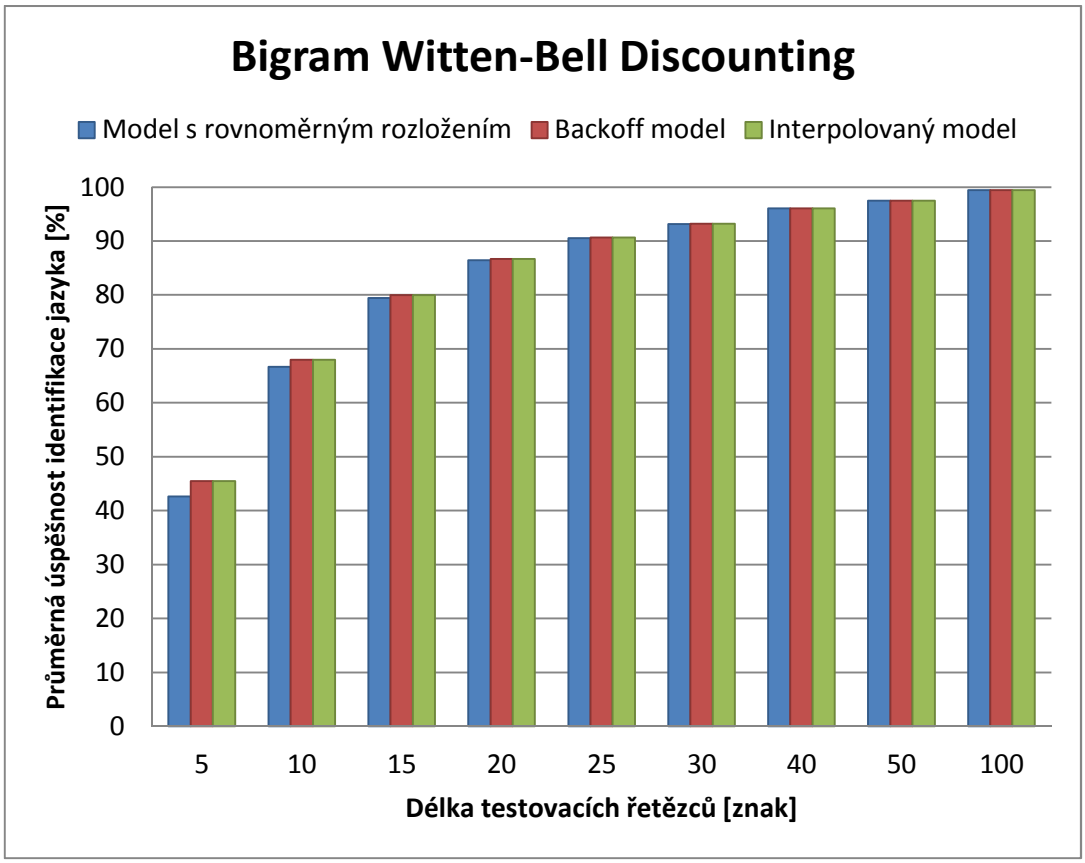
Pro backoff model a interpolovaný model jsem využil projekt SRILM, který je má v sobě implementované. Model s rovnoměrným rozložením jsem naprogramoval sám, vytvořil jsem unigramový, bigramový a trigramový model.

6.2.1 Porovnání výsledků

Pro unigramový model je model backoff stejný jako model s rovnoměrným rozložením, protože rozděluje pravděpodobnost získanou vyhlazovací technikou mezi neviděné n-gramy podle předchozího nižšího stupně, v tomto případě podle 0-gramového modelu, tedy rovnoměrně. Interpolovaný model rozděluje pravděpodobnost získanou vyhlazovací technikou všem n-gramům, tedy viděným i neviděným, takže se částečně liší od modelu backoff, ale tento rozdíl se ve výsledkách neprojevil.

Průměrná úspěšnost identifikace jazyka u všech 3 typů unigramových modelů tak byla stejná. Už před testem jsem předpokládal podobný výsledek, takže hlavním cílem tohoto testu bylo spíše ověřit, že jsem model s rovnoměrným rozložením naprogramoval správně.

Následující graf znázorňuje výsledky bigramových modelů.



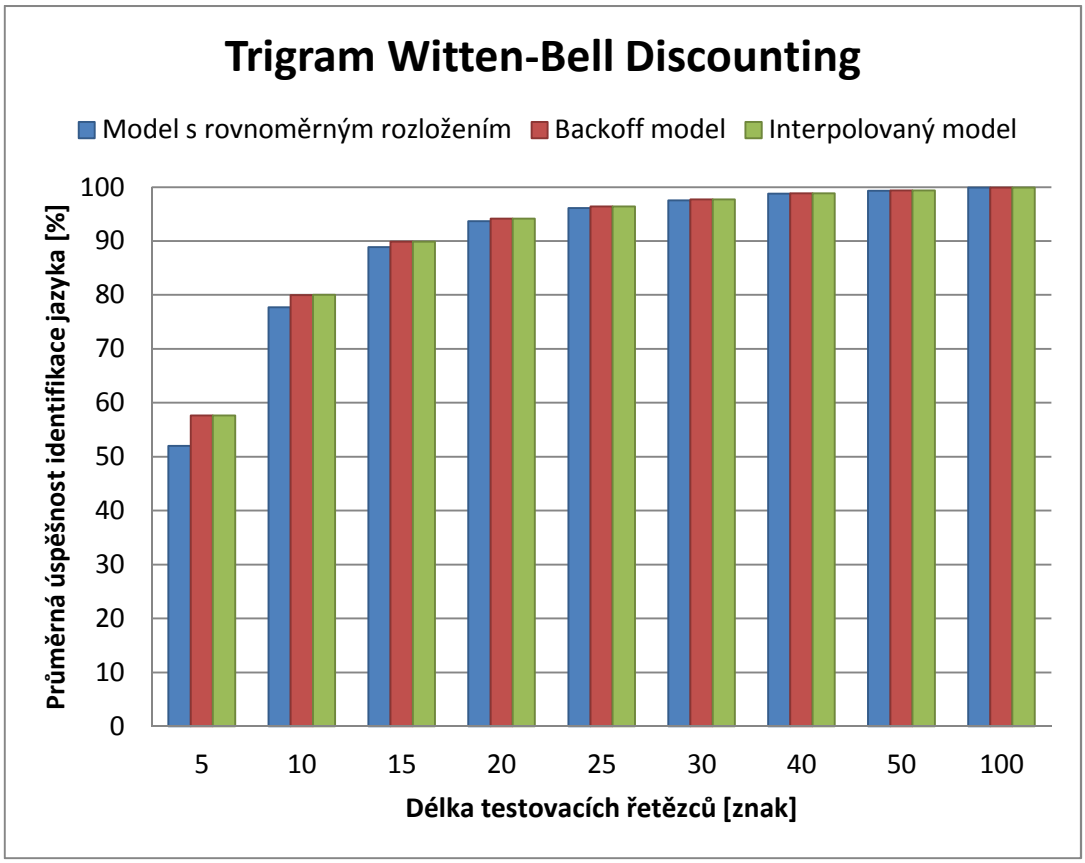
Graf 6.1 Porovnání typů modelů na bigramových modelech

Výsledky bigramových modelů ukazují, že model s rovnoměrným rozložením má proti backoff a interpolovanému modelu horší úspěšnost identifikace jazyka a to hlavně pro malé délky testovacích řetězců. U délky 5 znaků je rozdíl největší a je přibližně 3%, s rostoucí délkou testovacích řetězců se pak rozdíl snižuje, takže pro délky nad 50 znaků je výsledek totožný pro všechny modely.

U těchto výsledků se projevila skutečnost, že backoff a interpolovaný model využívají více informací oproti modelu s rovnoměrným rozložením, protože pro určení pravděpodobností svých n-gramů využívají bigramový i unigramový model, naopak model s rovnoměrným rozložením využívá pouze bigramový model. Největší vliv této ztráty informace je pak logicky u nejkratších testovacích řetězců, kde se pro identifikaci jazyka používá jen pár znaků, tedy poměrně málo informací.

Co se týká srovnání backoff a interpolovaného modelu, dávají oba modely stejné výsledky pro všechny délky testovacích řetězců.

Následující graf znázorňuje výsledky trigramových modelů.



Graf 6.2 Porovnání typů modelů na trigramových modelech

Podobně jako u bigramových modelů, jsou výsledky modelu s rovnoměrným rozložením oproti backoff a interpolovanému modelu horší, ale rozdíl v průměrné úspěšnosti identifikace jazyka mezi těmito modely vzrostl. Opět je zřejmé, že při zvyšování délky testovacích řetězců se snižuje rozdíl, a tak největší rozdíl je vidět pro délku testovacích řetězců 5 znaků, který činí přibližně 6%. Od délky 100 znaků pak mezi výsledky modelů není žádný rozdíl.

Při srovnání backoff a interpolovaného modelu dávají oba modely podobné výsledky, interpolovaný model je nepatrně lepší, ale rozdíl je řádově v setinách procenta, kde jedna setina procenta odpovídá jednomu testovacímu řetězci, který byl identifikován správně.

6.2.2 Zhodnocení

Nejlepším typem modelu je interpolovaný model, který vždy dává lepší nebo minimálně stejný výsledek jako model backoff. Rozdíl mezi nimi je však zanedbatelný, v řádech jednotek setin procenta úspěšnosti identifikace jazyka.

Větší rozdíl je pak mezi modelem backoff a modelem s rovnoměrným rozložením a to hlavně při identifikaci jazyka na krátkých testovacích řetězcích a vyšším stupni modelu. Ačkoliv jsem provedl test maximálně na trigramovém modelu, předpokládám, že pro modely s vyšším stupněm bude rozdíl v úspěšnosti identifikace jazyka ještě větší, protože s vyšším stupněm n-gramů se v testovacích řetězcích bude objevovat více neviděných n-gramů, které zapříčiní, že informace získaná z nižších stupňů n-gramu bude mít větší vliv na výsledek identifikace jazyka, proto i rozdíl mezi rovnoměrným a backoff modelem vzroste. V případě, že by se v testovacích řetězcích neobjevovaly neviděné n-gramy, byly by výsledky obou modelů stejné, to však prakticky nemůže nastat, protože trénovací korpus je omezený.

6.3 Porovnání modelů podle vyhlazovací techniky

V této kapitole budu porovnávat modely s různým stupněm a s různou vyhlazovací technikou. V následující tabulce jsou uvedené vyhlazovací techniky, které jsem použil při testování a jejich zkratky, které používám v legendách grafů.

Tabulka 6.1 Vyhlazovací techniky a jejich zkratky v legendách

Vyhlazovací technika	Zkratka
Add-One smoothing	add1
Kneser-Ney discounting (Chen and Goodman's)	kn
Kneser-Ney discounting (Original)	ukn
Good-Turing discounting	gt
Ristad's natural discounting	n
Witten-Bell discounting	wb

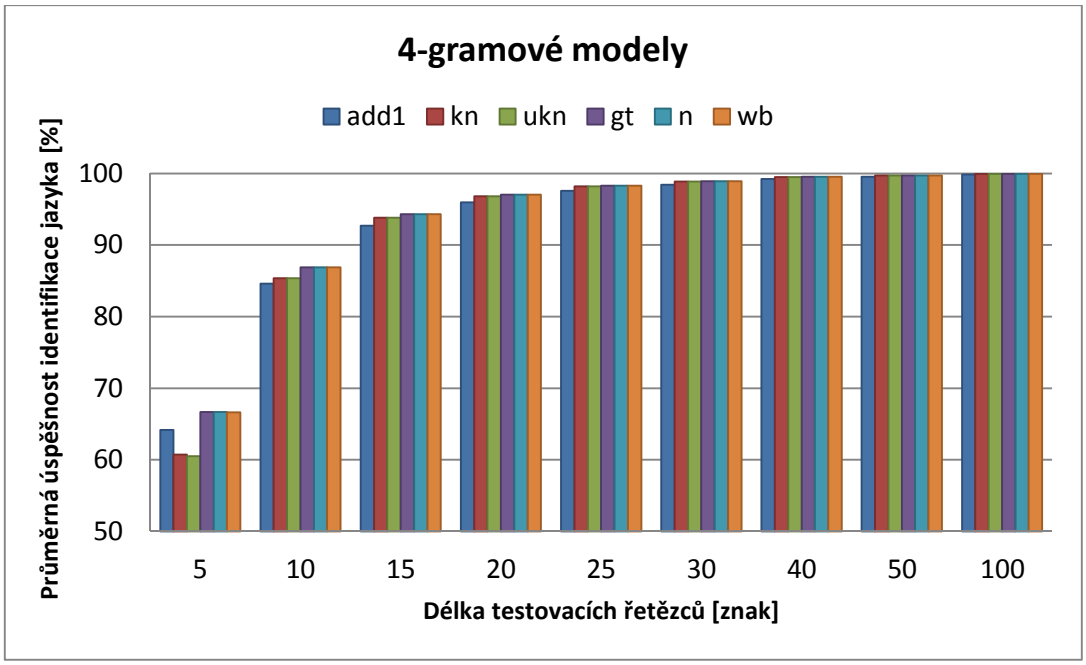
Vzhledem k tomu, že některé vyhlazovací techniky používají pro výpočet pravděpodobností počty n-gramů s určitou četností, které jsou v unigramovém modelu obvykle nulové nebo alespoň některé z nich, musel jsem pro tyto vyhlazovací techniky použít jinou vyhlazovací techniku pro unigramový model, pro ostatní vyšší stupně zůstává daná vyhlazovací technika. Jako alternativní vyhlazovací techniku jsem zvolil Witten-Bell discounting. Tento problém se týká vyhlazovací techniky Good-Turing discounting, kde jsou počty n-gramů s určitou četností přímo ve vztahu pro výpočet pravděpodobnosti, a vyhlazovací techniky Kneser-Ney discounting, kde se počty n-gramů s určitou četností využívají pro výpočet konstanty pro odečítání, viz kapitola Kneser-Ney discounting. U ostatních vyhlazovacích technik tento problém nenastává.

U vyhlazovacích technik, pro které má SRILM implementovaný interpolovaný model, jsem ho použil, protože by měl dávat stejné nebo lepší výsledky než model backoff, u ostatních jsem použil backoff model. Interpolovaný model jsem tak použil pro vyhlazovací techniku Witten-Bell discounting a Kneser-Ney discounting.

6.3.1 Porovnání výsledků

V této části uvedu grafy průměrné úspěšnosti identifikace jazyka v závislosti na délce testovacích řetězců a při daném stupni modelů. Protože grafy mají podobný průběh pro všechny stupně modelů, uvedu zde jen některé z nich, většinu grafů pak uvedu v příloze A.

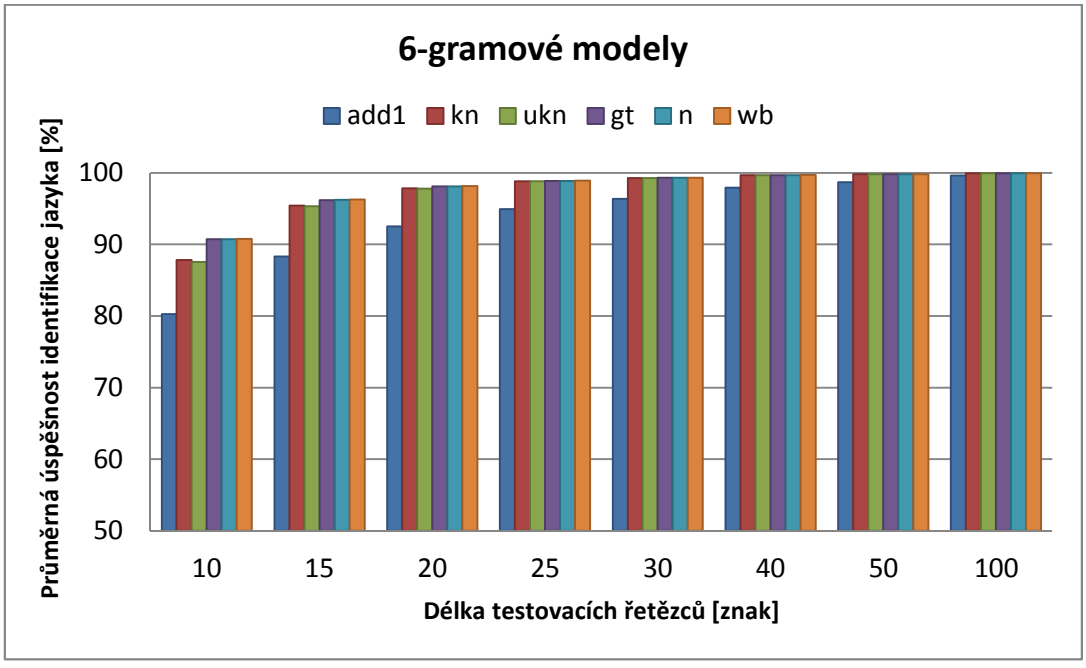
Unigramové a bigramové modely mají pro všechny vyhlazovací techniky téměř stejné výsledky, rozdíly jsou maximálně v setinách procenta úspěšnosti identifikace jazyka, proto jejich grafy zde neuvedu. U trigramového modelu se pomalu začínají projevovat rozdíly mezi vyhlazovacími technikami, ale jen v malé míře, proto jako první uvedu graf 4-gramových modelů.



Graf 6.3 Porovnání vyhlazovacích technik na 4-gramových modelech

Největší rozdíly mezi vyhlazovacími technikami se vyskytují u nejkratších testovacích řetězců délky 5 znaků. Při této délce nejvíce ztrácí vyhlazovací techniky Kneser-Ney, následuje pak Add-One a nejlépe identifikují jazyk techniky Good-Turing, Ristad Natural a Witten-Bell. S rostoucí délkou testovacích řetězců se pak rozdíly snižují a od délky 30 znaků jsou výsledky téměř totožné, jen Add-One trochu ztrácí.

Následující graf znázorňuje výsledky 6-gramových modelů.

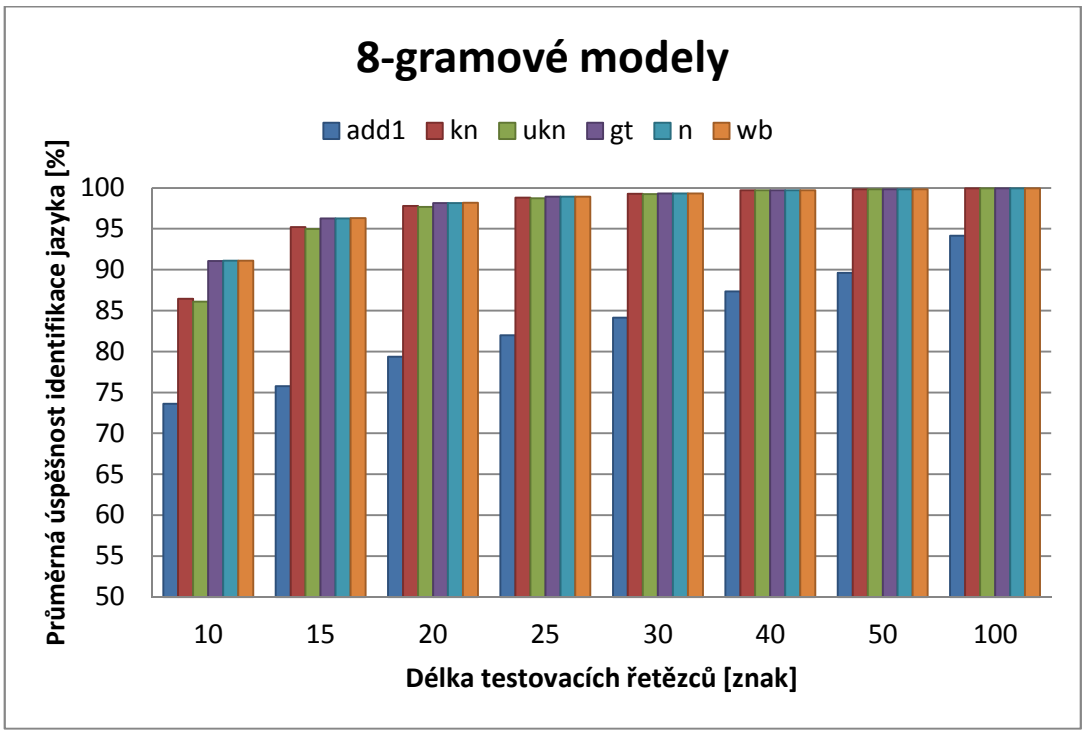


Graf 6.4 Porovnání vyhlazovacích technik na 6-gramových modelech

Podobně jako u 4-gramových modelů je největší rozdíl u nejkratších délek testovacích řetězců, s rostoucí délkou se pak tento rozdíl snižuje.

Nejvíce ztrácí vyhlazovací technika Add-One, následuje pak Kneser-Ney a zbylé vyhlazovací techniky Good-Turing, Ristad Natural a Witten-Bell opět identifikují jazyk nejlépe a mají přibližně stejné výsledky pro všechny délky testovacích řetězců.

Jako poslední uvedu graf nejvyššího stupně n-gramových modelů, který jsem testoval a to 8-gramových modelů.



Graf 6.5 Porovnání vyhlazovacích technik na 8-gramových modelech

Podobný průběh jako u předchozích výsledků jsem dostal i pro 8-gramové modely, jen rozdíly mezi vyhlazovacími technikami vzrostly.

Z grafu je patrné, že nejvíce ztrácela opět technika Add-One a to už poměrně hodně, kdy u délky testovacích řetězců 10 znaků byl rozdíl proti nejlepším technikám přes 17% úspěšnosti identifikace jazyka, a ani s rostoucí délkou se to příliš nezlepšilo.

Technika Kneser-Ney si vedla mnohem líp, ztrácela na nejlepší jen u nejkratších testovacích řetězců a to maximálně 5% a s rostoucí délkou se téměř vyrovnala těm nejlepším.

Nejlepších výsledků pak dosáhly opět techniky Good-Turing, Ristad Natural a Witten-Bell, které měly téměř shodné výsledky identifikace jazyka pro všechny délky testovacích řetězců.

6.3.2 Zhodnocení

Z předchozích grafů je patrné, že s rostoucím stupněm modelů rostou i rozdíly mezi vyhlazovacími technikami. U unigramových a bigramových modelů nejsou rozdíly mezi vyhlazovacími technikami téměř žádné, od trigramového modelu už se rozdíly pomalu ukazují a s rostoucím stupněm n-gramů pak tyto rozdíly ještě rostou. Nejvíce jsou pak patrné u modelů s nejvyšším stupněm, tedy u 8-gramovým modelů. Z tohoto tedy plyne, že pokud chceme pro identifikaci jazyka používat modely s nižším stupněm, tak tolik nezáleží na zvolené vyhlazovací technice, ale v případě, že chceme používat modely s vyšším stupněm, tak zvolená vyhlazovací technika už bude mít značný vliv na úspěšnost identifikace jazyka.

Za nejlepší vyhlazovací techniky podle výsledků lze považovat Witten-Bell, Good-Turing a Ristad Natural discounting. Jejich výsledky jsou téměř shodné pro všechny stupně modelů a všechny délky testovacích řetězců, liší se maximálně v setinách procenta úspěšnosti identifikace jazyka, ale většinou jsou jejich výsledky totožné. Z tohoto důvodu nelze určit jednoznačně vítěze, ale lze doporučit tyto tři vyhlazovací techniky pro identifikaci jazyka.

Na další místo bych pak umístil vyhlazovací techniku Kneser-Ney, která pro delší testovací řetězce dávala podobné výsledky jako tři nejlepší, ale rozdíl byl hlavně u kratších testovacích řetězců okolo 5 až 10 znaků, kde tato technika ztrácela poměrně hodně. Co se týká srovnání této metody při použití původní verze (ukn), která počítá pouze s jednou konstantou pro odečítání, a modifikovanou verzí Chen and Goodman (kn), která počítá se třemi konstantami pro odečítání, tak výsledky ukazují, že modifikovaná verze většinou dávala trochu lepší výsledky než původní, ale rozdíl nebyl příliš velký.

Za nejhorší vyhlazovací techniku pak lze určit Add-One a to hlavně pro modely s vyšším stupněm, kde tato technika ztrácela na všech délkách testovacích řetězců i proti technice Kneser-Ney discounting. Nejvíce je to pak vidět na výsledcích 8-gramových modelů.

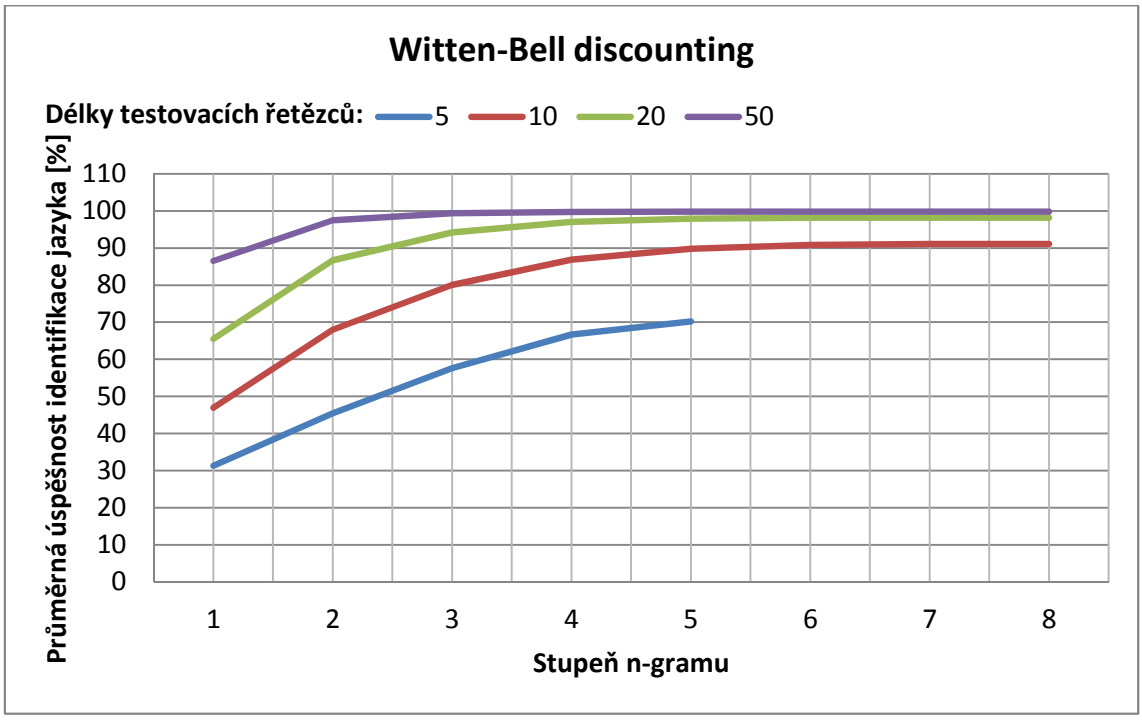
Jako výchozí vyhlazovací techniku pro další testování v této kapitole jsem zvolil techniku Witten-Bell discounting pro její dobré výsledky, jednoduchost a rozšířenost.

6.4 Porovnání modelů podle jejich stupně

V této kapitole budu porovnávat n-gramové modely podle jejich stupně. Jako vyhlazovací techniku jsem použil Witten-Bell discounting. V první části použiji průměrnou úspěšnost všech 34 jazyků, které mám k dispozici, v druhé části pak zobrazím výsledky identifikace jazyka při použití pouze dvou jazyků a to češtiny a slovenštiny.

6.4.1 Porovnání výsledků identifikace všech jazyků

Následující graf zobrazuje průměrnou úspěšnost identifikace jazyka v závislosti na stupni n-gramového modelu. Pro přehlednost zobrazuji jen výsledky některých délek testovacích řetězců a to 5, 10, 20 a 50 znaků. Ostatní grafy uvedu v příloze B.



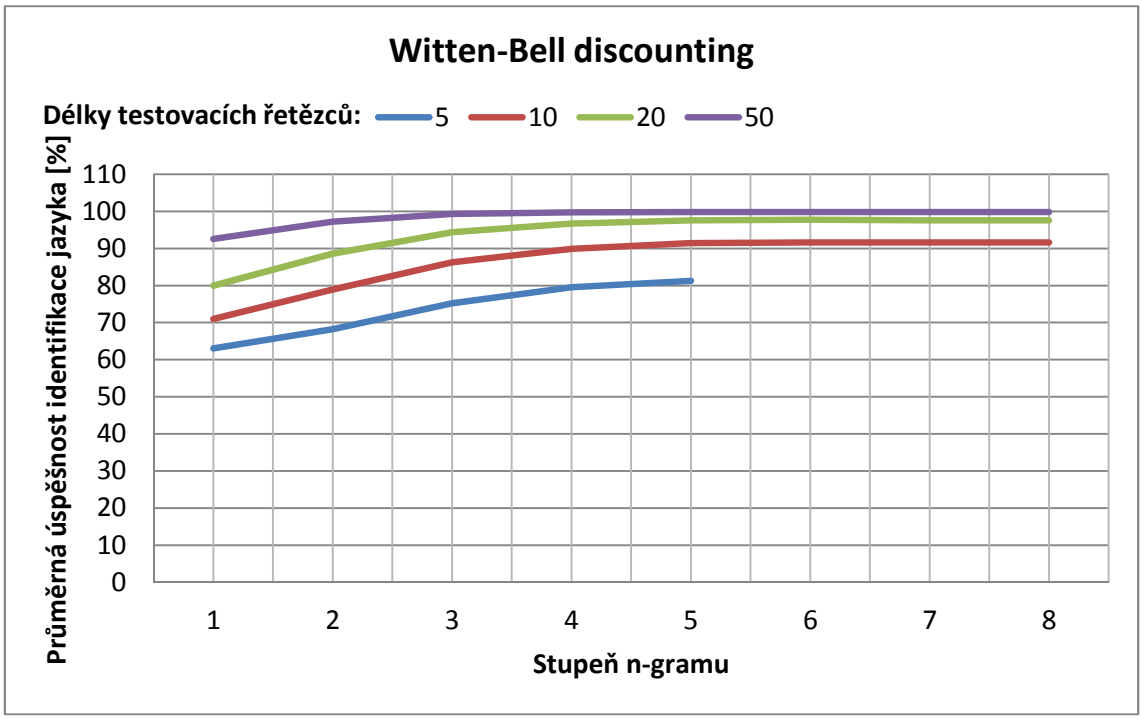
Graf 6.6 Porovnání modelů podle jejich stupně

Z předchozího grafu je patrné, že stupeň n-gramového modelu v rozmezí 1 až 5 má značný vliv na výsledky identifikace jazyka. Od stupně 6 a výše jsou pak výsledky identifikace jazyka téměř shodné a liší se řádově v jednotkách setin procenta.

6.4.2 Porovnání výsledků identifikace češtiny a slovenštiny

Výsledky identifikace těchto dvou jazyků zde zobrazuji kvůli aplikaci pro třídění textů podle jazyka, kterou je popsaná v kapitole Vytvořené aplikace, kde byl záměr odfiltrovat právě české a slovenské texty. Zároveň tyto výsledky ukazují, jak přesně lze identifikovat jazyk u dvou podobných jazyků.

Následující graf zobrazuje průměrnou úspěšnost identifikace češtiny a slovenštiny v závislosti na stupni n-gramového modelu. Pro přehlednost zde opět zobrazuji jen výsledky některých délek testovacích řetězců a to 5, 10, 20 a 50 znaků. Ostatní grafy uvedu v příloze B.



Graf 6.7 Porovnání modelů podle jejich stupně

Opět je zřejmé, že pro stupně n-gramového modelu v rozmezí 1 až 5 má tento parametr celkem silný vliv na výsledky identifikace jazyka a od stupně 6 a výše tento vliv prakticky mizí, rozdíly jsou pak opět v jednotkách setin procenta úspěšnosti identifikace jazyka.

6.4.3 Zhodnocení

Obecně lze říci, že použitím vyššího stupně n-gramového modelu dostaneme lepší výsledky, od určitého stupně však další zvyšování už nemá smysl, vzhledem k omezené velikosti trénovacího korpusu, kdy hodně n-gramů, které by se mohly objevit v daném jazyce, se neobjeví, a tak při vyhodnocování často dochází k použití nižšího stupně, viz. kapitoly backoff a interpolovaný model. Vyšší stupeň pak pouze zvyšuje výpočtovou náročnost. Proto lze při mé velikosti trénovacího korpusu, která činí 11 miliónů znaků, doporučit 6-gramový model, který dává téměř shodné výsledky jako vyšší stupně a přitom je proti nim méně náročný na výpočet.

6.5 Porovnání modelů podle diakritiky

V této části posuzuji vliv diakritiky na úspěšnost identifikace jazyka. Pro testování jsem omezil počet jazyků, které se budou identifikovat, na jazyky používající latinku, tak aby šlo použít texty jak s diakritikou tak bez ní. V následující tabulce je jejich seznam.

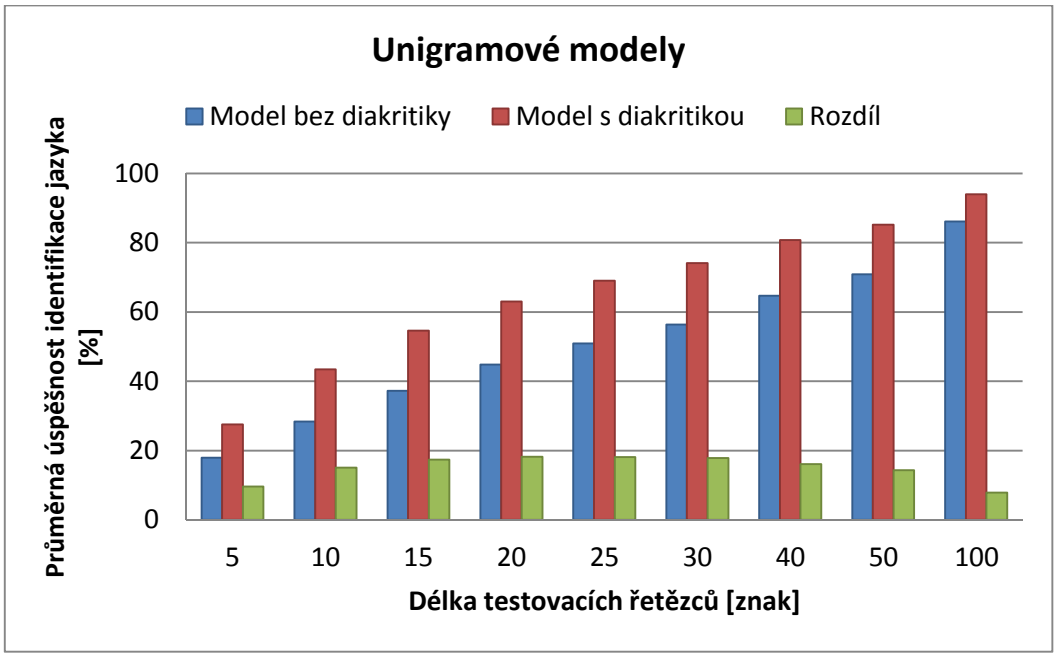
Tabulka 6.2 Jazyky používající latinku

Jazyk		Jazyk	
1	albánština	15	lotyština
2	angličtina	16	maďarština
3	baskičtina	17	němčina
4	čeština	18	norština
5	dánština	19	polština
6	estonština	20	portugalština
7	finština	21	rumunština
8	francouzština	22	slovenština
9	holandština	23	slovinština
10	chorvatština	24	španělština
11	islandština	25	švédština
12	italština	26	turečtina
13	katalánština	27	vietnamština
14	litevština		

Pro testování jsem použil interpolovaný n-gramový model s vyhlazováním Witten-Bell discounting se stupněm 1 až 5. Trénování modelů proběhlo na stejných textových datech, jen s rozdílem diakritiky. To samé platí pro testovací řetězce, které byly náhodně vygenerované z testovacích vět a pro danou délku stejné, jen s rozdílem diakritiky.

6.5.1 Porovnání výsledků

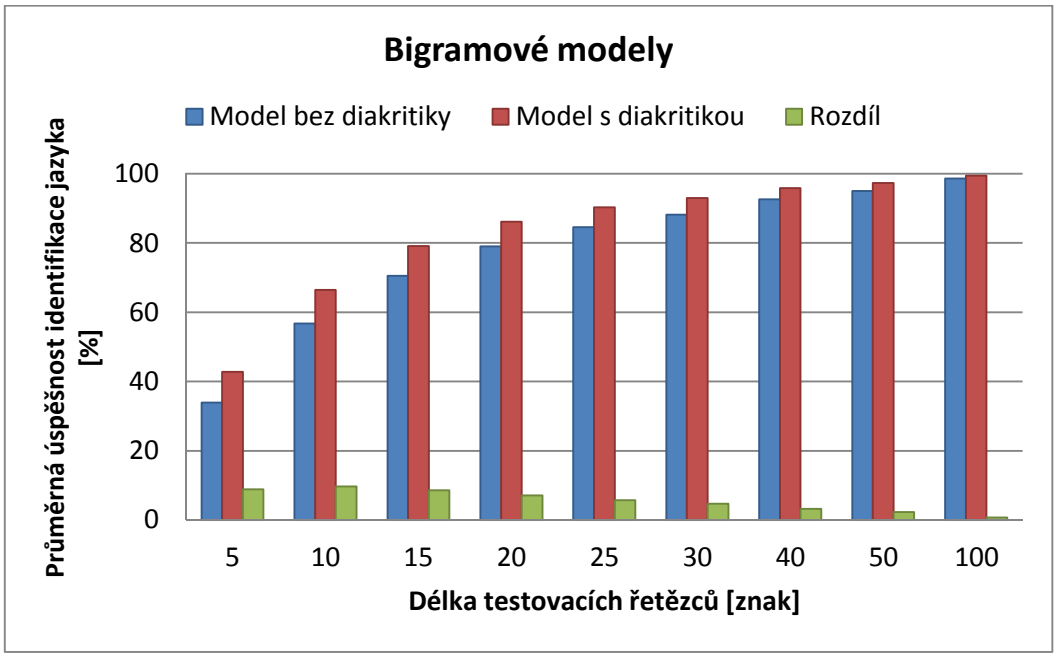
Následující graf znázorňuje výsledky unigramových modelů.



Graf 6.8 Porovnání unigramových modelů podle diakritiky

U unigramových modelů jsou rozdíly v úspěšnosti identifikace jazyka celkem velké, průměrně 15%, nejmenší rozdíl je pro délku testovacích řetězců 100 znaků, kde činí přibližně 8%, největší je pak pro délku testovacích řetězců 20 znaků a to přes 18%.

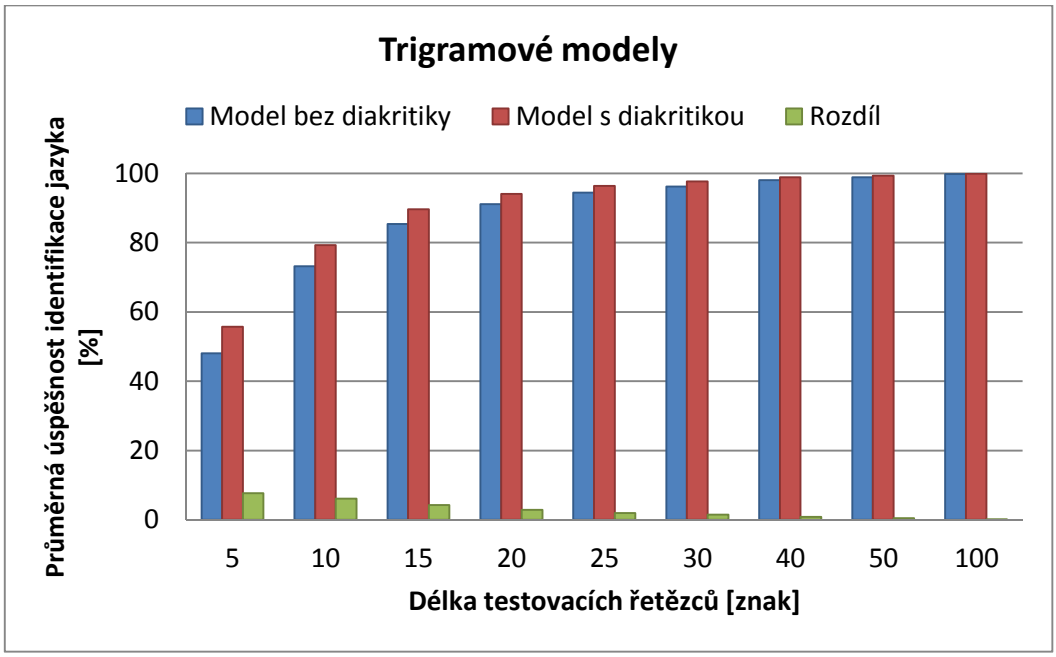
Následující graf znázorňuje výsledky bigramových modelů.



Graf 6.9 Porovnání bigramových modelů podle diakritiky

U bigramových modelů se rozdíl mezi modelem s diakritikou a bez diakritiky snížil, průměrně činil přibližně 5,5%. Největší rozdíl byl pro délku 10 znaků a to necelých 10%, nejmenší pak byl pro nejdelší testovací řetězce 100 znaků, kdy činil necelé procento průměrné úspěšnosti identifikace jazyka.

Jako poslední uvedu následující graf trigramových modelů.



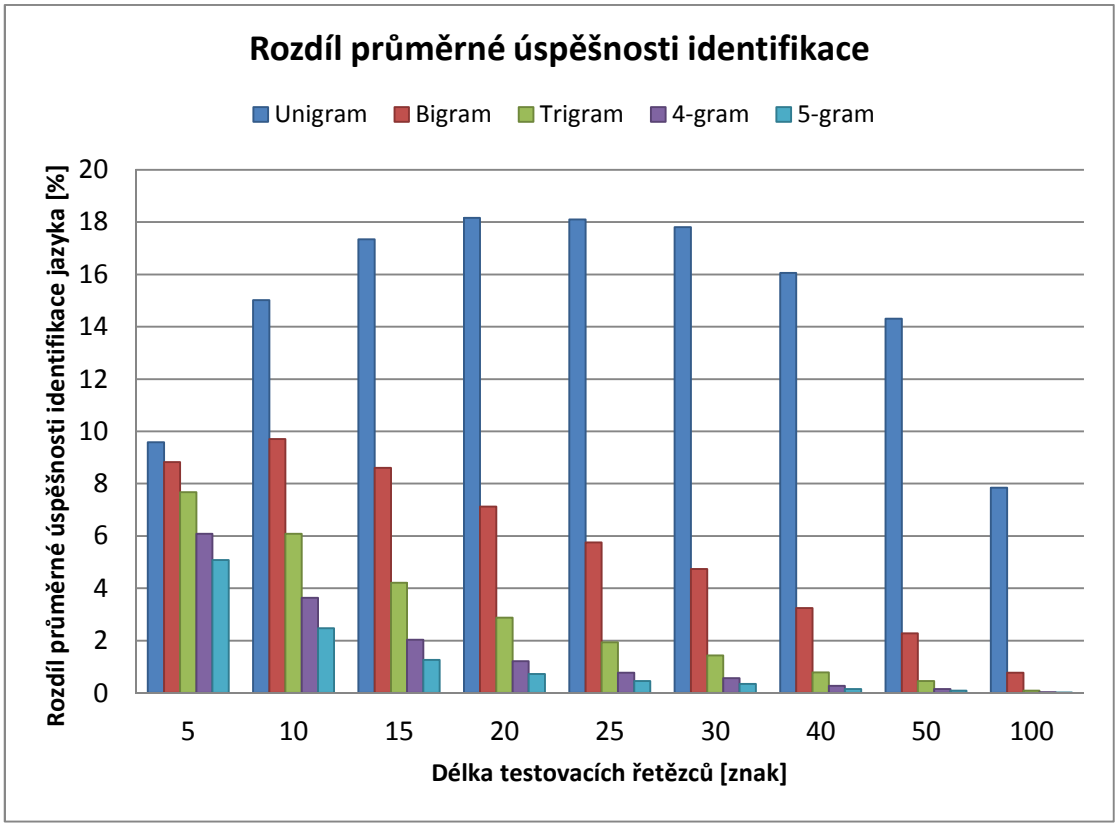
Graf 6.10 Porovnání trigramových modelů podle diakritiky

Zde už se průměrný rozdíl mezi modelem s diakritikou a bez diakritiky snížil na necelá 3%. Nejmenší rozdíl byl pro testovací řetězce délky 5 znaků a to přibližně 7,5%, nejmenší rozdíl opět pro nejdelší testovací řetězce délky 100 znaků a to 0,1%.

Průběhy vyšších stupňů n-gramových modelů měly podobný průběh jako trigramový model, jen se snížily rozdíly mezi modely, průměrný rozdíl modelů se pohyboval kolem 0,5%. Tyto grafy zde neuvedu, ale jsou v příloze C.

6.5.2 Zhodnocení

Následující graf ukazuje rozdíl průměrné úspěšnosti identifikace jazyka modelu s diakritikou a modelu bez diakritiky pro různé stupně n-gramových modelů.



Graf 6.11 Porovnání rozdílu úspěšnosti identifikace jazyka modelů s diakritikou a bez diakritiky

Používání textů bez diakritiky není nic jiného, než snížení počtu možných znaků, tedy snížení počtu znaků slovníku modelu. Odstraněním diakritiky tak ztrácíme část informace, která se nutně musí projevit na úspěšnosti identifikace jazyka. Vliv této ztráty je pak nejvíce vidět na unigramovém modelu, který vychází pouze z četností jednotlivých znaků slovníku, který jsme mu omezili.

Pro vyšší stupně modelů vliv diakritiky postupně klesá a to hlavně pro delší testovací řetězce. Pokud se podíváme na výsledky 5-gramového modelu, tak od délky testovacích řetězců 20 znaků se rozdíl v úspěšnosti mezi modelem s diakritikou a modelem bez diakritiky pohybuje kolem 0,7% a pro delší testovací řetězce ještě klesá, což už může být zanedbatelné.

Pokud tedy mám zvážit vliv diakritiky na identifikaci jazyka lze říci, že v případě, že chceme identifikovat jazyk s modelem s nižším stupněm nebo pokud chceme identifikovat jazyk na krátkých textech, je vliv diakritiky celkem velký a není ji vhodné zanedbávat. Naopak pro modely s vyšším stupněm a délky textů zhruba od 20 znaků je vliv diakritiky zanedbatelný.

6.6 Konfuzní matice

Konfuzní matice zobrazují přesnější informace o identifikaci jazyka. Ukazují nejen úspěšnost identifikace daného jazyka, ale také s jakými jazyky se nejvíce zaměňoval. Konfuzní matice je tedy čtvercová matice, kde sloupce představují rozpoznávané jazyky a řádky představují testovací řetězce v daném jazyce. Hodnoty konfuzní matice tedy říkají, v kolika procentech byl identifikován jazyk daného sloupce při identifikaci na testovacích řetězcích jazyka daného řádku.

6.6.1 Uspořádání jazyků

Vzhledem k tomu, že k záměně jazyků dochází hlavně u jazyků ze stejné skupiny jazyků, není vhodné uspořádat jazyky v matici podle abecedy, ale právě podle skupin, do kterých patří. V následující tabulce uvádím rozdělení jazyků do skupin, v těchto skupinách jsou jazyky uspořádany už podle abecedy. Jazyky ze skupiny ostatní patří každý do jiné jazykové skupiny, s tím že řečtina nepoužívá latinku, ale řecké písmo, proto je jako poslední.

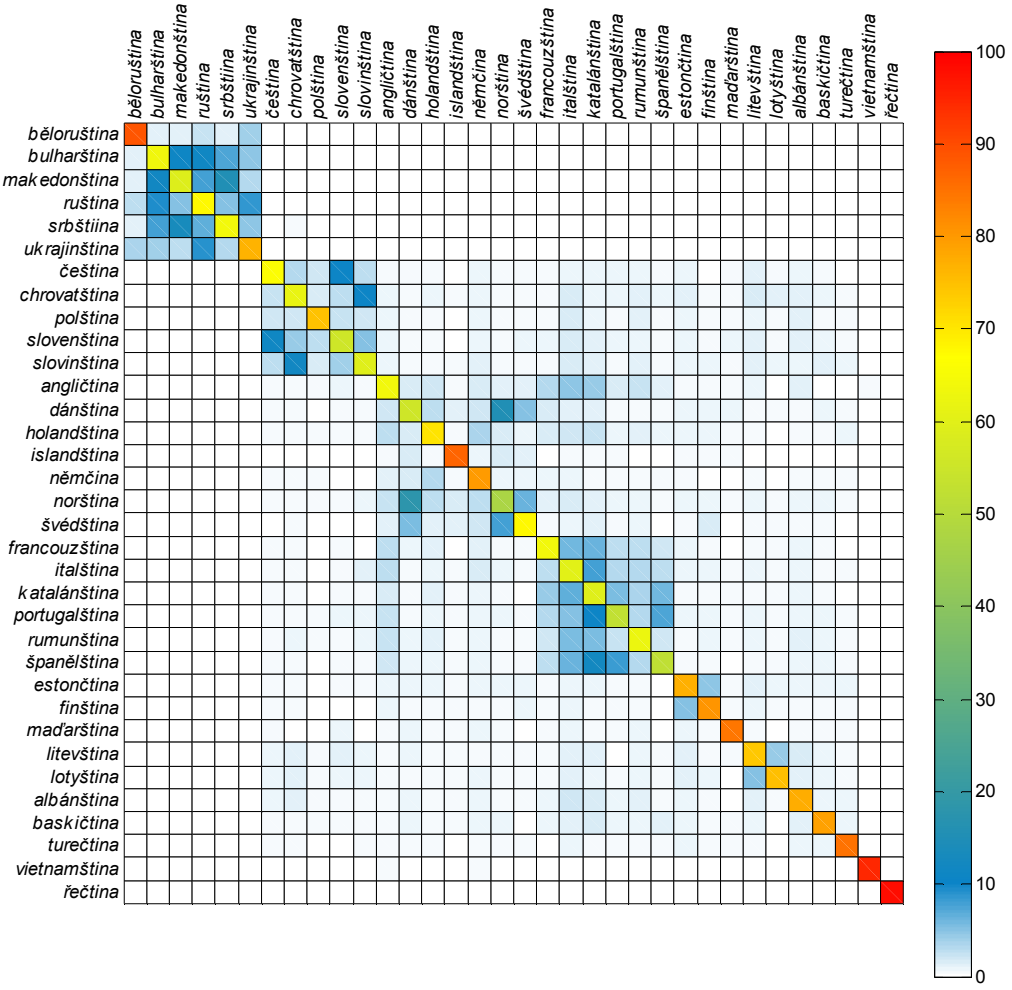
Tabulka 6.3 Rozdělení jazyků podle jazykových skupin

Slovanské *cyrilice*	Slovanské *latinka*	Germánské	Románské	Ugrofinské	Baltské	Ostatní
běloruština	čeština	angličtina	francouzština	estonština	litevština	albánština
bulharština	chorvatština	dánština	italština	finština	lotyština	baskičtina
makedonština	polština	holandština	katalánština	maďarština		turečtina
ruština	slovenština	islandština	portugalština			vietnamština
srbština	slovinština	němčina	rumunština			řečtina
ukrajinština		norština	španělština			
		švédština				

6.6.2 Vybraná konfuzní matice

V následující části zobrazím graf vybrané konfuzní matice, ostatní konfuzní matice včetně jejich tabulek s daty pak uvedu v příloze D.

Následující graf zobrazuje konfuzní matici 6-gramového modelu s vyhlazováním Wittem-Bell discounting při identifikaci jazyka na testovacích řetězcích délky 5 znaků.



Graf 6.12 Konfuzní matice 6-gramového modelu Witten-Bell discounting s délkou testovacích řetězců 5 znaků

Z grafu lze vyčíst, s kterým jazykem se každý jazyk nejvíce zaměňoval, např. čeština se nejvíce zaměňovala se slovenštinou a to přibližně v 10 % případů, dále se zaměňovala s chorvatštinou, polštinou a slovinštinou. Obdobně lze podobné informace vyčíst i pro ostatní jazyky. Z grafu lze vyčíst pouze přibližné hodnoty úspěšnosti identifikace daného jazyka, přesné hodnoty lze vyčíst přímo z tabulky matice, kterou jsem zde kvůli úspoře místa, nezobrazil. Tuto i ostatní tabulky konfuzních matic včetně jejich grafické podoby jsem umístil do přílohy D.

7 Vytvořené aplikace

V této části popíši aplikace vytvořené v rámci této diplomové práce. Všechny tyto aplikace jsou konsolové aplikace vytvořené v programovacím jazyku C# a jsou založené na projektu SRILM. V následující tabulce je seznam těchto aplikací včetně stručného popisu. Manuály k těmto aplikacím jsou umístěny v příloze E.

Tabulka 7.1 Seznam vytvořených aplikací

Aplikace	Popis
Model Creator	Aplikace pro vytváření jazykových modelů.
Language Recognizer	Aplikace pro třídění textů podle jazyka.
Language Recognizer View	Aplikace pro zobrazení výsledků identifikace jazyka.

7.1 Model Creator

Tato aplikace slouží k vytvoření modelu z trénovacího korpusu. V podstatě přeposílá požadavky do aplikace ngram-count, která slouží k vytváření modelů v rámci projektu SRILM. Znaky trénovacího korpusu rozdělí mezerami, tak jak to vyžaduje projekt SRILM, aby pracoval se znaky, protože SRILM pracuje s tím, co je odděleno mezerami, obvykle se slovy. Původní mezery nahradí za podtržítko „_“, proto by se v trénovacím korpusu nemělo podtržítko objevovat.

7.2 Language Recognizer

Tato aplikace slouží pro třídění textů podle jazyka. Aplikaci v parametru předáte soubor s texty, který chcete roztřídit, a řeknete, které modely má použít při identifikaci jazyka. Dále aplikaci předáte seznam oddělovačů, podle kterých má rozdělovat text na části, na kterých se bude jazyk identifikovat. Implicitně se oddělují pouze odstavce. Aplikaci lze předat ještě mnohem více parametrů, které ovlivňují výsledky identifikace jazyka jako např. stupeň n-gramového modelu, který se má použít, kódování textu, minimální délka vět, kdy při nedostatečné délce se ignoruje následující oddělovač a další. Všechny tyto parametry jsou uvedeny v manuálu této aplikace, v příloze E. Výsledkem jsou roztříděné texty umístěné v jednotlivých textových souborech a označené identifikovaným modelem.

7.3 Language Recognizer View

Tato aplikace slouží k zobrazení výsledků identifikace jazyka zvoleného textu. Na konsoly, případně do souboru vypíše věty a hodnoty jejich pravděpodobností v jednotlivých modelech, hodnotu s největší pravděpodobností barevně zvýrazní (identifikovaný model jazyka).

Závěr

Práce byla zaměřena na identifikaci jazyka textového dokumentu pomocí n-gramových modelů. Seznámil jsem se tak s problematikou n-gramových modelů, kterou jsem popsal v teoretické části této práce.

Jedním z cílů práce bylo posoudit vliv jednotlivých parametrů n-gramových modelů na úspěšnost identifikace jazyka textového dokumentu. Tuto část jsem řešil v kapitole 6, kam jsem umístil výsledky porovnání různých n-gramových modelů. Modely jsem porovnal podle vyhlazovací techniky, stupně a typu modelu. Dále jsem v této kapitole posoudil vliv diakritiky na úspěšnost identifikace jazyka. V každé části této kapitoly jsem umístil nejdůležitější výsledky provedených testů, které jsem poté zhodnotil. Zbylé výsledky a hlavně konfuzní matice, které přináší detailní informace o úspěšnosti identifikace jednotlivých jazyků, jsem z důvodu rozsáhlosti umístil do příloh.

Z těchto výsledků lze pro identifikaci jazyka textového dokumentu doporučit 6-gramový interpolovaný model s vyhlazovací technikou Witten-Bell, případně Good-Turing nebo Ristad's natural discounting. Tyto tři vyhlazovací techniky měly v podstatě shodné výsledky. Použití vyššího stupně n-gramových modelů nepřineslo, při mé velikosti trénovacích korpusů, žádné výrazné zlepšení výsledků a jen zvýšilo výpočtovou náročnost.

Výsledkem práce je také několik aplikací, založených na projektu SRILM. Aplikace „Model Creator“ pro vytváření modelů, která upravuje vstupní text tak, aby s ním projekt SRILM pracoval po znacích. Aplikace „Language Recognizer View“ pro zobrazení výsledků identifikace jazyka na konsoly a aplikace „Language Recognizer“ pro třídění textů podle jazyka, která slouží hlavně pro filtrování textových dat.

Seznam použité literatury

[1] Jurafsky D., Martin J. H. *Speech and Language Processing.* Prentice Hall, A Simon and Schuster Company. New Jersey 2000.

[2] Manning Ch., Schütze H. *Foundations of Statistical Natural Language Processing.* The MIT Press Cambridge. London 1999.

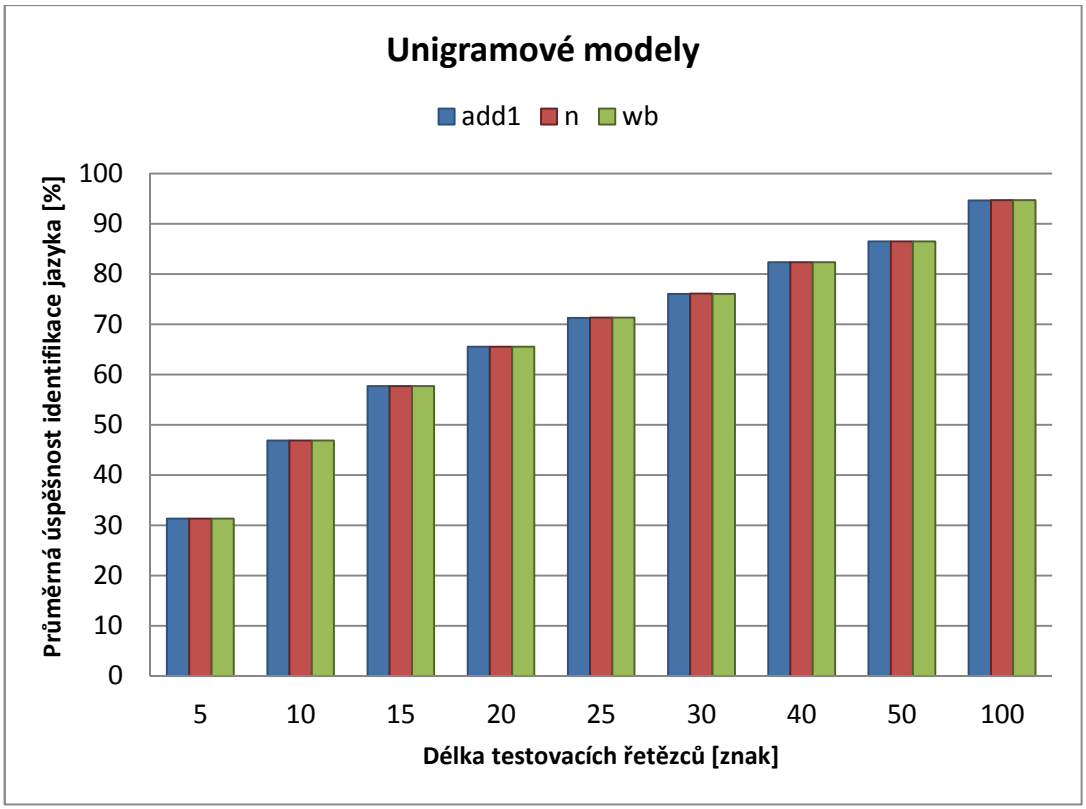
[3] Ristad E. S. *A Natural Law of Succesion.* http://arxiv.org/abs/cmp-lg/9508012v1 [online] [cit. 2012-05-05]. New Jersey 1995

[4] Frankie James. *Modified Kneser-Ney, Smoothing of n-gram Models.* 2000

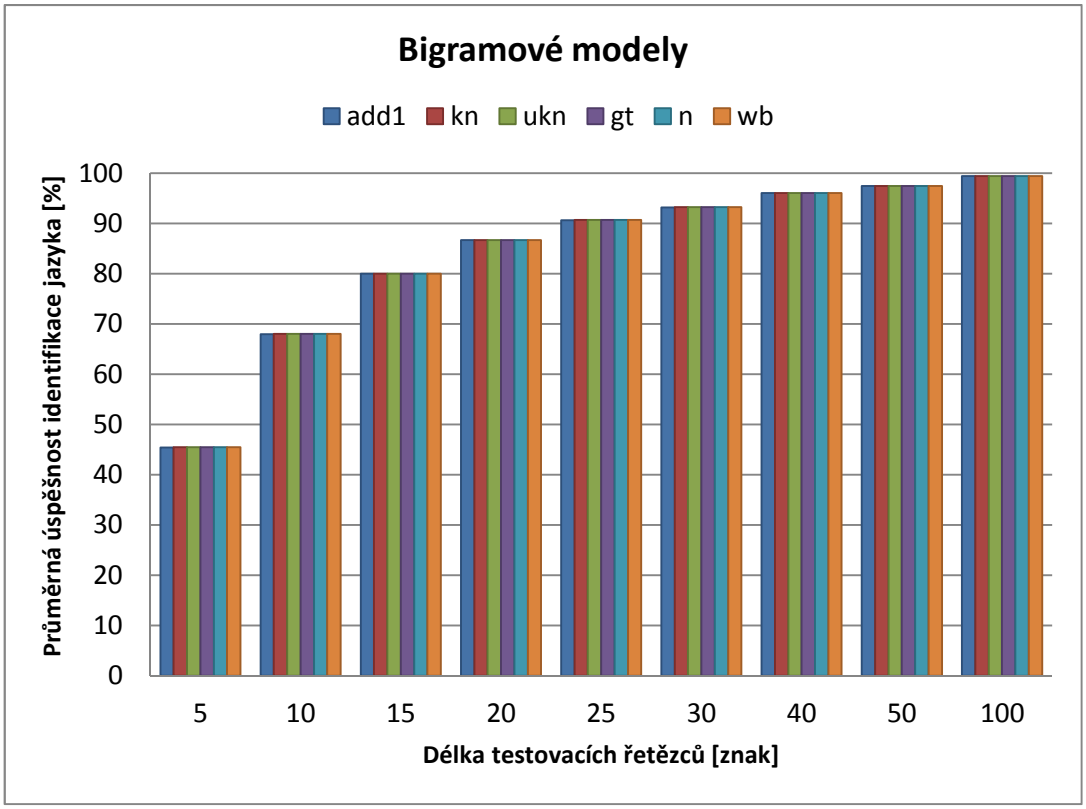
[5] Stolcke A. *http://www.speech.sri.com/projects/srilm/* [online] [cit. 2012-05-05]

[6] Ager S. *http://www.omniglot.com/* [online] [cit. 2012-05-05]

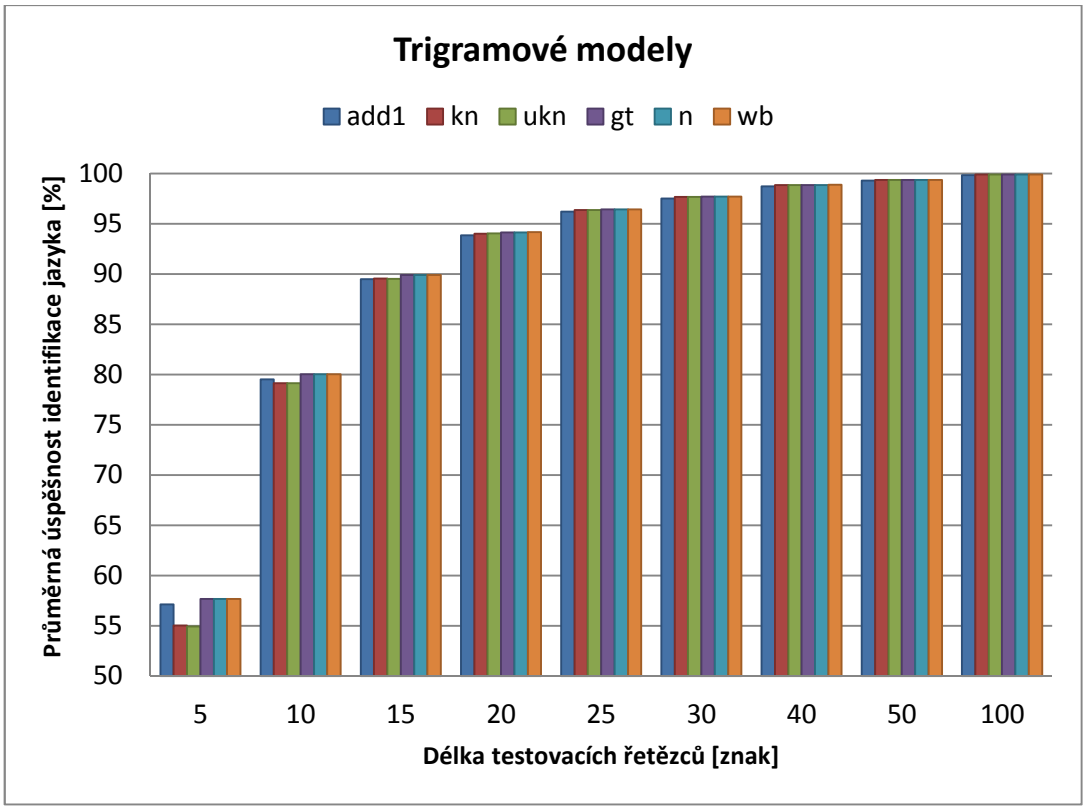
Příloha A
Porovnání modelů podle vyhlazovací techniky



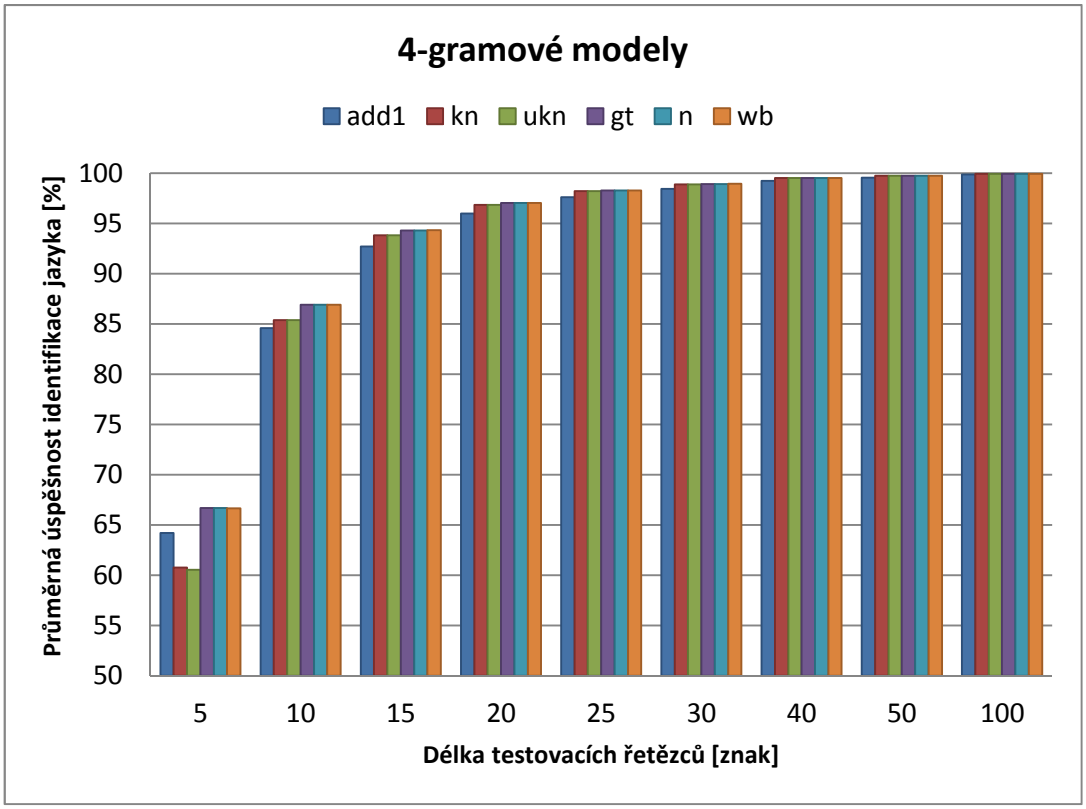
Graf A.1 Porovnání vyhlazovacích technik na unigramových modelech



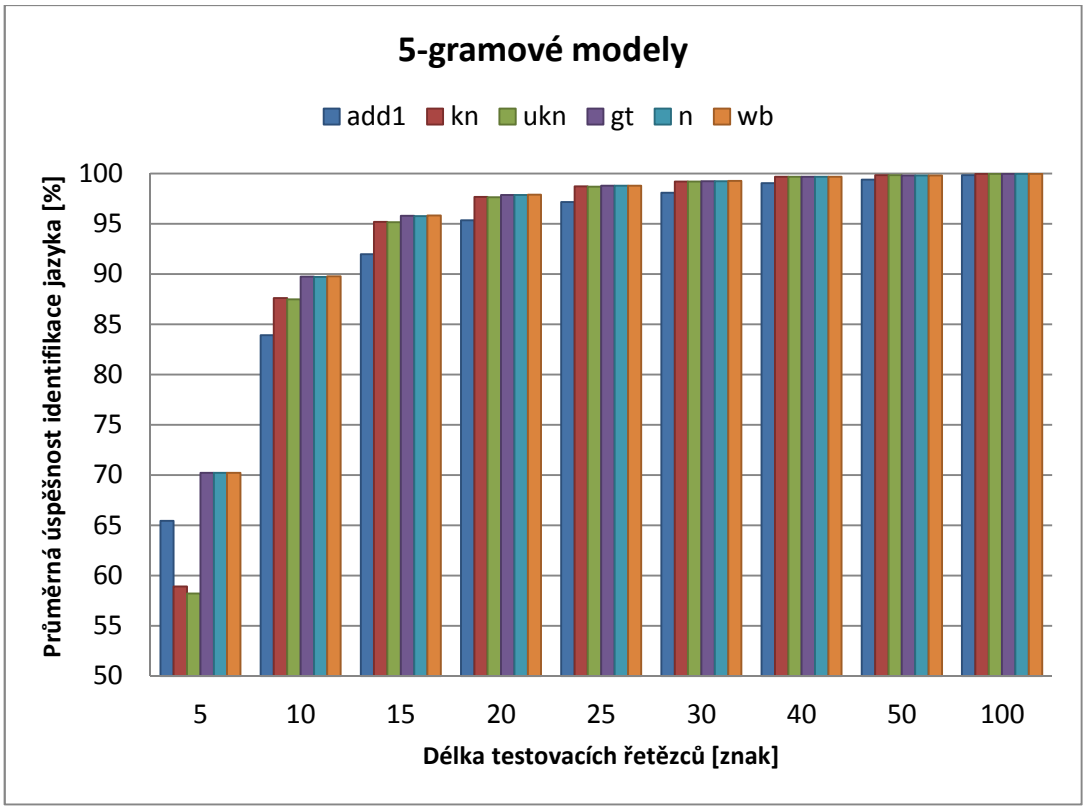
Graf A.2 Porovnání vyhlazovacích technik na bigramových modelech



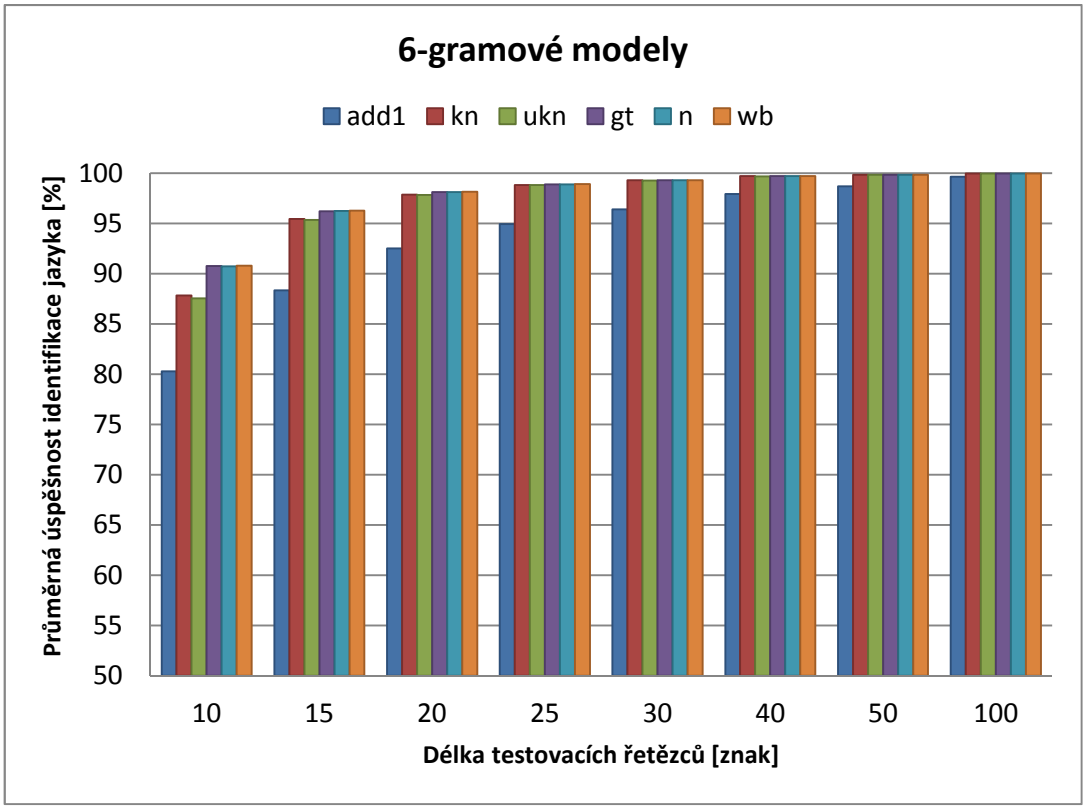
Graf A.3 Porovnání vyhlazovacích technik na trigramových modelech



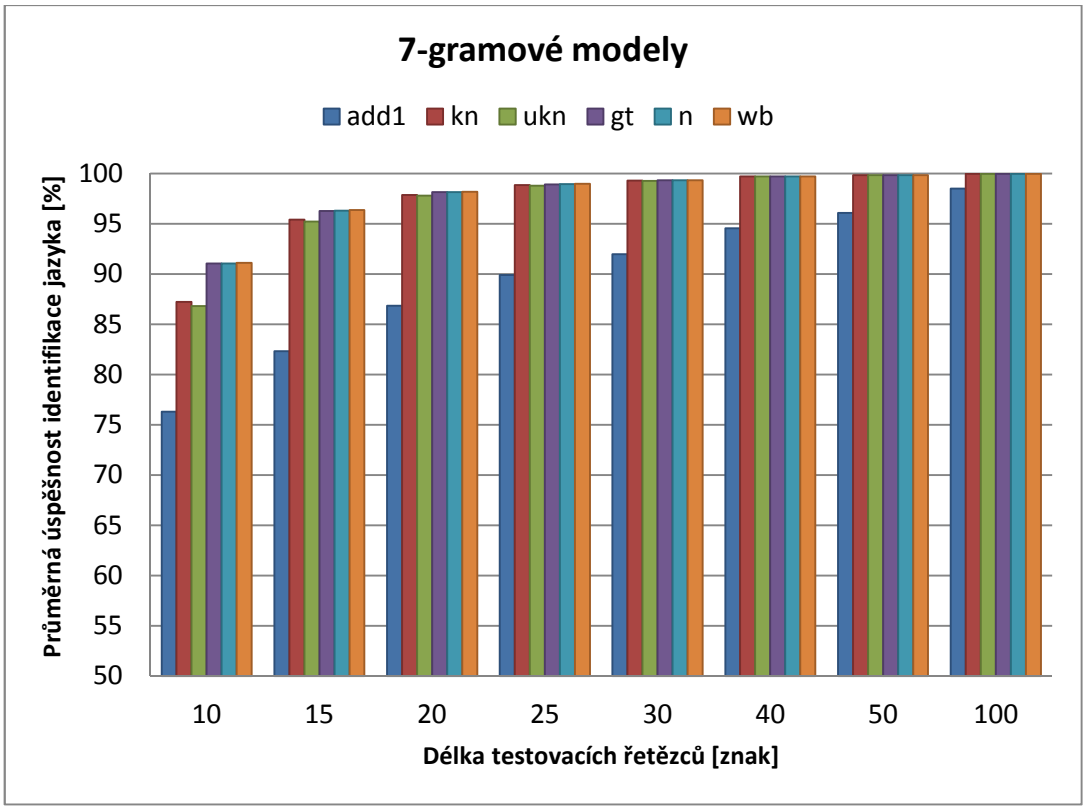
Graf A.4 Porovnání vyhlazovacích technik na 4-gramových modelech



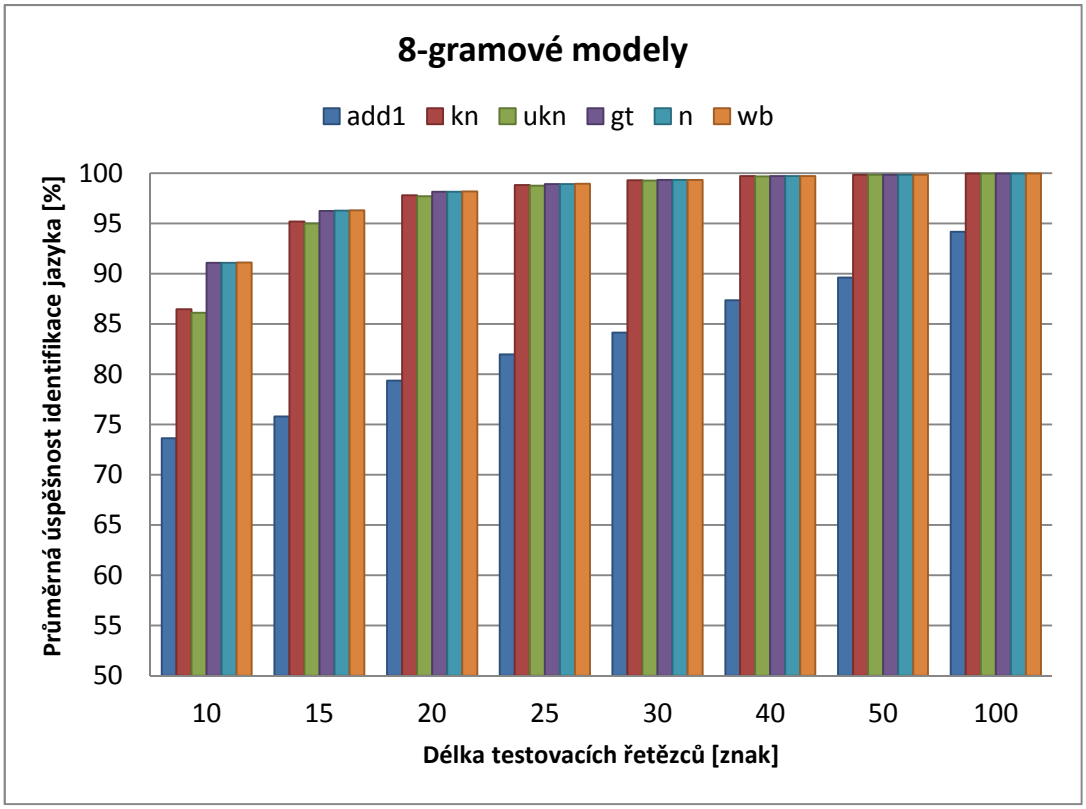
Graf A.5 Porovnání vyhlazovacích technik na 5-gramových modelech



Graf A.6 Porovnání vyhlazovacích technik na 6-gramových modelech

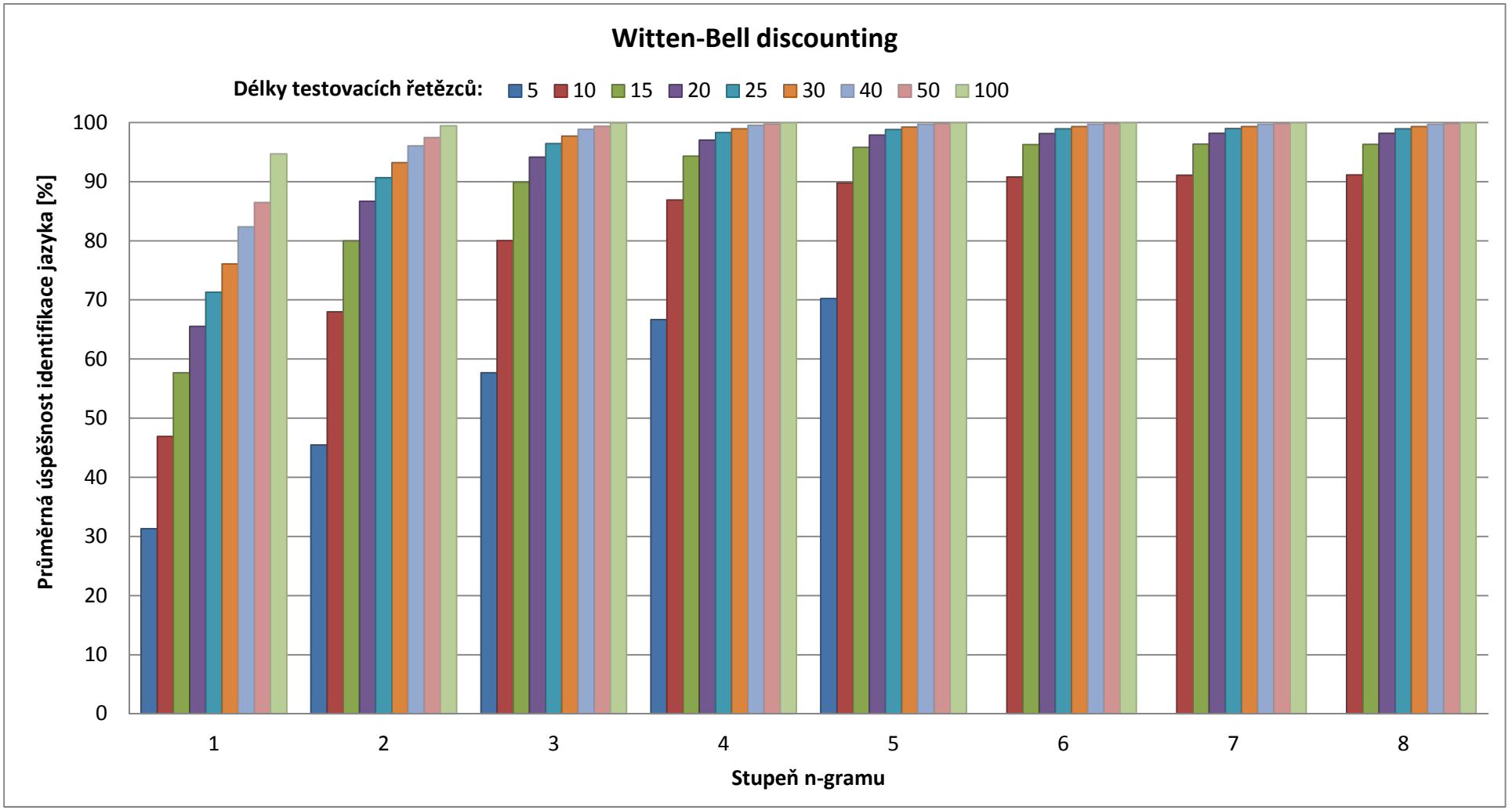


Graf A.7 Porovnání vyhlazovacích technik na 7-gramových modelech

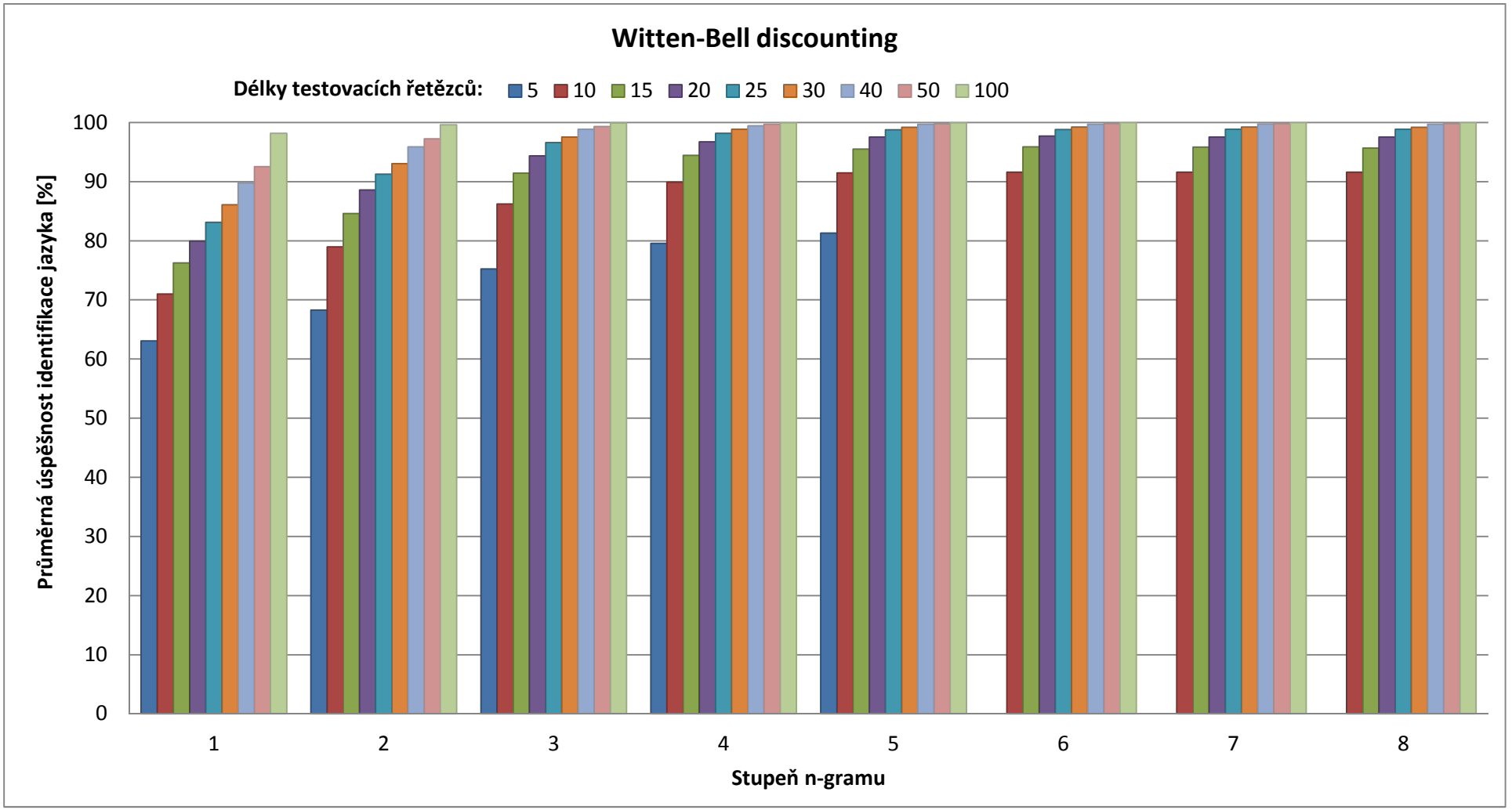


Graf A.8 Porovnání vyhlazovacích technik na 8-gramových modelech

Příloha B
Porovnání modelů podle stupně

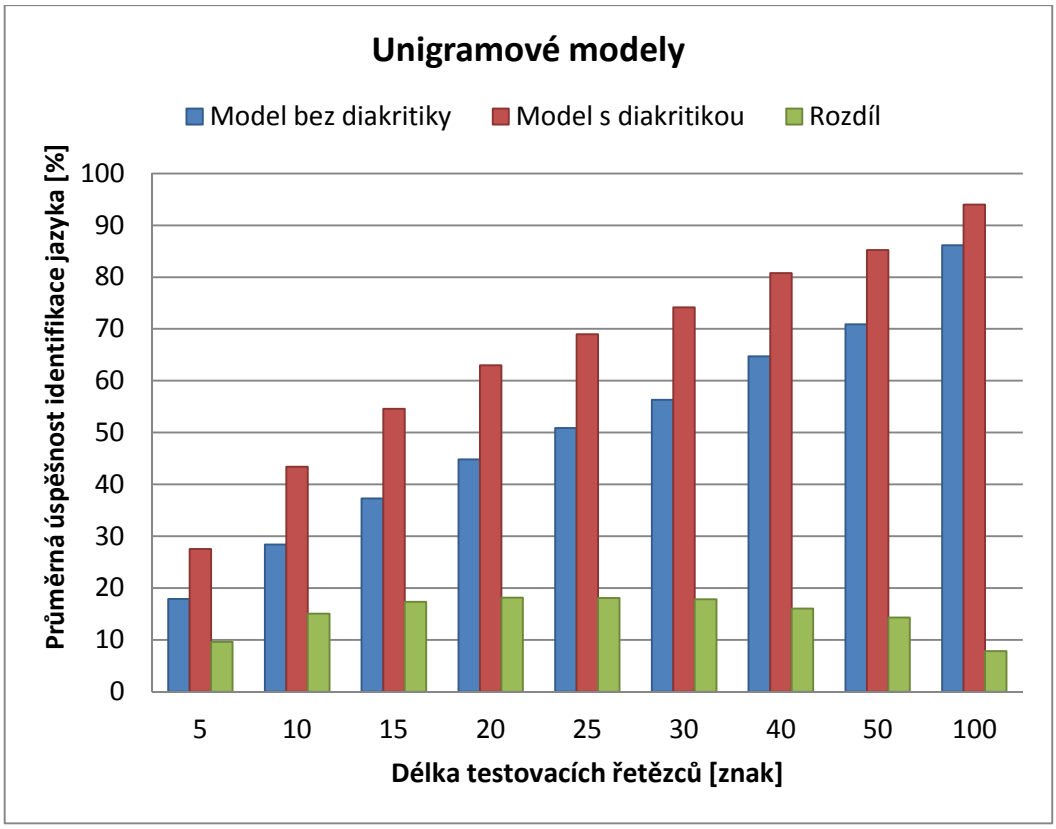


Graf B.1 Porovnání modelů podle stupně při identifikaci všech jazyků

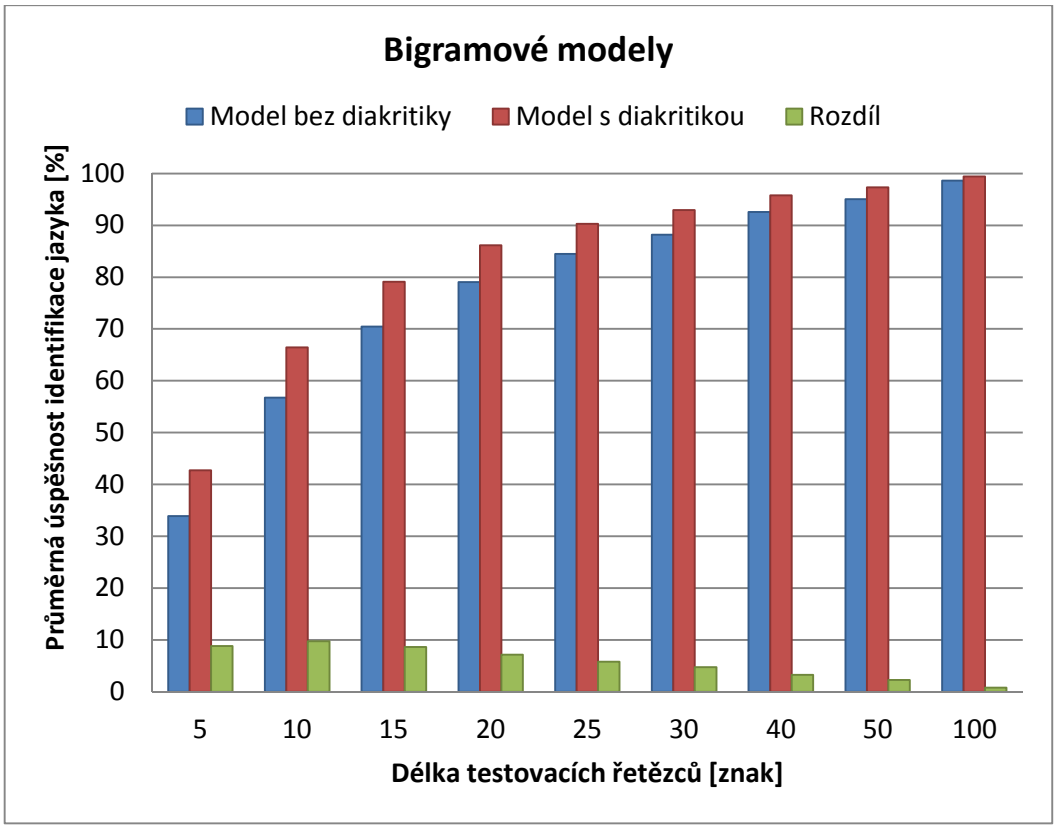


Graf B.2 Porovnání modelů podle stupně při identifikaci češtiny a slovenštiny

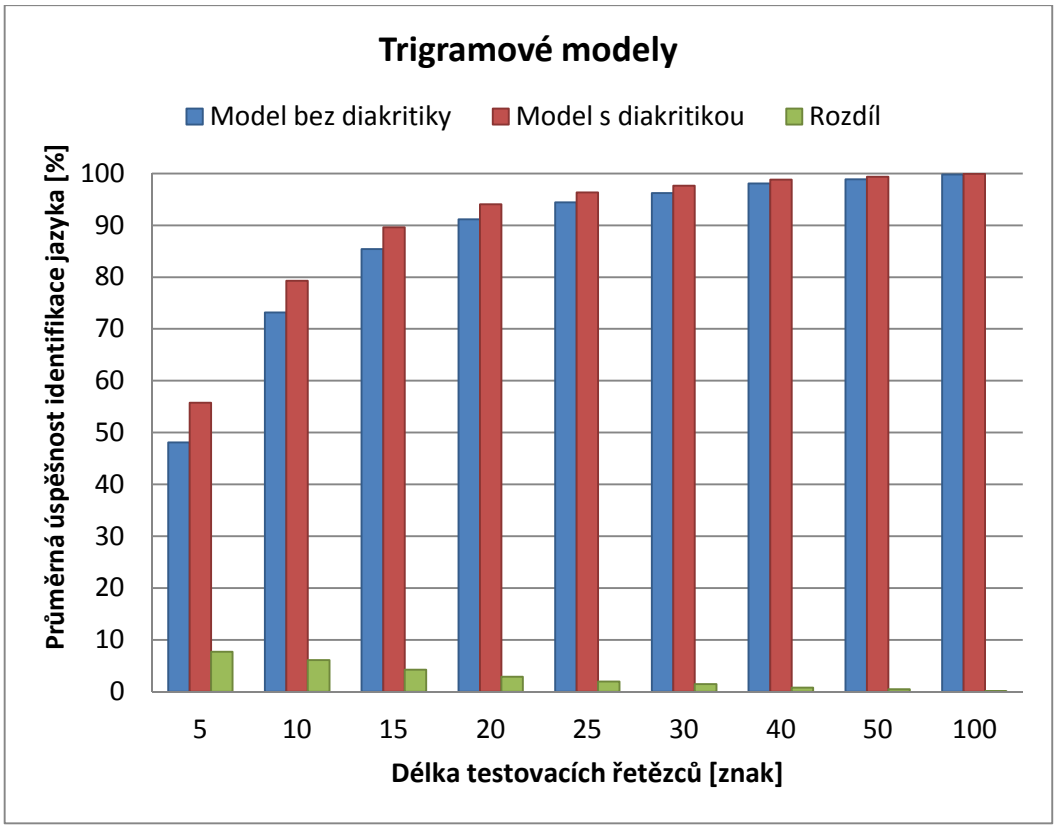
Příloha C
Porovnání modelů podle diakritiky



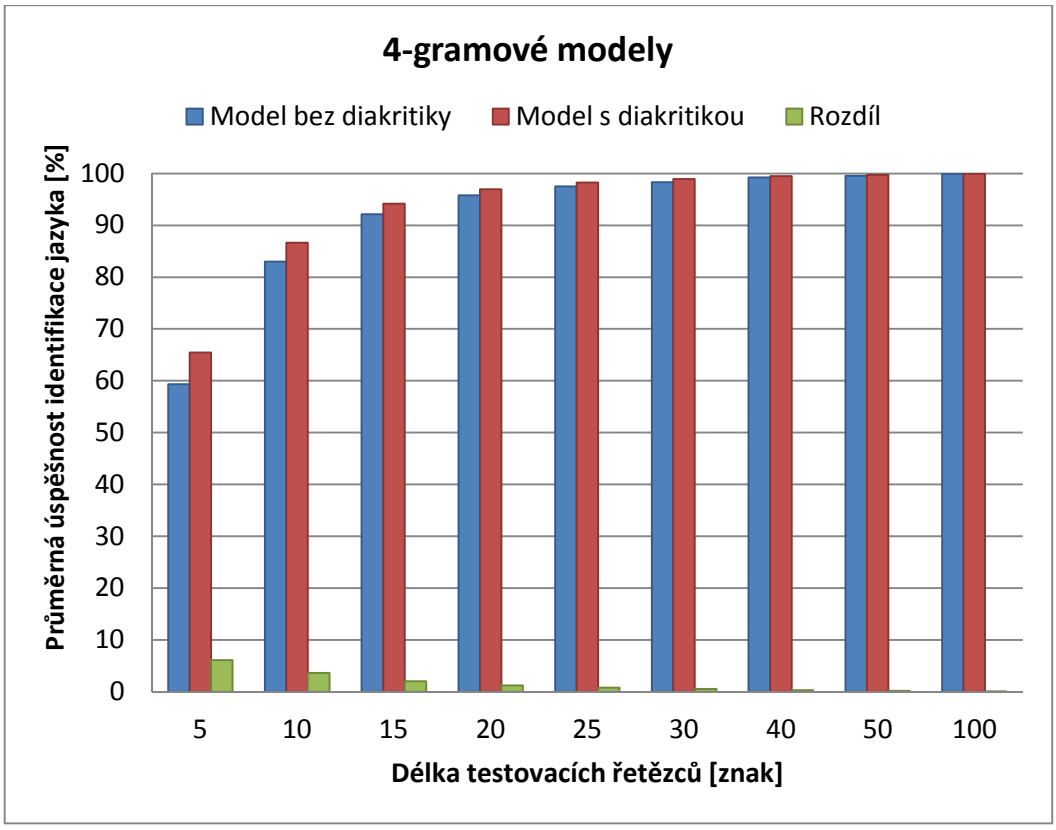
Graf C.1 Porovnání unigramových modelů podle diakritiky



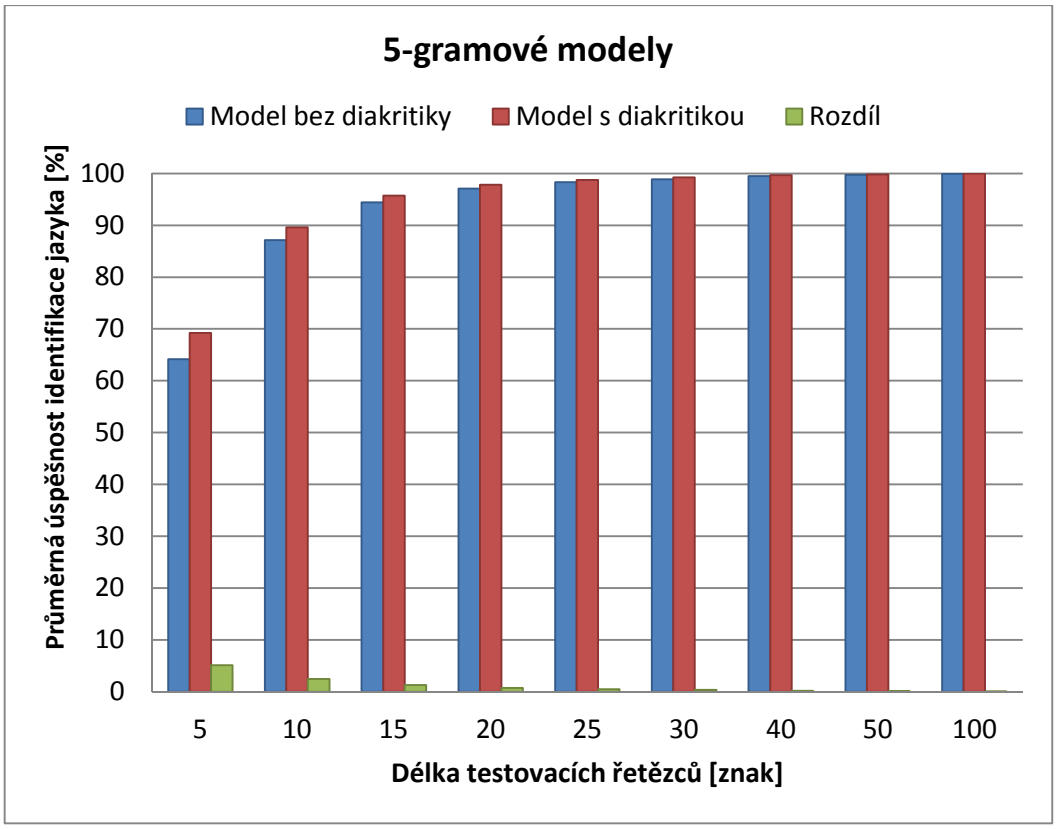
Graf C.2 Porovnání bigramových modelů podle diakritiky



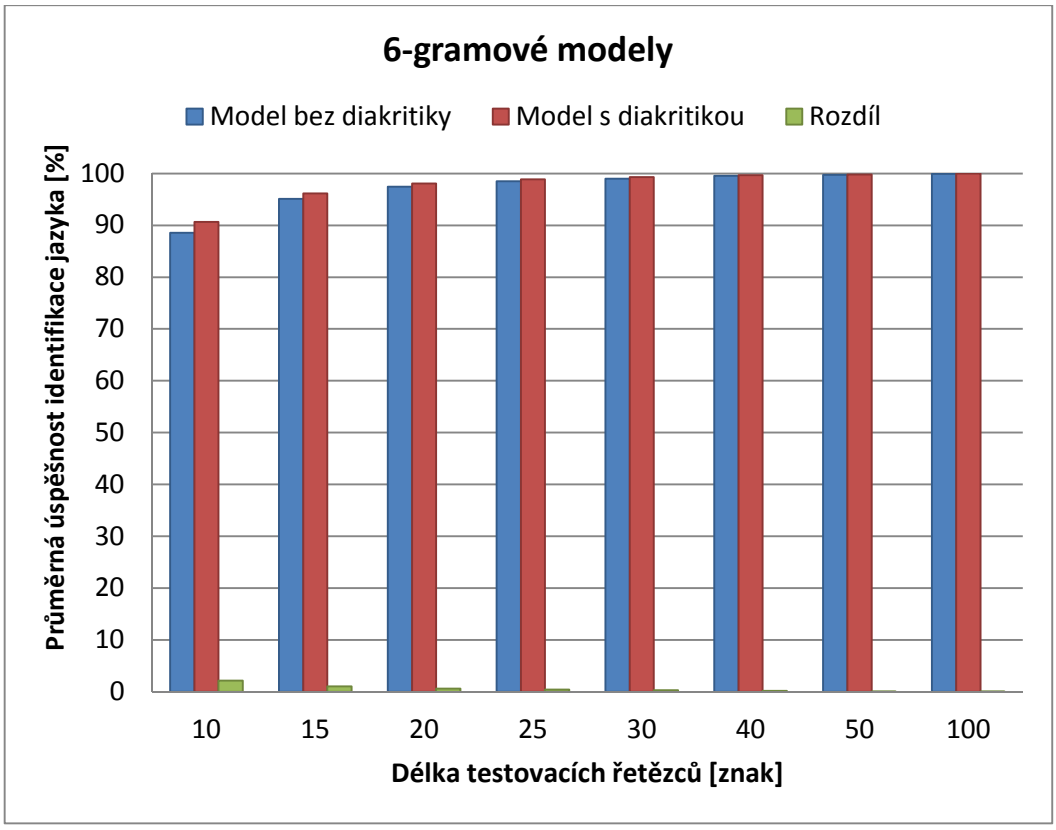
Graf C.3 Porovnání trigramových modelů podle diakritiky



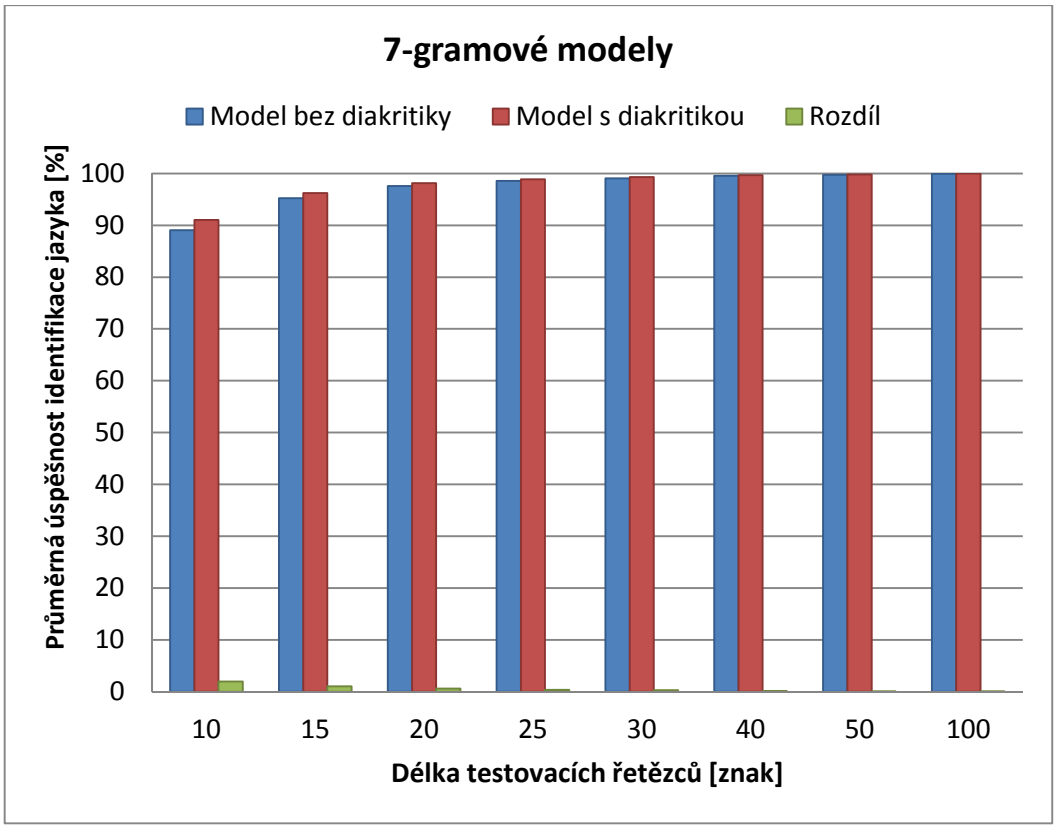
Graf C.4 Porovnání 4-gramových modelů podle diakritiky



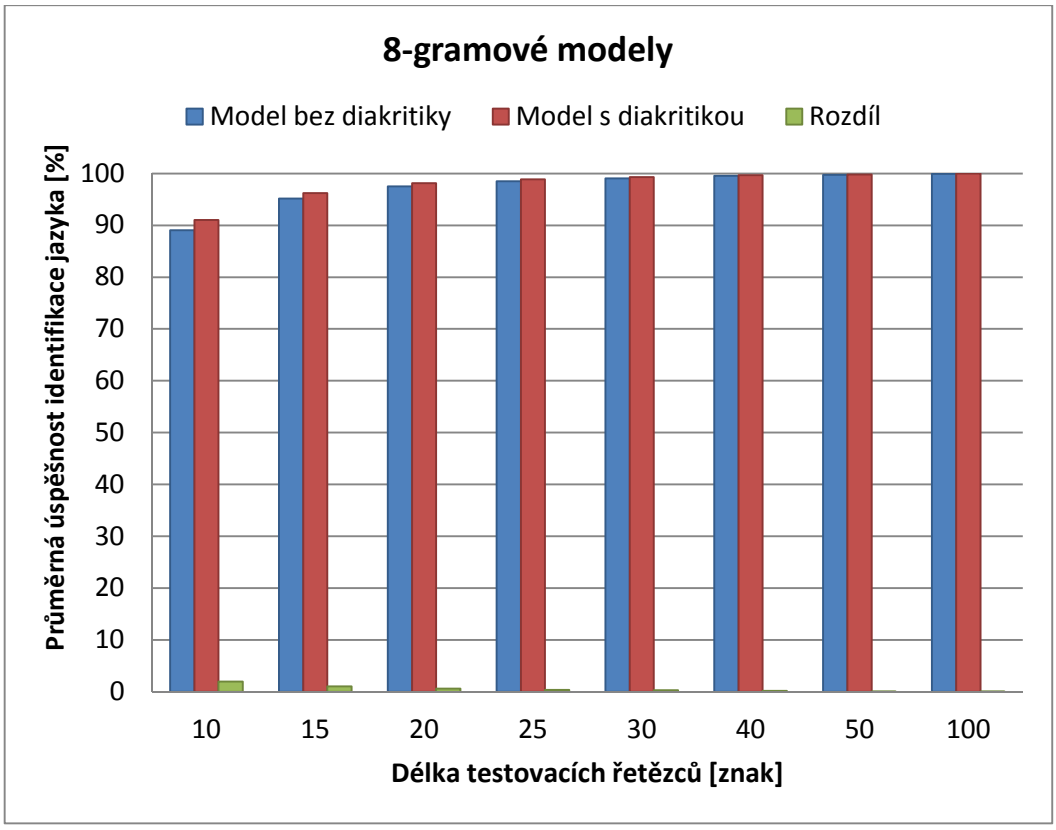
Graf C.5 Porovnání 5-gramových modelů podle diakritiky



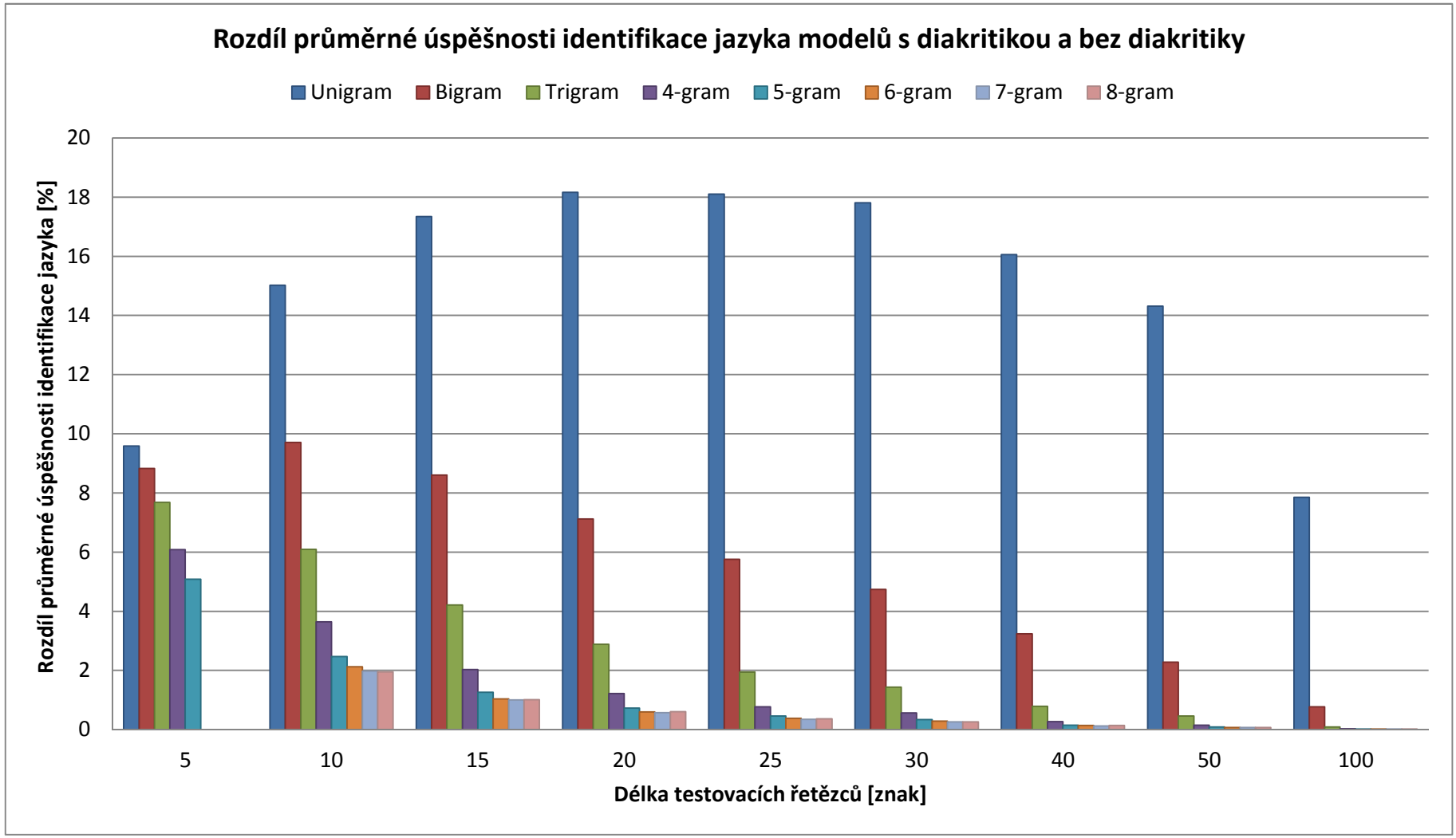
Graf C.6 Porovnání 6-gramových modelů podle diakritiky



Graf C.7 Porovnání 7-gramových modelů podle diakritiky



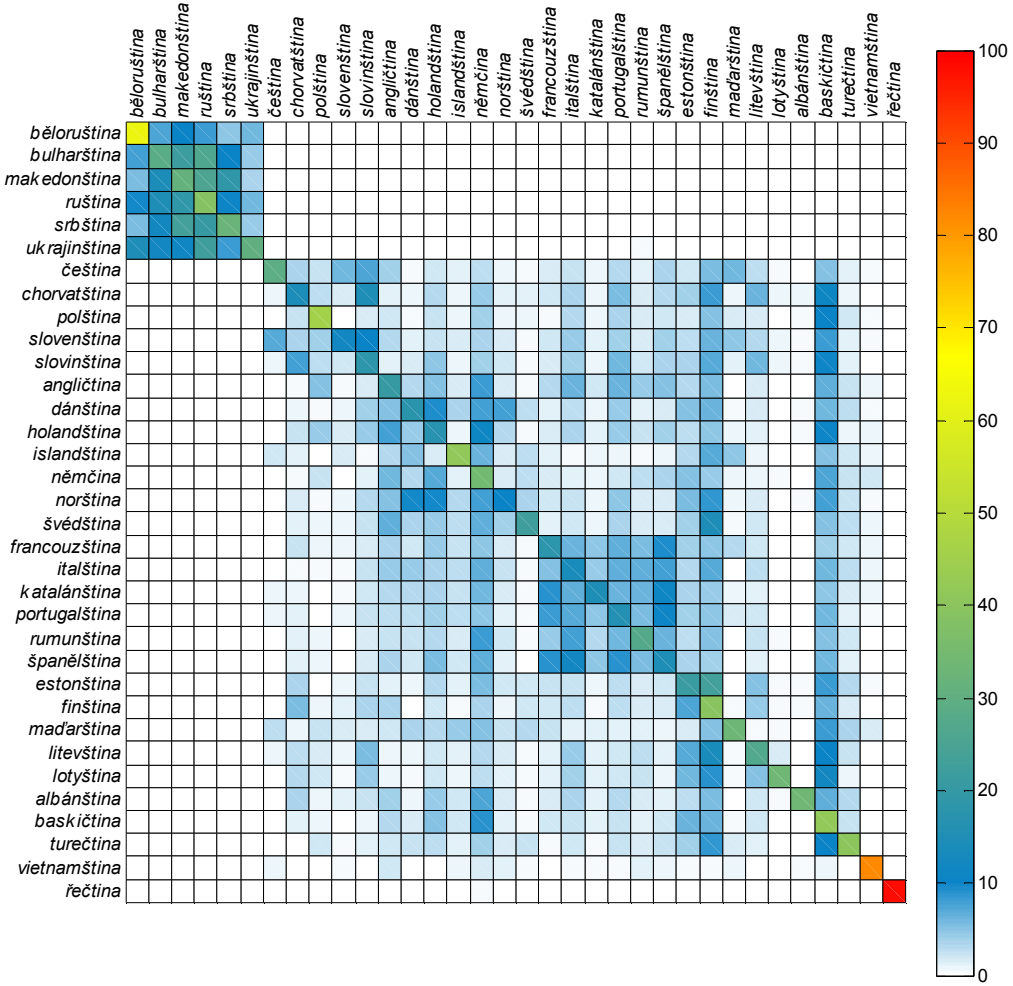
Graf C.8 Porovnání 8-gramových modelů podle diakritiky



Graf C.9 Porovnání rozdílů úspěšnosti identifikace jazyka modelů s diakritikou a bez diakritiky

Příloha D – Konfuzní matice

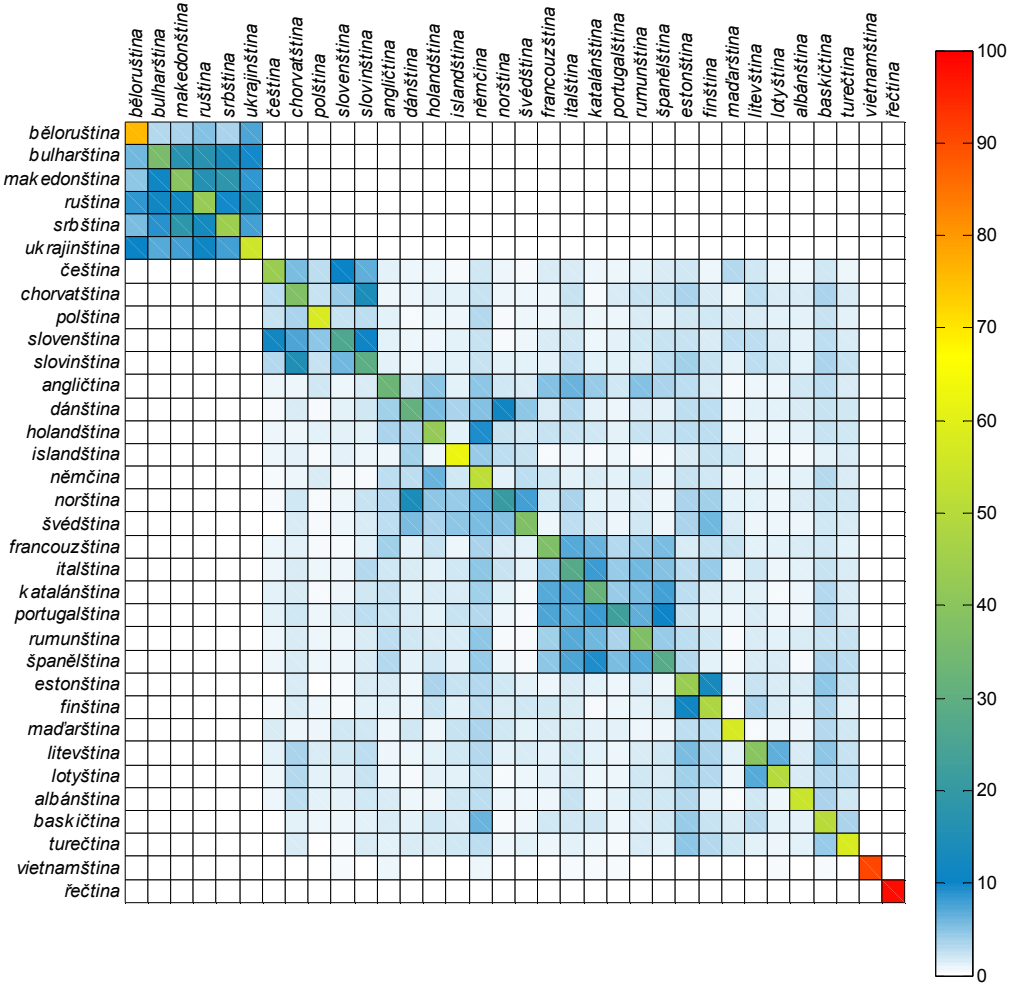
Graf	Tabulka	Stupeň modelu	Vyhlazovací technika	Délka testovacích řetězců [znak]	Strany
D.1	D.1	Unigramový	Witten-Bell	5	D2, D3
D.2	D.2	Bigramový	Witten-Bell	5	D4, D5
D.3	D.3	Trigramový	Witten-Bell	5	D6, D7
D.4	D.4	4-gramový	Witten-Bell	5	D8, D9
D.5	D.5	5-gramový	Witten-Bell	5	D10, D11
D.6	D.6	6-gramový	Witten-Bell	5	D12, D13
D.7	D.7	7-gramový	Witten-Bell	5	D14, D15
D.8	D.8	8-gramový	Witten-Bell	5	D16, D17
D.9	D.9	Unigramový	Witten-Bell	20	D18, D19
D.10	D.10	Bigramový	Witten-Bell	20	D20, D21
D.11	D.11	Trigramový	Witten-Bell	20	D22, D23
D.12	D.12	4-gramový	Witten-Bell	20	D24, D25
D.13	D.13	5-gramový	Witten-Bell	20	D26, D27
D.14	D.14	6-gramový	Witten-Bell	20	D28, D29
D.15	D.15	7-gramový	Witten-Bell	20	D30, D31
D.16	D.16	8-gramový	Witten-Bell	20	D32, D33



Graf D.1 Konfuzní matice unigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	62,23	7,79	10,26	8,21	4,90	6,00	0	0,02	0,04	0,02	0,01	0,02	0	0,01	0	0,10	0	0,01	0	0	0,01	0,01	0,23	0	0,01	0,04	0	0,01	0	0	0,04	0,01	0,01	0
bulharština	8,19	28,60	21,96	26,19	10,37	4,32	0	0	0,02	0	0,02	0,02	0	0	0	0,06	0,01	0	0	0,02	0	0	0,13	0,03	0,01	0	0,01	0,01	0	0	0,02	0	0,01	0
makedonština	5,67	14,18	30,89	25,51	18,98	3,79	0	0,02	0,04	0,04	0,04	0,04	0,03	0,03	0	0,05	0,01	0	0	0,01	0,04	0,05	0,34	0,08	0,03	0,05	0	0,02	0	0	0,05	0,01	0	0
ruština	10,08	14,89	18,85	39,00	10,70	5,98	0	0	0,03	0	0	0	0	0	0	0,13	0	0,01	0	0,02	0,01	0,02	0,23	0	0,03	0	0	0	0	0	0	0,01	0,01	0
srbština	5,49	12,21	23,30	20,65	32,50	4,62	0	0,16	0,05	0,02	0,12	0	0	0,07	0,01	0,09	0,03	0,01	0,04	0,07	0,01	0,11	0,12	0,03	0,04	0,10	0	0,01	0	0,01	0,12	0,01	0	0
ukrajinština	14,59	11,87	11,50	22,57	8,47	30,46	0	0	0	0	0,02	0,02	0	0,01	0	0,03	0	0	0	0,01	0	0	0,40	0	0,02	0,01	0	0	0,01	0	0	0	0,01	0
čeština	0	0	0	0	0	0	29,36	3,69	2,59	5,91	7,80	3,93	0,73	2,08	1,39	2,77	1,11	0,61	1,65	2,49	0,95	3,41	1,38	3,60	2,19	5,48	6,19	2,85	0,66	0,29	5,13	1,30	0,46	0
chorvatština	0	0	0	0	0	0	0,99	14,14	2,78	1,94	15,21	1,24	1,02	3,45	0,86	4,36	1,41	1,22	2,19	3,60	0,79	5,49	1,91	3,41	4,12	8,35	0,94	6,33	1,15	0,79	10,94	1,06	0,31	0
polština	0	0	0	0	0	0	0,01	2,43	46,06	0,30	1,95	2,23	0,57	2,36	0,88	4,29	0,98	0,79	0,48	3,37	1,03	3,72	1,77	2,17	1,82	5,10	1,88	1,92	0,22	0,60	10,54	2,13	0,40	0
slovenština	0	0	0	0	0	0	7,39	3,62	4,15	12,05	10,22	3,33	1,20	2,53	1,79	3,71	1,69	0,64	2,15	4,66	1,33	4,12	1,79	3,98	3,08	6,49	4,95	3,40	0,80	0,42	8,38	1,36	0,77	0
slovinština	0	0	0	0	0	0	1,06	7,91	3,06	2,24	18,54	1,49	1,74	4,69	1,15	4,03	2,08	0,57	1,83	3,94	1,16	6,01	2,09	3,88	3,76	7,20	1,21	6,09	1,15	0,76	10,93	1,34	0,09	0
angličtina	0	0	0	0	0	0	0,02	0,70	5,14	0,49	1,78	20,51	3,40	5,40	1,77	8,38	1,74	0,68	3,17	6,55	2,07	6,54	4,39	5,23	3,27	5,69	0,36	1,88	0,20	0,31	6,73	2,49	1,11	0
dánština	0	0	0	0	0	0	0	1,15	0,64	1,07	4,02	5,23	17,74	9,76	3,78	8,18	8,01	2,95	1,33	3,05	0,84	4,49	1,48	1,67	5,36	6,60	0,55	1,91	0,17	0,48	5,94	2,95	0,65	0
holandština	0	0	0	0	0	0	0,10	2,54	4,51	1,94	4,48	7,99	4,56	17,01	1,13	11,01	3,26	0,54	1,64	3,63	1,18	4,68	2,37	3,91	2,95	5,05	0,82	1,21	0,10	0,59	10,71	1,17	0,92	0
islandština	0	0	0	0	0	0	2,21	1,22	0,39	1,72	0,76	3,27	5,11	1,64	41,56	6,49	1,58	2,90	1,31	0,58	0,87	1,08	0,74	0,88	3,39	7,31	4,95	1,05	0,34	0,33	5,27	2,03	1,02	0
němčina	0	0	0	0	0	0	0,02	0,62	2,70	0,39	1,45	6,17	3,20	7,29	1,19	34,70	1,62	2,88	1,62	1,54	1,09	2,08	2,96	3,81	5,09	4,39	0,82	1,06	0,53	0,33	7,59	2,72	2,14	0
norština	0	0	0	0	0	0	0,03	1,57	0,54	1,03	3,50	5,33	9,97	10,11	3,20	7,99	10,35	3,60	1,96	2,72	0,89	4,79	1,69	1,70	5,74	8,81	0,88	1,81	0,25	0,51	7,86	2,62	0,55	0
švédština	0	0	0	0	0	0	0,03	1,21	1,00	1,10	2,46	6,97	3,55	4,36	3,05	6,80	4,12	22,62	1,31	2,02	1,12	3,53	1,60	1,83	4,24	14,52	0,66	2,07	0,35	0,33	5,21	2,81	1,13	0
francouzština	0	0	0	0	0	0	0,25	2,71	0,91	0,85	1,68	3,88	2,28	4,38	2,39	4,94	1,80	0,66	18,15	6,48	4,71	6,90	5,62	9,66	3,91	4,80	3,37	2,11	0,29	0,15	4,12	1,97	1,03	0
italština	0	0	0	0	0	0	0,01	0,64	0,77	0,69	2,54	4,56	4,61	3,55	3,02	6,82	2,69	0,45	5,10	13,99	4,57	7,03	6,85	8,06	3,48	7,15	0,25	2,81	0,35	0,06	6,19	2,97	0,79	0
katalánština	0	0	0	0	0	0	0,86	0,96	0,31	0,59	2,01	3,13	3,41	3,81	2,53	6,02	1,91	0,45	9,28	6,93	14,91	6,46	6,36	11,25	3,67	4,65	0,85	1,55	0,23	0,38	4,80	1,77	0,92	0
portugalština	0	0	0	0	0	0	1,14	1,34	0,34	1,09	2,40	2,83	3,05	4,02	2,82	5,01	1,51	0,42	8,29	7,17	4,96	15,68	5,68	10,67	4,13	4,75	1,81	2,20	0,23	0,28	5,95	1,48	0,75	0
rumunština	0	0	0	0	0	0	0	1,18	0,99	0,25	1,63	2,67	2,36	3,16	1,89	8,43	2,06	0,57	4,55	8,00	3,28	6,01	27,45	6,55	2,86	5,44	0,19	2,40	0,41	0,05	5,30	1,97	0,35	0
španělština	0	0	0	0	0	0	0,02	1,22	1,04	0,33	1,74	3,64	2,17	5,71	1,98	6,70	1,34	0,23	9,16	11,94	4,81	9,36	5,47	14,93	3,82	4,27	0,30	1,54	0,14	0,35	5,86	1,54	0,39	0
estonština	0	0	0	0	0	0	0	3,79	0,05	0,85	2,43	1,20	0,98	3,34	1,18	5,59	2,19	2,19	2,43	2,57	1,17	2,74	1,60	2,32	21,21	23,06	0,09	5,30	0,70	0,58	8,33	3,44	0,67	0
finština	0	0	0	0	0	0	0,03	5,57	1,08	1,35	3,85	3,88	0,36	2,15	0,48	3,62	0,91	2,18	2,28	2,96	0,61	2,90	1,60	1,92	7,78	39,58	0,61	4,46	0,75	0,40	6,63	1,85	0,21	0
maďarština	0	0	0	0	0	0	2,83	1,06	2,66	1,61	1,77	2,18	3,59	3,35	4,40	5,12	2,71	3,41	2,45	1,56	1,45	1,54	1,42	1,06	1,57	5,09	33,56	0,63	0,51	0,19	8,59	3,88	1,81	0
litevština	0	0	0	0	0	0	0,87	2,98	1,88	0,90	5,69	0,84	0,88	2,09	1,54	3,24	1,68	0,63	1,23	4,32	1,49	2,27	2,76	1,21	7,29	13,68	0,67	26,73	1,84	0,17	10,52	2,57	0,03	0
lotyština	0	0	0	0	0	0	0,10	3,28	1,98	0,43	4,32	0,89	0,58	2,16	0,90	2,83	1,23	0,42	1,18	4,27	1,34	2,13	2,64	1,17	5,86	9,37	0,41	5,19	33,61	0,20	12,45	0,93	0,13	0
albánština	0	0	0	0	0	0	0,02	3,73	1,10	1,32	2,73	4,26	0,93	4,64	2,19	7,75	1,74	0,78	1,64	3,83	1,53	3,14	1,83	1,24	3,03	5,62	0,30	2,26	0,51	33,83	6,67	3,23	0,15	0
baskičtina	0	0	0	0	0	0	0	1,55	1,15	0,21	0,88	3,23	1,91	5,45	2,13	9,05	1,40	0,42	2,31	2,58	1,25	2,54	1,39	2,24	6,49	6,38	0,49	2,09	0,14	0,16	41,82	2,38	0,36	0
turečtina	0	0	0	0	0	0	0,04	0,29	2,31	0,51	1,30	2,30	2,42	2,77	1,55	4,29	1,72	2,70	0,50	2,10	0,66	2,36	1,87	2,66	4,23	8,72	1,79	1,36	0,32	0,11	10,44	40,44	0,24	0
vietnamština	0	0	0	0	0	0	1,02	0,20	0,35	0,56	0,15	2,10	0,05	0,23	0,92	1,92	1,42	0,42	0,35	0,55	0,65	0,48	1,37	0,83	0,48	0,70	1,00	0,11	0,08	0,44	1,00	0,31	82,30	0
řečtina	0	0	0	0	0	0	0	0,02	0,05	0,03	0,01	0,11	0,01	0,08	0,03	0,42	0	0,02	0,02	0,06	0,04	0,06	0,31	0,15	0,05	0,08	0,01	0,01	0	0,02	0,17	0,11	0,01	98,11

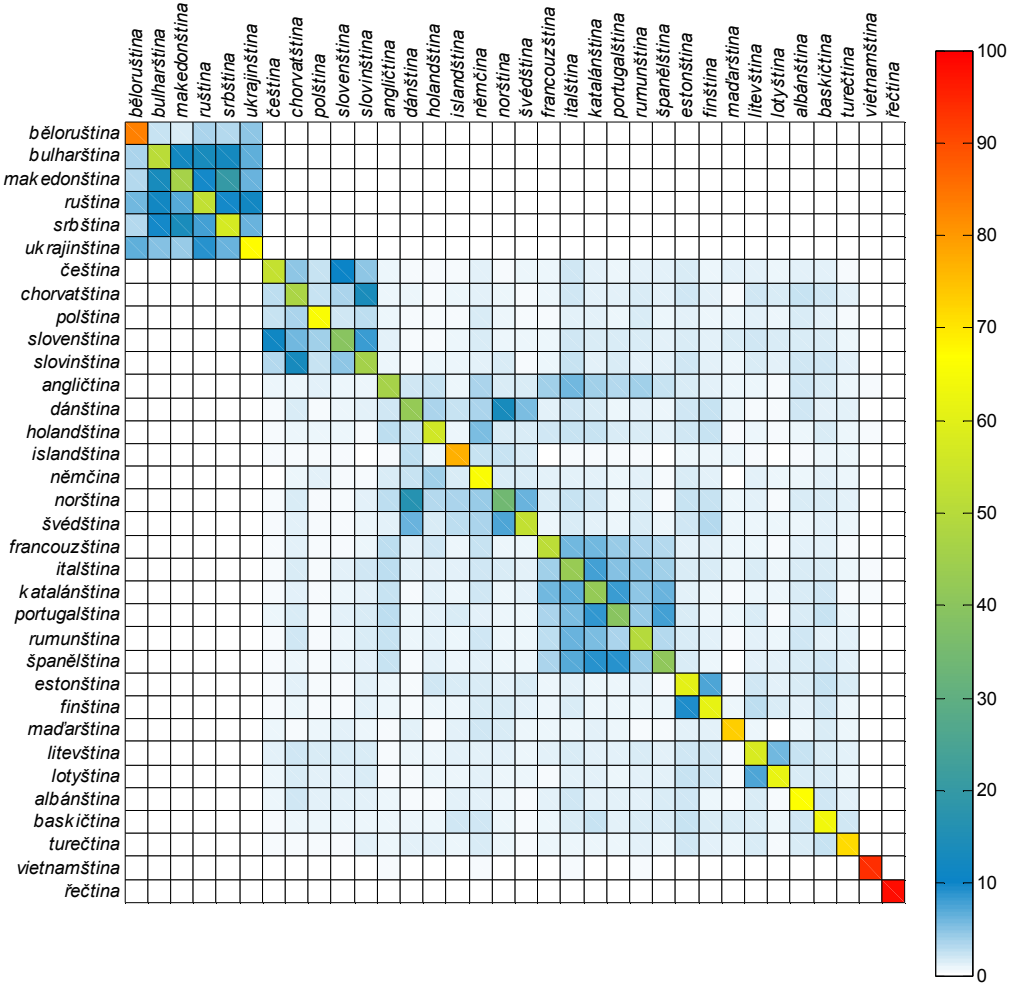
Tabulka D.1 Konfuzní matice unigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.2 Konfuzní matice bigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	75,83	3,19	3,79	5,27	3,74	7,71	0,01	0,03	0,03	0,01	0,04	0,01	0,01	0,01	0	0,05	0,01	0,03	0,03	0,02	0,01	0,01	0,01	0,01	0,02	0	0,03	0,02	0,01	0,02	0,02	0,01	0	0
bulharština	6,13	36,31	16,87	16,94	13,29	10,12	0	0,02	0	0	0,02	0,03	0	0,03	0	0,03	0	0,03	0,03	0,01	0,01	0,01	0,04	0,01	0,02	0	0	0,01	0,01	0,01	0	0,01	0,01	0
makedonština	4,82	11,66	39,54	16,04	18,30	8,78	0,02	0,06	0,02	0,04	0,05	0,05	0	0,03	0,01	0,02	0,03	0,04	0,03	0,04	0,03	0,07	0,10	0,04	0,03	0,02	0,01	0,01	0,01	0,01	0,02	0,03	0,03	0,01
ruština	8,76	11,67	12,26	43,24	9,92	13,80	0	0	0,02	0	0	0,03	0	0,03	0	0,04	0,01	0,01	0,01	0	0,02	0,03	0,05	0,01	0,01	0,01	0	0	0,01	0,01	0,01	0,01	0,03	0
srbština	5,50	9,06	18,58	13,25	44,29	8,09	0,02	0,31	0,01	0,07	0,16	0,01	0,02	0,08	0,01	0,04	0,01	0,01	0,01	0,01	0,04	0,02	0,07	0,06	0,08	0,02	0,03	0,03	0,01	0,02	0,05	0,03	0	0
ukrajinština	10,18	7,23	7,96	11,25	7,93	55,08	0	0,01	0	0	0,01	0,02	0,02	0,07	0	0,03	0	0,03	0,03	0	0	0	0,01	0,01	0,01	0	0,03	0	0	0,01	0,01	0	0,06	0
čeština	0,01	0	0	0,04	0,02	0,04	43,55	5,85	3,02	10,48	6,81	1,42	0,97	0,79	0,69	2,07	0,90	0,72	1,59	1,62	1,03	1,10	1,48	1,75	2,13	1,06	3,18	2,29	0,86	1,09	2,20	1,17	0,06	0,01
chorvatština	0,01	0	0,01	0,01	0,05	0	2,86	38,05	2,40	4,35	14,27	0,93	0,83	1,23	1,21	2,39	1,08	0,84	1,12	2,60	0,63	1,57	2,38	2,42	3,70	1,76	1,11	3,11	1,83	1,69	3,70	1,65	0,17	0,04
polština	0,01	0,01	0,01	0	0,01	0	2,49	3,68	58,06	2,64	2,95	1,33	0,43	1,16	0,81	3,28	0,46	0,85	0,98	1,66	0,81	1,01	1,75	1,49	2,02	1,97	1,62	1,71	1,47	1,43	2,55	1,26	0,06	0,02
slovenština	0,02	0	0,01	0,04	0,02	0	11,84	7,45	4,78	26,42	10,84	1,52	0,99	0,87	1,21	2,39	0,70	1,08	1,61	2,34	1,15	1,52	2,29	2,40	2,73	1,65	3,01	2,97	1,92	1,49	2,95	1,57	0,21	0,01
slovinština	0,02	0	0,04	0,03	0,03	0,03	3,18	15,54	2,52	6,03	29,65	1,38	1,16	1,53	1,31	2,40	1,40	1,20	1,40	3,00	1,44	1,28	1,92	2,87	4,06	2,36	1,31	3,03	2,26	1,43	3,55	2,56	0,07	0,01
angličtina	0,02	0	0,01	0,03	0,01	0,01	1,06	1,09	2,19	0,90	1,57	32,98	2,71	4,76	1,54	5,06	2,23	1,63	5,40	6,55	4,33	2,29	5,45	3,52	3,05	1,87	0,74	1,10	0,85	2,14	2,77	1,82	0,29	0,03
dánština	0,05	0,03	0	0,05	0,01	0	0,62	1,65	0,72	1,47	2,10	4,01	31,14	5,79	3,81	5,26	11,52	5,04	1,68	3,36	1,55	1,09	1,59	1,45	3,06	2,90	0,99	1,18	1,28	1,81	2,41	2,21	0,13	0,04
holandština	0,03	0,01	0,01	0,04	0,04	0,07	1,14	0,84	1,21	1,21	1,36	3,80	3,66	42,33	1,23	9,67	2,51	2,09	2,46	2,36	2,00	1,39	2,58	2,28	2,88	3,10	1,13	1,05	1,09	1,49	2,71	1,97	0,23	0,03
islandština	0,01	0	0,01	0,05	0	0,01	1,12	1,40	0,41	1,33	0,82	0,80	4,17	1,11	62,36	4,41	3,06	2,50	0,62	0,96	0,64	1,03	0,69	0,40	1,75	2,35	2,24	0,93	0,69	0,61	1,52	1,82	0,16	0,01
němčina	0,05	0,02	0,02	0	0	0,01	0,56	0,82	1,79	0,68	0,67	2,82	2,99	6,26	2,03	51,91	1,46	2,82	2,27	1,56	1,93	1,46	2,30	0,98	2,60	2,28	0,82	1,21	1,06	1,34	3,32	1,71	0,23	0,01
norština	0,02	0	0	0,01	0,01	0,02	0,70	2,27	0,64	0,96	2,39	3,30	14,15	4,78	4,54	6,96	21,06	7,93	2,01	3,76	1,39	1,38	1,86	0,92	3,75	4,04	1,45	1,54	1,11	1,86	2,65	2,34	0,18	0,02
švédština	0	0,01	0	0,04	0	0,01	0,37	1,74	0,62	0,98	1,75	2,81	5,58	3,74	4,73	5,59	5,30	37,85	1,10	3,02	1,77	1,12	2,06	1,17	3,85	5,88	1,68	0,81	1,04	0,99	2,31	1,87	0,19	0,02
francouzština	0,03	0	0	0,02	0	0,02	0,94	1,36	0,53	0,64	1,33	3,98	1,44	2,35	1,11	3,77	1,66	1,19	37,49	7,41	6,40	3,13	4,62	5,65	1,58	2,68	2,38	1,46	1,33	1,62	2,08	1,40	0,37	0,03
italština	0,02	0	0,03	0,06	0	0,01	0,81	1,95	1,11	1,08	3,21	2,34	1,80	1,56	2,10	4,91	2,68	1,46	5,01	27,94	8,44	4,39	6,07	5,37	2,92	4,46	0,89	2,00	0,92	1,55	2,64	1,86	0,39	0,02
katalánština	0	0	0	0,01	0	0	1,27	1,60	0,66	1,10	2,14	2,41	1,30	1,89	1,74	4,02	1,52	0,78	7,34	7,50	31,75	4,61	5,65	8,12	2,96	1,85	1,48	1,24	0,68	0,87	3,16	1,94	0,39	0,02
portugalština	0,01	0,02	0,02	0,02	0,03	0,01	1,44	2,28	0,97	1,61	2,79	2,51	1,65	1,20	2,66	3,44	1,03	0,55	7,64	7,25	8,59	22,78	6,67	10,62	2,51	1,50	1,15	1,66	0,85	1,16	3,43	1,75	0,17	0,03
rumunština	0,04	0	0	0,01	0,01	0	0,81	1,84	1,17	0,93	1,72	3,05	2,02	1,67	1,85	5,06	0,73	0,67	3,97	7,16	6,03	3,52	37,52	4,50	3,07	2,09	0,61	1,83	1,39	1,58	2,38	2,41	0,31	0,05
španělština	0	0	0	0,02	0	0	0,87	1,78	1,12	0,98	1,90	3,27	1,31	2,11	1,44	4,68	0,89	0,63	4,90	7,55	9,48	5,61	7,15	28,26	3,29	1,35	0,49	1,66	1,77	0,71	3,59	2,90	0,28	0,01
estonština	0,03	0	0	0	0,02	0,01	0,38	1,72	0,27	0,53	1,95	1,83	0,97	3,52	2,68	3,43	2,17	1,21	1,08	1,66	1,34	1,14	1,88	0,94	43,64	13,04	0,94	2,52	1,67	1,93	4,96	2,44	0,09	0,01
finština	0,05	0	0,02	0,01	0,01	0,02	0,69	1,71	1,07	0,67	1,39	1,47	0,80	2,45	1,48	3,09	1,64	2,08	2,04	1,80	0,47	1,12	1,42	1,76	11,46	48,31	0,73	3,76	1,83	1,32	3,89	1,32	0,12	0
maďarština	0	0	0,01	0,02	0,02	0,03	1,73	0,89	1,17	1,99	2,08	0,85	2,21	0,99	2,59	3,55	2,05	1,21	1,76	1,63	1,35	1,28	0,90	0,80	2,91	2,82	57,25	0,67	1,07	0,81	3,18	2,01	0,15	0,01
litevština	0,01	0	0	0	0	0,02	1,18	3,54	1,75	1,96	2,91	1,06	1,08	1,56	2,10	3,23	1,48	1,85	1,35	2,04	1,25	1,05	1,85	2,33	5,55	3,64	1,24	39,83	6,98	1,73	5,06	2,36	0,01	0
lotyština	0	0	0	0	0,01	0	0,90	3,33	1,25	1,41	2,66	0,81	0,72	1,20	1,48	2,50	0,71	0,96	1,19	1,80	1,08	1,00	2,01	1,78	4,21	3,42	0,93	7,08	49,71	1,79	3,14	2,84	0,08	0
albánština	0	0	0,01	0,01	0,02	0,01	0,86	3,05	1,25	1,54	2,16	1,80	1,01	1,15	2,17	2,84	1,11	1,05	1,06	2,71	0,96	1,44	2,04	2,07	3,50	1,41	0,77	2,25	1,16	54,48	3,82	2,13	0,15	0
baskičtina	0	0	0	0	0	0	0,34	1,23	0,79	1,06	1,23	1,73	1,45	1,97	1,68	6,56	0,62	0,87	2,07	2,08	1,99	0,98	1,72	2,32	4,30	2,48	1,69	3,31	1,41	1,55	50,74	3,76	0,05	0,01
turečtina	0	0	0,01	0,04	0	0,01	0,34	1,67	0,22	0,56	1,62	1,24	1,86	1,79	2,00	3,12	1,08	1,23	1,01	1,49	1,02	0,73	1,77	1,32	4,71	3,25	2,14	1,19	0,81	1,19	4,66	57,85	0,03	0,03
vietnamština	0	0	0	0	0	0,01	0,26	0,12	0,15	0,57	0,10	1,17	0,14	0,29	0,13	0,88	0,20	0,23	0,35	0,46	0,48	0,61	0,34	0,34	0,17	0,19	0,15	0,10	0,40	0,28	0,62	0,24	91,00	0,01
řečtina	0,01	0	0	0,03	0	0,01	0,02	0,03	0,04	0	0,05	0,11	0,01	0,13	0,01	0,20	0,06	0,10	0,09	0,12	0,05	0,02	0,07	0,06	0,07	0,03	0,06	0,05	0,03	0,09	0,11	0,08	0,06	98,19

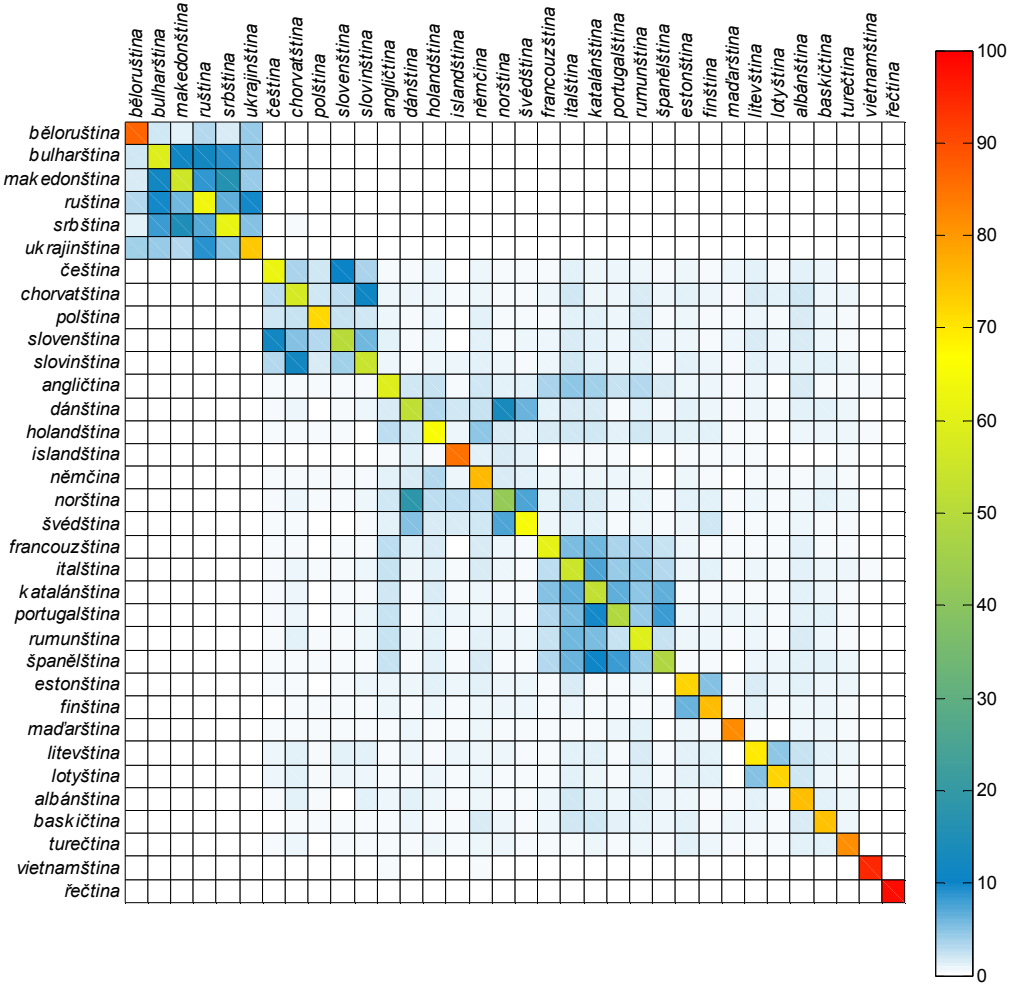
Tabulka D.2 Konfuzní matice bigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.3 Konfuzní matice trigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	83,36	2,50	1,75	3,65	3,29	4,97	0,01	0,02	0,04	0,01	0,06	0,03	0,01	0,01	0	0,01	0	0,01	0,03	0	0,02	0,01	0	0,04	0,03	0,02	0,02	0	0,02	0,02	0,02	0,01	0,01	0,01
bulharština	3,81	50,86	12,26	13,36	12,50	6,89	0,02	0,02	0	0,02	0,03	0,03	0	0,01	0	0,01	0	0,02	0	0	0,01	0,01	0,05	0,01	0,02	0,01	0,02	0	0	0,01	0	0	0,01	0
makedonština	3,41	13,29	46,19	9,78	19,99	6,53	0,01	0,07	0,03	0,04	0,09	0,04	0,02	0,01	0,02	0	0,01	0,02	0,04	0,07	0,01	0,07	0,07	0,02	0,02	0,01	0,01	0,04	0,01	0,02	0,03	0,01	0,02	0
ruština	5,91	11,81	7,35	52,71	9,98	11,95	0	0	0,01	0,02	0	0,01	0,02	0,01	0	0,02	0	0	0,01	0,01	0,04	0,01	0,02	0	0,01	0,01	0,01	0	0,01	0	0,01	0	0,04	0,01
srbština	3,43	9,97	13,70	7,97	57,27	6,51	0,05	0,31	0,03	0,03	0,19	0,01	0	0,02	0,02	0,02	0,01	0	0	0,04	0,01	0,02	0,08	0,02	0,03	0	0,01	0,02	0,03	0,08	0,06	0,04	0,01	0
ukrajinština	6,66	5,35	4,59	9,36	6,53	67,23	0	0,01	0,01	0	0	0,04	0	0,03	0	0,03	0,01	0,01	0,03	0,01	0	0	0,03	0	0,01	0	0,01	0	0	0	0	0	0,04	0
čeština	0	0,01	0	0,06	0,01	0,03	53,42	4,70	2,36	10,32	4,87	1,17	0,53	0,72	0,54	1,46	0,69	0,98	0,91	2,03	1,27	0,96	1,27	1,27	1,82	0,90	1,56	1,50	0,98	1,52	1,38	0,60	0,13	0,02
chorvatština	0,01	0	0,01	0,09	0,05	0,01	2,86	47,37	2,51	3,66	13,79	1,00	0,81	0,76	0,85	1,29	0,87	0,70	0,90	2,16	1,45	1,49	1,73	1,49	2,12	1,22	0,74	2,05	1,67	2,45	2,26	1,28	0,31	0,04
polština	0	0,01	0	0,04	0	0,01	2,37	3,55	65,68	2,16	2,92	0,95	0,55	0,72	0,46	1,80	0,79	0,63	0,78	1,48	1,24	1,05	1,68	1,03	1,30	1,11	1,01	1,37	1,14	1,93	1,42	0,68	0,12	0,01
slovenština	0,01	0	0,01	0,03	0,02	0,02	11,33	6,15	3,91	39,47	8,41	1,41	0,48	0,74	0,81	1,79	0,82	1,00	0,96	2,32	1,92	1,58	1,79	1,21	1,83	1,20	1,84	2,27	1,88	1,88	1,90	0,82	0,17	0,02
slovinština	0,03	0	0	0,04	0,03	0,03	3,20	13,15	2,43	4,76	45,76	0,91	0,73	0,87	0,98	1,50	1,65	0,76	1,12	2,39	1,49	1,65	1,55	1,36	2,08	1,48	0,97	1,86	1,52	1,98	2,05	1,35	0,31	0,01
angličtina	0,02	0	0,01	0,08	0	0,01	0,83	1,08	1,32	1,17	1,08	46,14	2,24	2,61	1,02	3,77	1,69	1,63	4,28	5,99	4,12	3,27	3,97	2,36	1,92	1,32	0,83	1,16	0,58	2,04	1,74	1,08	0,61	0,03
dánština	0,05	0,02	0	0,07	0,01	0,04	0,59	1,72	0,59	0,86	1,20	1,97	42,34	3,71	2,57	3,89	13,40	5,77	1,54	2,15	1,95	1,14	1,38	0,91	2,05	2,37	1,13	0,77	0,62	2,22	1,50	1,24	0,19	0,04
holandština	0,01	0,01	0	0,20	0	0,10	0,73	1,11	0,98	1,06	0,72	2,82	2,40	56,02	1,13	5,62	1,89	1,76	2,10	2,61	2,37	1,59	1,84	1,45	2,25	2,45	0,66	0,93	0,71	1,07	1,95	0,93	0,44	0,08
islandština	0	0	0	0,11	0	0,03	0,50	0,62	0,41	0,63	0,32	0,77	2,85	0,92	77,00	2,44	2,56	1,83	0,33	0,45	0,54	0,50	0,46	0,30	1,09	0,95	1,16	0,65	0,35	0,43	0,80	0,80	0,18	0,01
němčina	0,04	0	0,04	0,05	0	0,06	0,38	0,81	1,39	0,47	0,45	1,88	2,38	4,29	1,74	65,83	1,83	1,78	1,34	1,38	1,34	1,07	1,46	0,70	1,59	1,27	0,34	1,28	0,81	1,07	1,52	1,02	0,33	0,05
norština	0,01	0	0	0,04	0	0,02	0,63	1,82	0,77	0,72	1,32	2,76	16,69	3,36	3,56	4,46	34,13	6,60	1,65	2,52	1,97	1,14	1,63	0,69	2,39	2,38	1,11	1,42	0,76	1,91	1,74	1,37	0,39	0,04
švédština	0	0	0	0,06	0	0,01	0,43	1,29	0,52	0,46	1,04	1,35	6,27	1,66	2,96	3,60	7,75	53,03	0,92	1,85	1,44	1,05	1,78	0,97	2,08	3,28	0,94	0,86	0,81	1,01	1,40	0,94	0,21	0,02
francouzština	0	0,01	0	0,04	0	0,01	0,76	1,20	0,56	0,67	0,79	3,02	1,26	2,06	0,82	2,56	1,11	0,83	51,18	6,20	6,22	4,55	3,64	3,43	1,25	1,25	0,90	1,10	0,77	1,52	1,43	0,61	0,19	0,05
italština	0,02	0	0,03	0,14	0,01	0,03	0,50	1,77	0,51	1,21	2,18	2,75	1,20	1,35	1,40	2,28	1,64	1,51	4,22	43,24	8,01	5,11	5,06	3,94	1,85	1,69	0,94	1,58	0,98	1,77	1,57	0,99	0,47	0,05
katalánština	0	0	0	0,05	0,01	0,02	0,67	1,39	0,76	0,97	1,50	2,46	0,72	1,42	1,02	2,03	1,09	1,25	5,90	7,01	42,30	8,40	4,95	6,32	1,19	1,03	0,77	1,14	0,79	1,60	1,97	0,74	0,51	0,02
portugalština	0	0,01	0,01	0,03	0	0,01	0,88	1,77	0,78	1,19	1,94	3,06	1,07	1,43	1,61	1,52	0,80	1,12	3,85	5,70	8,77	39,70	4,63	8,15	1,57	1,00	1,14	1,79	0,65	1,86	2,58	1,09	0,25	0,04
rumunština	0,04	0,03	0	0,07	0	0,03	0,61	2,07	0,70	0,99	1,82	2,40	1,10	1,35	0,94	1,96	1,16	1,12	3,12	6,40	5,69	3,58	49,79	3,48	1,69	1,42	0,74	1,33	1,16	2,25	1,20	1,45	0,28	0,03
španělština	0	0	0	0,04	0,01	0	0,44	1,13	0,59	0,91	1,46	2,63	0,60	1,52	1,07	1,55	0,87	1,15	3,60	7,41	9,26	9,17	4,30	41,10	1,58	1,10	0,31	1,26	1,40	1,95	2,04	1,24	0,30	0,01
estonština	0,02	0,01	0	0	0,02	0,01	0,41	1,29	0,75	0,75	1,37	0,93	0,65	2,04	1,66	1,74	1,55	1,70	1,09	1,45	1,07	0,99	1,22	0,61	60,84	7,72	0,61	2,08	1,44	1,75	2,44	1,62	0,15	0,01
finština	0,04	0	0	0,09	0	0,03	0,64	1,10	0,64	0,65	1,44	1,04	0,40	1,05	1,04	1,67	1,71	1,32	1,12	1,60	0,94	0,99	0,93	0,92	9,47	61,42	0,95	2,81	1,62	1,20	2,13	0,80	0,20	0,03
maďarština	0,02	0	0,01	0,09	0,01	0,04	0,97	0,67	0,85	1,33	0,88	0,78	1,35	0,78	1,25	2,03	1,67	0,83	0,84	1,17	1,30	1,04	0,69	0,42	1,30	1,34	73,18	0,85	0,36	1,03	1,75	0,81	0,30	0,05
litevština	0	0	0	0	0	0	1,29	2,08	1,64	1,63	1,92	0,59	1,07	0,87	1,43	1,31	1,49	0,83	1,21	1,64	1,52	1,31	1,83	1,42	2,31	2,22	0,64	57,91	6,01	2,44	1,95	1,33	0,09	0,01
lotyština	0	0	0	0,01	0,01	0,01	0,94	1,92	1,29	1,39	1,78	0,44	0,61	0,82	0,86	1,29	1,00	0,89	0,77	1,39	1,32	1,11	1,49	1,43	2,50	2,06	0,58	7,63	61,94	1,80	1,81	0,79	0,09	0,02
albánština	0,01	0	0,01	0,01	0	0,04	1,06	2,00	1,35	1,19	1,29	1,12	0,60	0,73	1,39	1,26	0,94	0,59	1,19	2,27	1,56	1,52	1,49	1,69	2,00	0,89	0,52	1,83	0,78	66,68	2,16	1,56	0,23	0,03
baskičtina	0	0	0	0,02	0	0	0,41	1,02	0,81	0,86	0,92	1,14	1,05	0,89	2,03	2,23	1,14	0,58	1,10	1,63	2,58	1,46	1,64	1,95	2,73	1,66	1,64	1,63	1,00	2,00	63,68	2,00	0,18	0,01
turečtina	0	0,01	0	0,05	0	0,04	0,48	0,70	0,50	0,58	1,25	1,03	1,20	1,28	0,80	1,80	1,11	0,59	0,86	1,03	1,09	1,01	1,39	0,83	2,00	1,26	0,90	1,62	0,69	1,72	2,37	71,65	0,12	0,03
vietnamština	0	0	0	0,01	0	0,01	0,15	0,18	0,12	0,09	0,12	0,58	0,14	0,13	0,07	0,55	0,19	0,24	0,36	0,63	0,27	0,26	0,54	0,33	0,12	0,15	0,24	0,13	0,13	0,17	0,15	0,04	93,87	0,02
řečtina	0,01	0	0	0,05	0	0,03	0,05	0,04	0,07	0,03	0	0,18	0,01	0,07	0,06	0,24	0,02	0,08	0,08	0,08	0,13	0,04	0,05	0,06	0,07	0,03	0,02	0,03	0,02	0,07	0,03	0,05	0,09	98,20

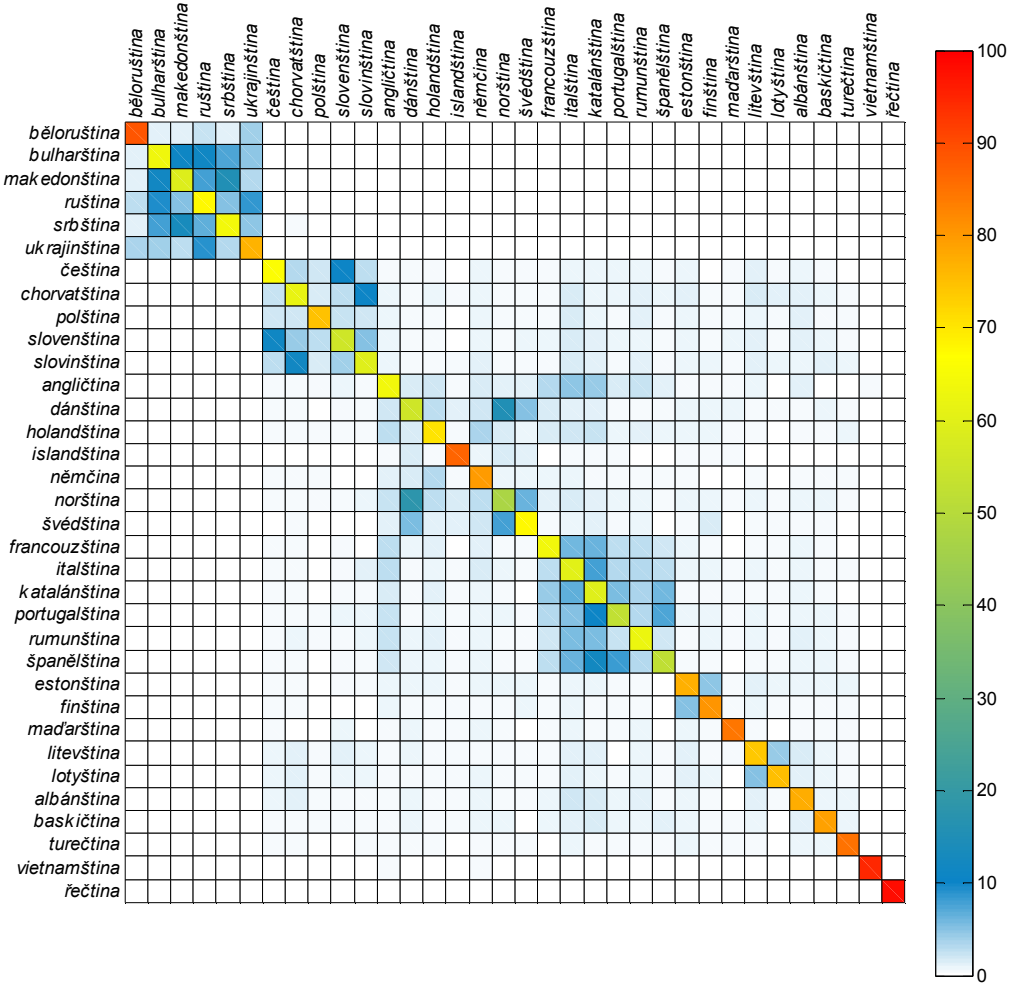
Tabulka D.3 Konfuzní matice trigramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.4 Konfuzní matice 4-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	86,75	2,11	1,38	3,30	1,62	4,37	0,03	0,01	0,05	0,01	0,04	0,02	0,01	0,02	0	0,04	0	0	0,02	0	0,01	0,01	0,02	0,02	0,03	0,02	0,04	0	0,02	0,01	0,01	0,02	0	0
bulharština	2,00	59,34	11,44	12,37	9,21	5,36	0,02	0	0	0,04	0,03	0,01	0	0,01	0	0	0,01	0,02	0	0	0	0	0,03	0,02	0,03	0	0,03	0,02	0	0	0	0	0	0
makedonština	1,66	12,05	55,79	8,81	16,59	4,31	0,03	0,11	0,02	0,07	0,06	0,02	0,02	0,02	0	0,05	0,01	0	0,02	0,07	0,02	0,04	0,05	0,02	0,02	0	0,01	0,03	0,02	0,03	0,01	0,02	0,01	0
ruština	3,35	10,08	5,87	63,55	6,84	10,03	0,01	0	0,01	0,01	0	0,03	0,01	0,07	0,03	0,02	0	0,01	0,01	0	0	0,01	0	0	0	0,01	0,02	0	0,01	0	0	0	0	0,01
srbština	1,53	8,33	14,49	7,34	61,88	5,23	0,05	0,40	0,03	0,08	0,09	0,01	0	0,02	0,01	0,01	0,01	0,01	0	0,06	0,02	0	0,06	0,03	0,05	0,02	0,02	0,03	0,04	0,04	0,05	0,02	0,02	0,01
ukrajinština	4,26	4,50	3,32	9,01	4,73	73,94	0,01	0	0	0	0,01	0,03	0	0,07	0	0,04	0	0	0	0,01	0	0	0,02	0	0,01	0	0,01	0	0,01	0	0	0	0,01	0
čeština	0	0,01	0,01	0,06	0,01	0,02	62,31	3,82	2,18	10,44	3,87	0,65	0,49	0,85	0,30	1,01	0,41	0,54	0,71	1,40	1,15	0,96	0,88	0,49	1,07	0,54	0,90	1,46	0,73	1,36	0,80	0,38	0,16	0,02
chorvatština	0,01	0	0	0,02	0,02	0	2,84	57,14	2,07	2,99	11,08	0,99	0,84	0,88	0,55	1,07	0,63	0,71	0,85	2,22	1,09	0,86	1,67	0,84	1,49	0,94	0,59	1,88	1,27	2,11	1,14	0,87	0,32	0,01
polština	0,01	0,01	0	0,03	0,02	0,02	2,22	2,52	71,96	2,45	1,96	1,13	0,78	0,81	0,33	1,24	0,52	0,62	0,62	1,55	1,20	0,88	1,60	0,51	0,92	0,53	0,57	1,04	0,78	1,62	0,88	0,56	0,09	0,01
slovenština	0,01	0	0,04	0,05	0,02	0,02	11,84	5,16	3,22	50,45	5,99	1,25	0,68	0,92	0,38	1,25	0,62	0,84	1,07	1,83	1,22	1,13	1,72	0,91	1,11	0,67	0,87	1,86	1,06	1,91	1,12	0,60	0,17	0
slovinština	0,01	0	0,02	0,05	0	0,03	3,22	12,35	1,79	4,24	54,39	1,01	0,71	1,04	0,82	1,32	1,06	0,66	0,99	2,07	1,31	1,10	1,37	0,63	1,38	0,97	0,64	1,45	0,86	1,53	1,50	1,09	0,37	0,01
angličtina	0	0	0,01	0,06	0	0,03	0,60	0,62	0,77	0,75	0,62	59,18	1,99	2,56	0,73	2,01	1,39	1,36	3,75	4,99	4,18	2,37	3,35	1,92	1,00	0,82	0,55	0,91	0,37	1,57	0,54	0,40	0,56	0,03
dánština	0	0,01	0	0,05	0	0,01	0,59	0,81	0,24	0,48	0,90	1,73	51,97	3,23	2,30	2,44	13,34	6,46	1,44	1,64	1,62	0,60	1,30	0,66	1,27	1,09	0,77	0,79	0,52	1,22	1,19	1,07	0,23	0,02
holandština	0	0	0	0,07	0	0,08	0,48	0,70	0,59	0,48	0,30	2,82	2,34	65,75	0,84	4,69	1,79	1,23	1,71	2,32	1,99	1,18	2,02	1,31	1,20	1,36	0,48	0,74	0,39	0,95	0,87	0,85	0,45	0,01
islandština	0	0,01	0	0,07	0	0,02	0,22	0,39	0,13	0,32	0,25	0,53	1,42	0,70	84,90	1,55	1,83	1,40	0,34	0,54	0,46	0,54	0,25	0,11	0,66	0,59	0,73	0,51	0,23	0,26	0,32	0,51	0,19	0,01
němčina	0,05	0	0,02	0,03	0,01	0,02	0,46	0,52	0,62	0,42	0,52	1,49	1,62	3,45	0,88	76,13	1,58	1,47	1,03	0,97	0,92	0,96	1,13	0,35	0,95	0,46	0,35	0,55	0,45	0,64	0,90	0,64	0,38	0,02
norština	0,01	0	0	0,02	0	0,02	0,52	1,01	0,63	0,49	0,96	2,34	18,37	3,04	2,85	3,03	42,25	7,70	1,24	2,06	1,71	1,05	1,46	0,56	1,38	1,18	0,71	1,08	0,58	1,15	1,23	0,97	0,37	0,03
švédština	0	0	0	0,03	0	0,01	0,46	0,59	0,26	0,31	0,47	1,26	5,26	1,77	1,61	2,33	7,53	65,18	0,82	1,33	1,34	0,59	0,88	0,71	1,12	2,02	0,53	0,78	0,53	0,80	0,68	0,57	0,22	0
francouzština	0	0	0	0,06	0	0,02	0,50	0,46	0,32	0,45	0,44	2,84	1,20	1,66	0,34	1,64	0,81	0,62	61,20	5,69	5,91	3,53	3,87	2,44	1,02	0,74	0,47	0,68	0,60	1,21	0,63	0,41	0,23	0
italština	0,02	0	0,02	0,09	0,01	0,06	0,57	0,97	0,48	0,74	0,96	2,64	1,03	1,21	0,62	1,67	1,13	0,65	3,08	54,99	7,62	4,52	4,82	3,50	1,13	1,47	0,74	1,06	0,67	1,31	0,89	0,86	0,41	0,05
katalánština	0	0,01	0	0,02	0	0,05	0,48	0,94	0,34	0,78	0,45	2,02	0,56	1,58	0,51	1,09	0,78	0,66	5,24	6,86	52,67	6,84	4,84	6,95	0,52	0,74	0,52	0,56	0,58	1,54	0,94	0,52	0,38	0,02
portugalština	0	0	0,01	0,04	0	0,01	0,77	0,98	0,38	0,67	0,83	2,60	0,74	1,30	1,01	1,00	0,93	0,53	3,29	5,48	9,78	49,09	4,56	8,28	0,93	0,90	0,81	1,14	0,40	1,33	1,18	0,73	0,30	0
rumunština	0,01	0	0	0,05	0	0,04	0,76	1,20	0,56	0,78	0,86	2,60	0,97	1,56	0,48	1,53	0,79	0,79	2,35	5,96	5,59	2,70	59,61	2,42	1,03	1,00	0,64	1,08	0,73	1,80	1,10	0,64	0,33	0,03
španělština	0	0	0	0,01	0	0	0,40	0,77	0,43	0,58	0,60	2,48	0,77	1,28	0,62	1,61	0,64	0,67	3,25	6,41	10,75	8,49	4,37	48,64	0,65	0,66	0,33	0,95	0,98	1,18	1,36	0,80	0,29	0,03
estonština	0,02	0	0,02	0,02	0	0,02	0,42	0,64	0,31	0,47	0,80	0,79	1,04	1,33	0,85	1,26	0,95	0,98	0,67	1,72	0,69	0,61	0,98	0,37	72,43	5,44	0,68	1,72	1,05	1,45	1,12	0,99	0,15	0
finština	0,03	0	0	0,03	0	0,01	0,32	0,57	0,56	0,54	0,56	0,81	0,73	0,83	0,60	1,00	0,91	0,94	0,69	1,11	0,51	0,68	0,77	0,46	6,25	75,45	0,69	1,43	0,71	0,82	1,17	0,62	0,15	0,04
maďarština	0	0	0,01	0,08	0,02	0,02	0,58	0,45	0,43	0,74	0,48	0,70	0,83	0,75	0,76	1,10	0,73	0,53	0,71	0,77	0,78	0,82	1,24	0,36	0,80	0,57	82,01	0,73	0,23	0,80	1,05	0,63	0,28	0
litevština	0	0	0	0,02	0	0	0,87	1,19	0,69	1,23	1,29	0,68	0,92	0,74	0,89	0,84	0,96	0,63	0,62	1,26	1,53	0,61	1,61	0,68	1,49	1,23	0,38	69,82	4,73	2,56	1,29	1,12	0,10	0,01
lotyština	0	0	0,01	0,01	0,03	0,01	0,92	1,21	0,58	0,82	1,12	0,41	0,66	0,92	0,55	0,94	0,55	0,67	0,51	1,27	1,02	0,62	1,48	0,78	1,34	1,38	0,36	5,34	72,53	2,01	1,10	0,71	0,12	0,01
albánština	0	0	0,01	0,01	0	0,05	0,73	1,19	0,49	0,37	1,22	0,87	1,18	0,91	0,59	1,01	0,57	0,59	1,06	2,07	1,51	0,99	1,93	0,93	0,93	0,69	0,40	1,32	0,69	75,05	1,45	0,95	0,22	0,01
baskičtina	0,01	0	0	0,01	0	0	0,34	0,72	0,72	0,48	0,53	0,73	1,06	0,77	0,65	1,58	0,99	0,54	0,85	2,15	2,04	1,22	1,38	1,13	1,36	0,73	0,85	0,81	0,58	1,66	74,90	1,09	0,10	0,01
turečtina	0,01	0	0	0,04	0,01	0,01	0,56	0,89	0,28	0,40	0,55	0,59	0,76	0,91	0,44	0,94	0,82	0,45	0,43	0,57	0,80	0,63	1,02	0,63	1,49	0,99	0,59	0,60	0,50	1,24	1,20	81,55	0,07	0,02
vietnamština	0	0	0	0,01	0	0,04	0,13	0,13	0,06	0,12	0,12	0,49	0,15	0,24	0,11	0,61	0,14	0,17	0,27	0,35	0,32	0,24	0,33	0,19	0,10	0,13	0,15	0,09	0,08	0,12	0,15	0,03	94,88	0,04
řečtina	0	0	0,02	0,02	0	0,05	0,02	0,03	0,04	0,02	0,04	0,22	0,06	0,15	0,02	0,21	0,03	0,07	0,08	0,09	0,06	0,04	0,08	0,06	0,04	0,04	0,02	0,03	0,02	0,04	0,02	0,05	0,05	98,27

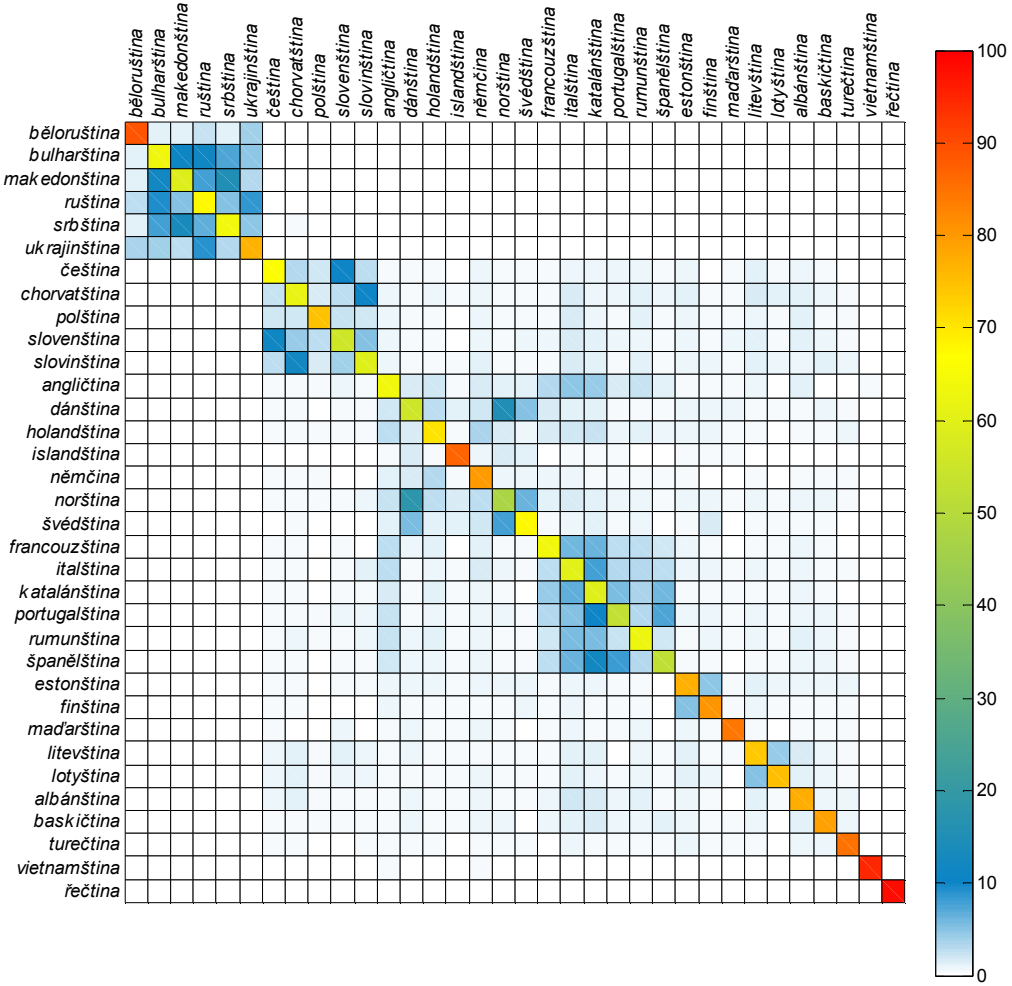
Tabulka D.4 Konfuzní matice 4-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.5 Konfuzní matice 5-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	88,76	1,53	1,20	2,73	1,30	4,01	0,03	0,02	0,06	0	0,05	0,02	0,01	0,02	0	0,02	0	0	0,03	0,01	0,01	0,01	0,02	0,03	0,01	0,02	0,04	0,01	0,02	0,01	0,01	0	0	0
bulharština	1,31	63,33	11,26	11,56	7,51	4,75	0,02	0	0	0,05	0,03	0,01	0	0,01	0	0	0,01	0,01	0	0	0,01	0	0,03	0,01	0,02	0	0,03	0,02	0	0,01	0	0	0	0
makedonština	1,37	11,75	59,25	7,88	15,49	3,46	0,02	0,11	0,02	0,04	0,10	0,02	0,03	0,03	0,01	0,05	0,01	0,02	0,02	0,08	0,01	0,04	0,06	0,01	0,01	0	0,01	0,02	0,01	0,03	0,01	0,02	0	0
ruština	2,82	9,51	5,34	67,93	5,35	8,75	0	0,01	0,01	0	0	0,03	0	0,06	0,03	0,05	0	0	0	0,04	0,01	0	0,01	0	0,01	0,01	0,01	0	0,01	0	0	0	0	0
srbština	1,39	7,84	13,66	6,95	64,23	4,74	0,05	0,44	0,03	0,09	0,08	0,01	0,01	0,01	0	0,01	0,01	0,01	0	0,04	0,03	0,01	0,05	0,02	0,04	0,01	0,01	0,02	0,06	0,06	0,04	0,02	0,01	0,01
ukrajinština	3,65	4,23	2,77	8,99	3,47	76,64	0	0	0,01	0	0,01	0,01	0	0,09	0	0,05	0	0	0	0,02	0	0	0,02	0	0	0	0,01	0	0,01	0	0	0	0,01	0
čeština	0	0,01	0,01	0,06	0,01	0,02	66,17	3,32	2,04	10,84	2,76	0,72	0,47	0,76	0,22	1,05	0,45	0,41	0,66	1,15	0,96	0,85	0,91	0,42	0,92	0,36	0,63	1,26	0,53	0,98	0,60	0,35	0,09	0
chorvatština	0,01	0	0	0,01	0,03	0	2,44	61,90	1,72	2,81	10,67	1,04	0,75	0,90	0,40	0,95	0,57	0,55	0,72	1,91	1,05	0,91	1,34	0,79	1,20	0,70	0,61	1,62	1,21	1,48	0,86	0,66	0,17	0,01
polština	0,01	0,01	0	0,01	0	0,02	2,34	2,15	74,66	2,53	1,99	0,90	0,72	0,50	0,34	1,11	0,56	0,49	0,60	1,59	0,90	0,73	1,19	0,55	0,84	0,45	0,42	1,15	0,62	1,29	0,78	0,46	0,06	0,02
slovenština	0,01	0,01	0,02	0,03	0,02	0,02	11,64	4,65	3,08	55,80	5,38	1,13	0,62	0,63	0,34	1,11	0,49	0,81	1,00	1,62	1,21	0,90	1,11	0,59	1,06	0,64	1,01	1,56	0,61	1,45	0,81	0,49	0,14	0
slovinština	0,01	0	0,02	0,05	0	0,03	2,92	11,75	1,68	4,09	59,64	0,84	0,66	0,73	0,57	1,27	0,76	0,60	0,78	1,75	1,20	0,67	1,25	0,52	1,17	0,97	0,55	1,35	0,81	1,08	1,22	0,86	0,18	0,01
angličtina	0	0	0,01	0,03	0	0,03	0,61	0,57	0,69	0,79	0,69	64,04	1,74	2,27	0,42	1,68	1,29	1,23	3,30	4,88	4,38	1,83	2,61	1,33	0,78	0,74	0,47	0,84	0,25	1,35	0,38	0,31	0,42	0,03
dánština	0	0,01	0	0,05	0	0,01	0,67	0,78	0,31	0,53	0,46	2,06	55,47	2,76	1,31	2,29	15,31	5,17	1,61	1,55	1,47	0,51	0,60	0,42	1,11	0,97	0,89	0,66	0,44	0,70	0,99	0,74	0,13	0,01
holandština	0	0	0	0,06	0,01	0,08	0,56	0,51	0,47	0,66	0,42	2,90	1,86	70,45	0,62	3,69	1,66	1,09	1,68	2,33	2,42	0,92	1,37	0,92	0,89	0,94	0,44	0,48	0,31	0,67	0,51	0,85	0,19	0,03
islandština	0	0,01	0	0,06	0	0,01	0,17	0,33	0,13	0,31	0,21	0,52	1,57	0,59	86,78	1,11	1,66	1,42	0,31	0,44	0,46	0,42	0,27	0,10	0,52	0,50	0,63	0,34	0,13	0,29	0,30	0,23	0,16	0,01
němčina	0,04	0	0,02	0,03	0,01	0,02	0,41	0,48	0,46	0,36	0,31	1,51	1,78	3,42	0,66	79,91	1,00	1,08	0,89	0,90	0,75	0,69	0,73	0,40	0,70	0,36	0,32	0,47	0,33	0,40	0,78	0,50	0,26	0,01
norština	0	0	0,02	0,02	0	0,02	0,64	0,77	0,50	0,60	0,93	2,48	18,05	2,78	1,91	2,81	47,42	6,41	1,41	1,88	1,48	1,08	1,01	0,46	1,00	1,09	0,71	0,95	0,54	0,93	1,11	0,75	0,23	0,01
švédština	0	0	0	0,03	0	0,01	0,39	0,65	0,28	0,29	0,34	1,19	5,72	1,47	1,51	2,26	8,06	67,71	0,75	0,99	1,30	0,56	0,79	0,36	0,66	1,60	0,39	0,51	0,52	0,58	0,62	0,30	0,15	0
francouzština	0	0	0	0,05	0	0,02	0,63	0,49	0,37	0,55	0,33	3,09	1,12	1,47	0,16	1,33	0,68	0,50	63,96	5,93	6,49	2,82	2,93	2,18	0,81	0,54	0,38	0,71	0,57	0,88	0,52	0,33	0,15	0
italština	0,02	0	0,02	0,08	0,01	0,06	0,42	0,57	0,38	0,66	1,18	2,97	0,76	1,15	0,50	1,69	1,01	0,65	3,04	60,21	8,04	3,51	3,41	2,95	0,91	1,05	0,64	0,89	0,51	0,93	0,74	0,76	0,24	0,03
katalánština	0	0,01	0	0,01	0	0,05	0,46	0,56	0,31	0,71	0,60	1,87	0,54	1,39	0,57	1,08	0,44	0,38	4,36	6,81	59,40	5,79	3,53	5,99	0,53	0,62	0,43	0,61	0,51	1,12	0,69	0,38	0,22	0,02
portugalština	0	0	0,01	0,04	0	0,01	0,68	0,70	0,41	0,81	0,95	2,65	0,76	0,98	0,76	0,90	0,87	0,41	3,19	5,30	10,67	52,47	3,31	7,70	0,90	0,81	0,67	0,98	0,55	0,99	0,84	0,49	0,17	0,01
rumunština	0,01	0	0	0,01	0	0,03	0,65	1,11	0,51	0,74	1,12	2,47	0,90	1,41	0,46	1,15	0,52	0,53	2,30	5,85	5,57	2,52	62,84	2,14	0,77	0,81	0,77	1,00	0,73	1,38	0,85	0,62	0,19	0,03
španělština	0	0	0	0	0	0	0,52	0,60	0,33	0,66	0,68	2,05	0,83	1,13	0,52	0,91	0,60	0,42	3,11	6,34	12,11	8,54	3,26	52,05	0,57	0,49	0,31	0,66	0,58	0,79	1,04	0,65	0,21	0,03
estonština	0,02	0	0	0,02	0	0,02	0,44	0,58	0,28	0,67	0,48	0,94	1,00	0,98	0,49	1,01	0,83	0,82	0,41	1,01	0,87	0,62	0,76	0,34	77,00	4,84	0,46	1,44	0,95	0,85	0,83	0,96	0,07	0
finština	0,03	0	0	0,02	0	0	0,52	0,43	0,36	0,39	0,34	0,87	0,52	0,57	0,40	0,76	0,75	0,80	0,47	0,80	0,73	0,44	0,65	0,31	5,10	80,65	0,62	1,17	0,53	0,58	0,61	0,46	0,10	0,01
maďarština	0	0	0	0,05	0,02	0,02	0,72	0,34	0,32	0,91	0,32	0,77	0,83	0,72	0,55	0,93	0,43	0,39	0,52	1,15	0,72	0,77	0,83	0,39	0,79	0,43	84,49	0,56	0,22	0,47	0,70	0,49	0,14	0
litevština	0	0	0	0,02	0	0	1,01	1,28	0,75	1,22	0,98	0,60	0,89	0,54	0,49	0,64	0,53	0,52	0,72	1,43	1,52	0,38	1,07	0,51	1,51	0,72	0,42	73,96	4,67	1,78	1,14	0,66	0,03	0
lotyština	0	0	0,01	0,01	0,02	0,01	0,94	1,32	0,57	0,85	1,01	0,46	0,41	0,61	0,44	0,79	0,62	0,58	0,54	1,30	1,13	0,69	0,99	0,62	1,52	0,89	0,36	5,21	75,08	1,48	0,87	0,59	0,06	0,01
albánština	0	0	0,01	0,01	0	0,05	0,80	1,18	0,40	0,60	0,66	0,75	1,14	0,71	0,45	0,82	0,53	0,42	0,87	2,27	1,79	1,11	1,45	0,70	1,01	0,42	0,27	1,37	0,74	77,55	0,97	0,79	0,14	0,01
baskičtina	0,01	0	0	0,01	0	0	0,50	0,50	0,59	0,45	0,41	0,77	1,08	0,67	0,47	1,16	0,91	0,26	0,85	1,51	1,94	1,15	1,02	1,40	0,89	0,50	0,49	0,79	0,38	1,21	79,08	0,92	0,04	0,03
turečtina	0,01	0	0	0,04	0,01	0,01	0,47	0,60	0,24	0,33	0,46	0,49	0,66	0,76	0,35	0,74	0,48	0,44	0,38	0,88	0,62	0,58	0,69	0,62	1,17	0,57	0,51	0,59	0,27	1,12	0,84	85,01	0,04	0,01
vietnamština	0	0	0	0,01	0	0,04	0,17	0,11	0,10	0,11	0,11	0,58	0,16	0,31	0,04	0,55	0,14	0,17	0,20	0,35	0,33	0,21	0,36	0,23	0,08	0,12	0,13	0,09	0,08	0,11	0,15	0,04	94,89	0,02
řečtina	0	0	0,01	0,02	0	0,05	0,02	0,04	0,01	0,03	0,01	0,20	0,08	0,23	0,03	0,15	0,04	0,04	0,08	0,09	0,07	0,07	0,08	0,04	0,04	0	0,03	0,01	0,02	0,09	0,02	0,03	0,03	98,33

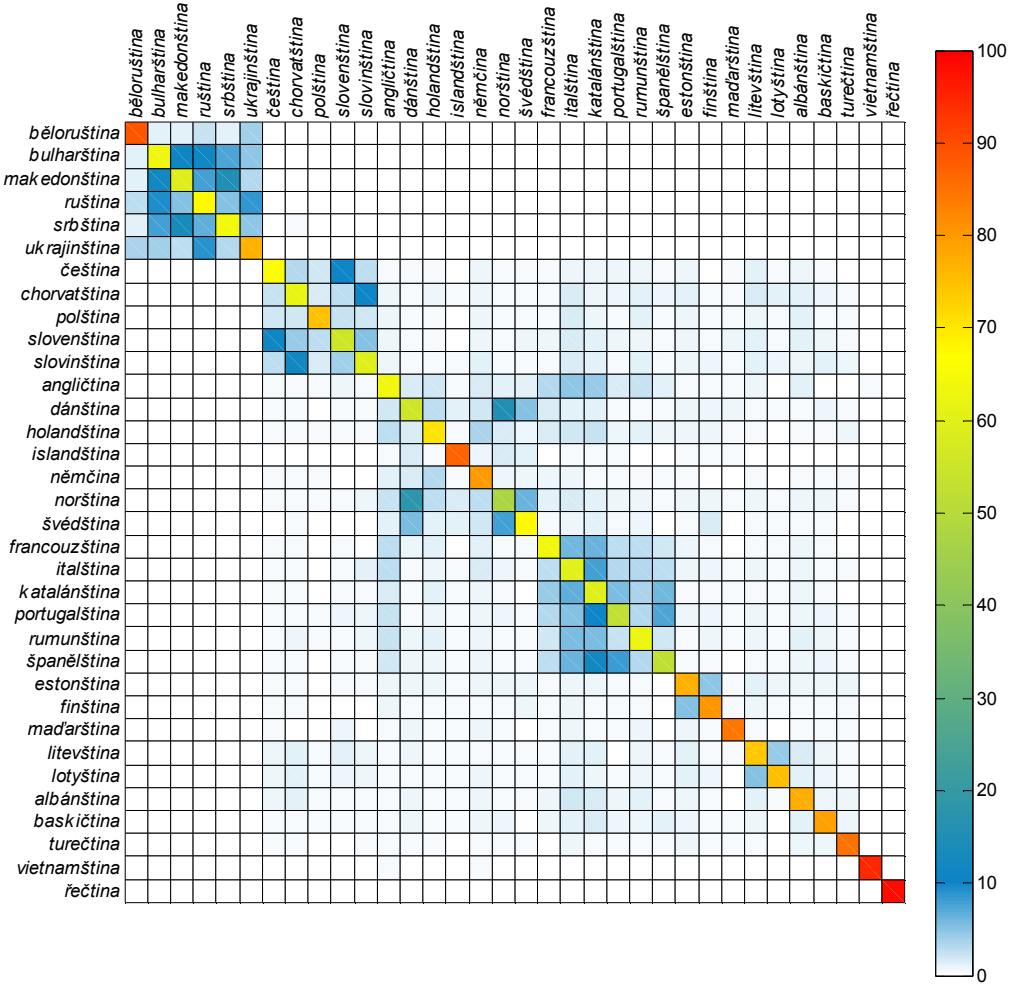
Tabulka D.5 Konfuzní matice 5-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.6 Konfuzní matice 6-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	88,76	1,53	1,20	2,73	1,30	4,01	0,03	0,02	0,06	0	0,05	0,02	0,01	0,02	0	0,02	0	0	0,03	0,01	0,01	0,01	0,02	0,03	0,01	0,02	0,04	0,01	0,02	0,01	0,01	0	0	0
bulharština	1,31	63,33	11,26	11,56	7,51	4,75	0,02	0	0	0,05	0,03	0,01	0	0,01	0	0	0,01	0,01	0	0	0,01	0	0,03	0,01	0,02	0	0,03	0,02	0	0,01	0	0	0	0
makedonština	1,37	11,75	59,25	7,88	15,49	3,46	0,02	0,11	0,02	0,04	0,10	0,02	0,03	0,03	0,01	0,05	0,01	0,02	0,02	0,08	0,01	0,04	0,06	0,01	0,01	0	0,01	0,02	0,01	0,03	0,01	0,02	0	0
ruština	2,82	9,51	5,34	67,93	5,35	8,75	0	0,01	0,01	0	0	0,03	0	0,06	0,03	0,05	0	0	0	0,04	0,01	0	0,01	0	0,01	0,01	0,01	0	0,01	0	0	0	0	0
srbština	1,39	7,84	13,66	6,95	64,23	4,74	0,05	0,44	0,03	0,09	0,08	0,01	0,01	0,01	0	0,01	0,01	0,01	0	0,04	0,03	0,01	0,05	0,02	0,04	0,01	0,01	0,02	0,06	0,06	0,04	0,02	0,01	0,01
ukrajinština	3,65	4,23	2,77	8,99	3,47	76,64	0	0	0,01	0	0,01	0,01	0	0,09	0	0,05	0	0	0	0,02	0	0	0,02	0	0	0	0,01	0	0,01	0	0	0	0,01	0
čeština	0	0,01	0,01	0,06	0,01	0,02	66,17	3,32	2,04	10,84	2,76	0,72	0,47	0,76	0,22	1,05	0,45	0,41	0,66	1,15	0,96	0,85	0,91	0,42	0,92	0,36	0,63	1,26	0,53	0,98	0,60	0,35	0,09	0
chorvatština	0,01	0	0	0,01	0,03	0	2,44	61,90	1,72	2,81	10,67	1,04	0,75	0,90	0,40	0,95	0,57	0,55	0,72	1,91	1,05	0,91	1,34	0,79	1,20	0,70	0,61	1,62	1,21	1,48	0,86	0,66	0,17	0,01
polština	0,01	0,01	0	0,01	0	0,02	2,34	2,15	74,66	2,53	1,99	0,90	0,72	0,50	0,34	1,11	0,56	0,49	0,60	1,59	0,90	0,73	1,19	0,55	0,84	0,45	0,42	1,15	0,62	1,29	0,78	0,46	0,06	0,02
slovenština	0,01	0,01	0,02	0,03	0,02	0,02	11,64	4,65	3,08	55,80	5,38	1,13	0,62	0,63	0,34	1,11	0,49	0,81	1,00	1,62	1,21	0,90	1,11	0,59	1,06	0,64	1,01	1,56	0,61	1,45	0,81	0,49	0,14	0
slovinština	0,01	0	0,02	0,05	0	0,03	2,92	11,75	1,68	4,09	59,64	0,84	0,66	0,73	0,57	1,27	0,76	0,60	0,78	1,75	1,20	0,67	1,25	0,52	1,17	0,97	0,55	1,35	0,81	1,08	1,22	0,86	0,18	0,01
angličtina	0	0	0,01	0,03	0	0,03	0,61	0,57	0,69	0,79	0,69	64,04	1,74	2,27	0,42	1,68	1,29	1,23	3,30	4,88	4,38	1,83	2,61	1,33	0,78	0,74	0,47	0,84	0,25	1,35	0,38	0,31	0,42	0,03
dánština	0	0,01	0	0,05	0	0,01	0,67	0,78	0,31	0,53	0,46	2,06	55,47	2,76	1,31	2,29	15,31	5,17	1,61	1,55	1,47	0,51	0,60	0,42	1,11	0,97	0,89	0,66	0,44	0,70	0,99	0,74	0,13	0,01
holandština	0	0	0	0,06	0,01	0,08	0,56	0,51	0,47	0,66	0,42	2,90	1,86	70,45	0,62	3,69	1,66	1,09	1,68	2,33	2,42	0,92	1,37	0,92	0,89	0,94	0,44	0,48	0,31	0,67	0,51	0,85	0,19	0,03
islandština	0	0,01	0	0,06	0	0,01	0,17	0,33	0,13	0,31	0,21	0,52	1,57	0,59	86,78	1,11	1,66	1,42	0,31	0,44	0,46	0,42	0,27	0,10	0,52	0,50	0,63	0,34	0,13	0,29	0,30	0,23	0,16	0,01
němčina	0,04	0	0,02	0,03	0,01	0,02	0,41	0,48	0,46	0,36	0,31	1,51	1,78	3,42	0,66	79,91	1,00	1,08	0,89	0,90	0,75	0,69	0,73	0,40	0,70	0,36	0,32	0,47	0,33	0,40	0,78	0,50	0,26	0,01
norština	0	0	0,02	0,02	0	0,02	0,64	0,77	0,50	0,60	0,93	2,48	18,05	2,78	1,91	2,81	47,42	6,41	1,41	1,88	1,48	1,08	1,01	0,46	1,00	1,09	0,71	0,95	0,54	0,93	1,11	0,75	0,23	0,01
švédština	0	0	0	0,03	0	0,01	0,39	0,65	0,28	0,29	0,34	1,19	5,72	1,47	1,51	2,26	8,06	67,71	0,75	0,99	1,30	0,56	0,79	0,36	0,66	1,60	0,39	0,51	0,52	0,58	0,62	0,30	0,15	0
francouzština	0	0	0	0,05	0	0,02	0,63	0,49	0,37	0,55	0,33	3,09	1,12	1,47	0,16	1,33	0,68	0,50	63,96	5,93	6,49	2,82	2,93	2,18	0,81	0,54	0,38	0,71	0,57	0,88	0,52	0,33	0,15	0
italština	0,02	0	0,02	0,08	0,01	0,06	0,42	0,57	0,38	0,66	1,18	2,97	0,76	1,15	0,50	1,69	1,01	0,65	3,04	60,21	8,04	3,51	3,41	2,95	0,91	1,05	0,64	0,89	0,51	0,93	0,74	0,76	0,24	0,03
katalánština	0	0,01	0	0,01	0	0,05	0,46	0,56	0,31	0,71	0,60	1,87	0,54	1,39	0,57	1,08	0,44	0,38	4,36	6,81	59,40	5,79	3,53	5,99	0,53	0,62	0,43	0,61	0,51	1,12	0,69	0,38	0,22	0,02
portugalština	0	0	0,01	0,04	0	0,01	0,68	0,70	0,41	0,81	0,95	2,65	0,76	0,98	0,76	0,90	0,87	0,41	3,19	5,30	10,67	52,47	3,31	7,70	0,90	0,81	0,67	0,98	0,55	0,99	0,84	0,49	0,17	0,01
rumunština	0,01	0	0	0,01	0	0,03	0,65	1,11	0,51	0,74	1,12	2,47	0,90	1,41	0,46	1,15	0,52	0,53	2,30	5,85	5,57	2,52	62,84	2,14	0,77	0,81	0,77	1,00	0,73	1,38	0,85	0,62	0,19	0,03
španělština	0	0	0	0	0	0	0,52	0,60	0,33	0,66	0,68	2,05	0,83	1,13	0,52	0,91	0,60	0,42	3,11	6,34	12,11	8,54	3,26	52,05	0,57	0,49	0,31	0,66	0,58	0,79	1,04	0,65	0,21	0,03
estonština	0,02	0	0	0,02	0	0,02	0,44	0,58	0,28	0,67	0,48	0,94	1,00	0,98	0,49	1,01	0,83	0,82	0,41	1,01	0,87	0,62	0,76	0,34	77,00	4,84	0,46	1,44	0,95	0,85	0,83	0,96	0,07	0
finština	0,03	0	0	0,02	0	0	0,52	0,43	0,36	0,39	0,34	0,87	0,52	0,57	0,40	0,76	0,75	0,80	0,47	0,80	0,73	0,44	0,65	0,31	5,10	80,65	0,62	1,17	0,53	0,58	0,61	0,46	0,10	0,01
maďarština	0	0	0	0,05	0,02	0,02	0,72	0,34	0,32	0,91	0,32	0,77	0,83	0,72	0,55	0,93	0,43	0,39	0,52	1,15	0,72	0,77	0,83	0,39	0,79	0,43	84,49	0,56	0,22	0,47	0,70	0,49	0,14	0
litevština	0	0	0	0,02	0	0	1,01	1,28	0,75	1,22	0,98	0,60	0,89	0,54	0,49	0,64	0,53	0,52	0,72	1,43	1,52	0,38	1,07	0,51	1,51	0,72	0,42	73,96	4,67	1,78	1,14	0,66	0,03	0
lotyština	0	0	0,01	0,01	0,02	0,01	0,94	1,32	0,57	0,85	1,01	0,46	0,41	0,61	0,44	0,79	0,62	0,58	0,54	1,30	1,13	0,69	0,99	0,62	1,52	0,89	0,36	5,21	75,08	1,48	0,87	0,59	0,06	0,01
albánština	0	0	0,01	0,01	0	0,05	0,80	1,18	0,40	0,60	0,66	0,75	1,14	0,71	0,45	0,82	0,53	0,42	0,87	2,27	1,79	1,11	1,45	0,70	1,01	0,42	0,27	1,37	0,74	77,55	0,97	0,79	0,14	0,01
baskičtina	0,01	0	0	0,01	0	0	0,50	0,50	0,59	0,45	0,41	0,77	1,08	0,67	0,47	1,16	0,91	0,26	0,85	1,51	1,94	1,15	1,02	1,40	0,89	0,50	0,49	0,79	0,38	1,21	79,08	0,92	0,04	0,03
turečtina	0,01	0	0	0,04	0,01	0,01	0,47	0,60	0,24	0,33	0,46	0,49	0,66	0,76	0,35	0,74	0,48	0,44	0,38	0,88	0,62	0,58	0,69	0,62	1,17	0,57	0,51	0,59	0,27	1,12	0,84	85,01	0,04	0,01
vietnamština	0	0	0	0,01	0	0,04	0,17	0,11	0,10	0,11	0,11	0,58	0,16	0,31	0,04	0,55	0,14	0,17	0,20	0,35	0,33	0,21	0,36	0,23	0,08	0,12	0,13	0,09	0,08	0,11	0,15	0,04	94,89	0,02
řečtina	0	0	0,01	0,02	0	0,05	0,02	0,04	0,01	0,03	0,01	0,20	0,08	0,23	0,03	0,15	0,04	0,04	0,08	0,09	0,07	0,07	0,08	0,04	0,04	0	0,03	0,01	0,02	0,09	0,02	0,03	0,03	98,33

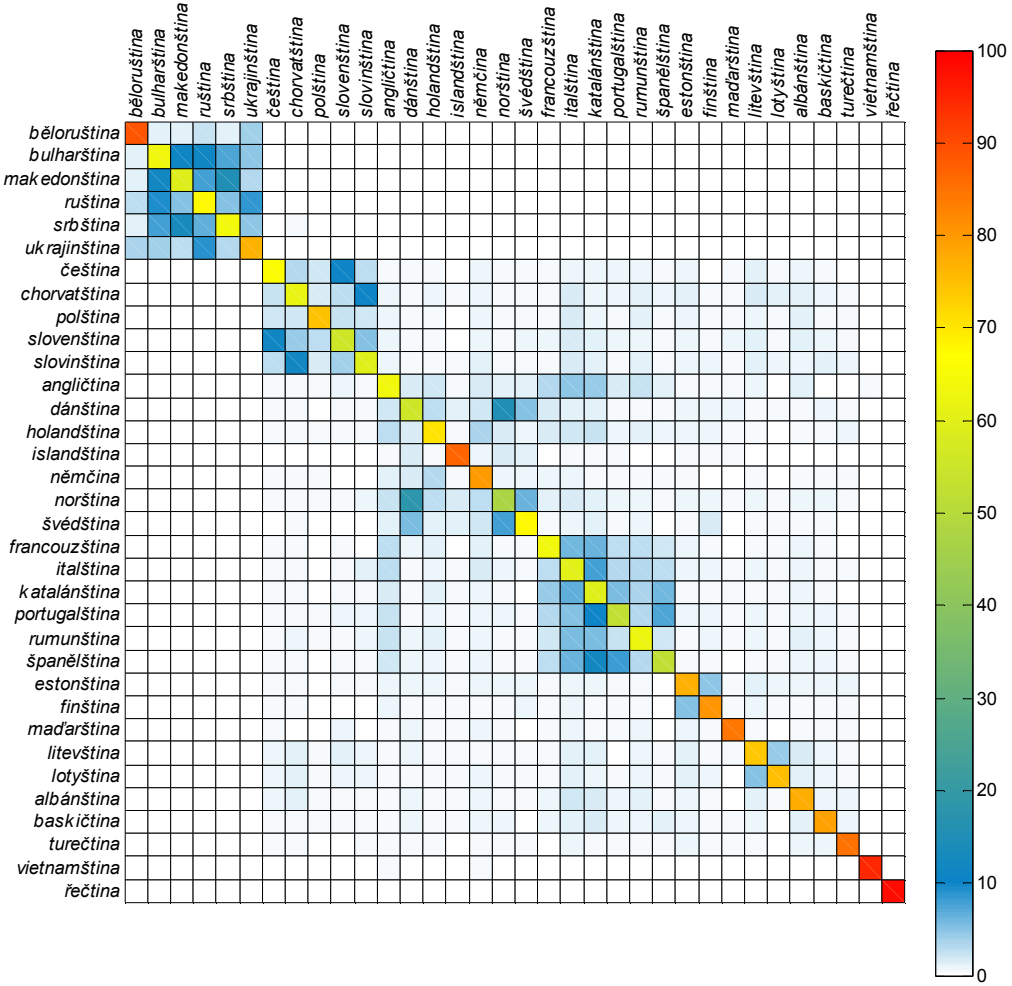
Tabulka D.6 Konfuzní matice 6-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.7 Konfuzní matice 7-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	88,76	1,53	1,20	2,73	1,30	4,01	0,03	0,02	0,06	0	0,05	0,02	0,01	0,02	0	0,02	0	0	0,03	0,01	0,01	0,01	0,02	0,03	0,01	0,02	0,04	0,01	0,02	0,01	0,01	0	0	0
bulharština	1,31	63,33	11,26	11,56	7,51	4,75	0,02	0	0	0,05	0,03	0,01	0	0,01	0	0	0,01	0,01	0	0	0,01	0	0,03	0,01	0,02	0	0,03	0,02	0	0,01	0	0	0	0
makedonština	1,37	11,75	59,25	7,88	15,49	3,46	0,02	0,11	0,02	0,04	0,10	0,02	0,03	0,03	0,01	0,05	0,01	0,02	0,02	0,08	0,01	0,04	0,06	0,01	0,01	0	0,01	0,02	0,01	0,03	0,01	0,02	0	0
ruština	2,82	9,51	5,34	67,93	5,35	8,75	0	0,01	0,01	0	0	0,03	0	0,06	0,03	0,05	0	0	0	0,04	0,01	0	0,01	0	0,01	0,01	0,01	0	0,01	0	0	0	0	0
srbština	1,39	7,84	13,66	6,95	64,23	4,74	0,05	0,44	0,03	0,09	0,08	0,01	0,01	0,01	0	0,01	0,01	0,01	0	0,04	0,03	0,01	0,05	0,02	0,04	0,01	0,01	0,02	0,06	0,06	0,04	0,02	0,01	0,01
ukrajinština	3,65	4,23	2,77	8,99	3,47	76,64	0	0	0,01	0	0,01	0,01	0	0,09	0	0,05	0	0	0	0,02	0	0	0,02	0	0	0	0,01	0	0,01	0	0	0	0,01	0
čeština	0	0,01	0,01	0,06	0,01	0,02	66,17	3,32	2,04	10,84	2,76	0,72	0,47	0,76	0,22	1,05	0,45	0,41	0,66	1,15	0,96	0,85	0,91	0,42	0,92	0,36	0,63	1,26	0,53	0,98	0,60	0,35	0,09	0
chorvatština	0,01	0	0	0,01	0,03	0	2,44	61,90	1,72	2,81	10,67	1,04	0,75	0,90	0,40	0,95	0,57	0,55	0,72	1,91	1,05	0,91	1,34	0,79	1,20	0,70	0,61	1,62	1,21	1,48	0,86	0,66	0,17	0,01
polština	0,01	0,01	0	0,01	0	0,02	2,34	2,15	74,66	2,53	1,99	0,90	0,72	0,50	0,34	1,11	0,56	0,49	0,60	1,59	0,90	0,73	1,19	0,55	0,84	0,45	0,42	1,15	0,62	1,29	0,78	0,46	0,06	0,02
slovenština	0,01	0,01	0,02	0,03	0,02	0,02	11,64	4,65	3,08	55,80	5,38	1,13	0,62	0,63	0,34	1,11	0,49	0,81	1,00	1,62	1,21	0,90	1,11	0,59	1,06	0,64	1,01	1,56	0,61	1,45	0,81	0,49	0,14	0
slovinština	0,01	0	0,02	0,05	0	0,03	2,92	11,75	1,68	4,09	59,64	0,84	0,66	0,73	0,57	1,27	0,76	0,60	0,78	1,75	1,20	0,67	1,25	0,52	1,17	0,97	0,55	1,35	0,81	1,08	1,22	0,86	0,18	0,01
angličtina	0	0	0,01	0,03	0	0,03	0,61	0,57	0,69	0,79	0,69	64,04	1,74	2,27	0,42	1,68	1,29	1,23	3,30	4,88	4,38	1,83	2,61	1,33	0,78	0,74	0,47	0,84	0,25	1,35	0,38	0,31	0,42	0,03
dánština	0	0,01	0	0,05	0	0,01	0,67	0,78	0,31	0,53	0,46	2,06	55,47	2,76	1,31	2,29	15,31	5,17	1,61	1,55	1,47	0,51	0,60	0,42	1,11	0,97	0,89	0,66	0,44	0,70	0,99	0,74	0,13	0,01
holandština	0	0	0	0,06	0,01	0,08	0,56	0,51	0,47	0,66	0,42	2,90	1,86	70,45	0,62	3,69	1,66	1,09	1,68	2,33	2,42	0,92	1,37	0,92	0,89	0,94	0,44	0,48	0,31	0,67	0,51	0,85	0,19	0,03
islandština	0	0,01	0	0,06	0	0,01	0,17	0,33	0,13	0,31	0,21	0,52	1,57	0,59	86,78	1,11	1,66	1,42	0,31	0,44	0,46	0,42	0,27	0,10	0,52	0,50	0,63	0,34	0,13	0,29	0,30	0,23	0,16	0,01
němčina	0,04	0	0,02	0,03	0,01	0,02	0,41	0,48	0,46	0,36	0,31	1,51	1,78	3,42	0,66	79,91	1,00	1,08	0,89	0,90	0,75	0,69	0,73	0,40	0,70	0,36	0,32	0,47	0,33	0,40	0,78	0,50	0,26	0,01
norština	0	0	0,02	0,02	0	0,02	0,64	0,77	0,50	0,60	0,93	2,48	18,05	2,78	1,91	2,81	47,42	6,41	1,41	1,88	1,48	1,08	1,01	0,46	1,00	1,09	0,71	0,95	0,54	0,93	1,11	0,75	0,23	0,01
švédština	0	0	0	0,03	0	0,01	0,39	0,65	0,28	0,29	0,34	1,19	5,72	1,47	1,51	2,26	8,06	67,71	0,75	0,99	1,30	0,56	0,79	0,36	0,66	1,60	0,39	0,51	0,52	0,58	0,62	0,30	0,15	0
francouzština	0	0	0	0,05	0	0,02	0,63	0,49	0,37	0,55	0,33	3,09	1,12	1,47	0,16	1,33	0,68	0,50	63,96	5,93	6,49	2,82	2,93	2,18	0,81	0,54	0,38	0,71	0,57	0,88	0,52	0,33	0,15	0
italština	0,02	0	0,02	0,08	0,01	0,06	0,42	0,57	0,38	0,66	1,18	2,97	0,76	1,15	0,50	1,69	1,01	0,65	3,04	60,21	8,04	3,51	3,41	2,95	0,91	1,05	0,64	0,89	0,51	0,93	0,74	0,76	0,24	0,03
katalánština	0	0,01	0	0,01	0	0,05	0,46	0,56	0,31	0,71	0,60	1,87	0,54	1,39	0,57	1,08	0,44	0,38	4,36	6,81	59,40	5,79	3,53	5,99	0,53	0,62	0,43	0,61	0,51	1,12	0,69	0,38	0,22	0,02
portugalština	0	0	0,01	0,04	0	0,01	0,68	0,70	0,41	0,81	0,95	2,65	0,76	0,98	0,76	0,90	0,87	0,41	3,19	5,30	10,67	52,47	3,31	7,70	0,90	0,81	0,67	0,98	0,55	0,99	0,84	0,49	0,17	0,01
rumunština	0,01	0	0	0,01	0	0,03	0,65	1,11	0,51	0,74	1,12	2,47	0,90	1,41	0,46	1,15	0,52	0,53	2,30	5,85	5,57	2,52	62,84	2,14	0,77	0,81	0,77	1,00	0,73	1,38	0,85	0,62	0,19	0,03
španělština	0	0	0	0	0	0	0,52	0,60	0,33	0,66	0,68	2,05	0,83	1,13	0,52	0,91	0,60	0,42	3,11	6,34	12,11	8,54	3,26	52,05	0,57	0,49	0,31	0,66	0,58	0,79	1,04	0,65	0,21	0,03
estonština	0,02	0	0	0,02	0	0,02	0,44	0,58	0,28	0,67	0,48	0,94	1,00	0,98	0,49	1,01	0,83	0,82	0,41	1,01	0,87	0,62	0,76	0,34	77,00	4,84	0,46	1,44	0,95	0,85	0,83	0,96	0,07	0
finština	0,03	0	0	0,02	0	0	0,52	0,43	0,36	0,39	0,34	0,87	0,52	0,57	0,40	0,76	0,75	0,80	0,47	0,80	0,73	0,44	0,65	0,31	5,10	80,65	0,62	1,17	0,53	0,58	0,61	0,46	0,10	0,01
maďarština	0	0	0	0,05	0,02	0,02	0,72	0,34	0,32	0,91	0,32	0,77	0,83	0,72	0,55	0,93	0,43	0,39	0,52	1,15	0,72	0,77	0,83	0,39	0,79	0,43	84,49	0,56	0,22	0,47	0,70	0,49	0,14	0
litevština	0	0	0	0,02	0	0	1,01	1,28	0,75	1,22	0,98	0,60	0,89	0,54	0,49	0,64	0,53	0,52	0,72	1,43	1,52	0,38	1,07	0,51	1,51	0,72	0,42	73,96	4,67	1,78	1,14	0,66	0,03	0
lotyština	0	0	0,01	0,01	0,02	0,01	0,94	1,32	0,57	0,85	1,01	0,46	0,41	0,61	0,44	0,79	0,62	0,58	0,54	1,30	1,13	0,69	0,99	0,62	1,52	0,89	0,36	5,21	75,08	1,48	0,87	0,59	0,06	0,01
albánština	0	0	0,01	0,01	0	0,05	0,80	1,18	0,40	0,60	0,66	0,75	1,14	0,71	0,45	0,82	0,53	0,42	0,87	2,27	1,79	1,11	1,45	0,70	1,01	0,42	0,27	1,37	0,74	77,55	0,97	0,79	0,14	0,01
baskičtina	0,01	0	0	0,01	0	0	0,50	0,50	0,59	0,45	0,41	0,77	1,08	0,67	0,47	1,16	0,91	0,26	0,85	1,51	1,94	1,15	1,02	1,40	0,89	0,50	0,49	0,79	0,38	1,21	79,08	0,92	0,04	0,03
turečtina	0,01	0	0	0,04	0,01	0,01	0,47	0,60	0,24	0,33	0,46	0,49	0,66	0,76	0,35	0,74	0,48	0,44	0,38	0,88	0,62	0,58	0,69	0,62	1,17	0,57	0,51	0,59	0,27	1,12	0,84	85,01	0,04	0,01
vietnamština	0	0	0	0,01	0	0,04	0,17	0,11	0,10	0,11	0,11	0,58	0,16	0,31	0,04	0,55	0,14	0,17	0,20	0,35	0,33	0,21	0,36	0,23	0,08	0,12	0,13	0,09	0,08	0,11	0,15	0,04	94,89	0,02
řečtina	0	0	0,01	0,02	0	0,05	0,02	0,04	0,01	0,03	0,01	0,20	0,08	0,23	0,03	0,15	0,04	0,04	0,08	0,09	0,07	0,07	0,08	0,04	0,04	0	0,03	0,01	0,02	0,09	0,02	0,03	0,03	98,33

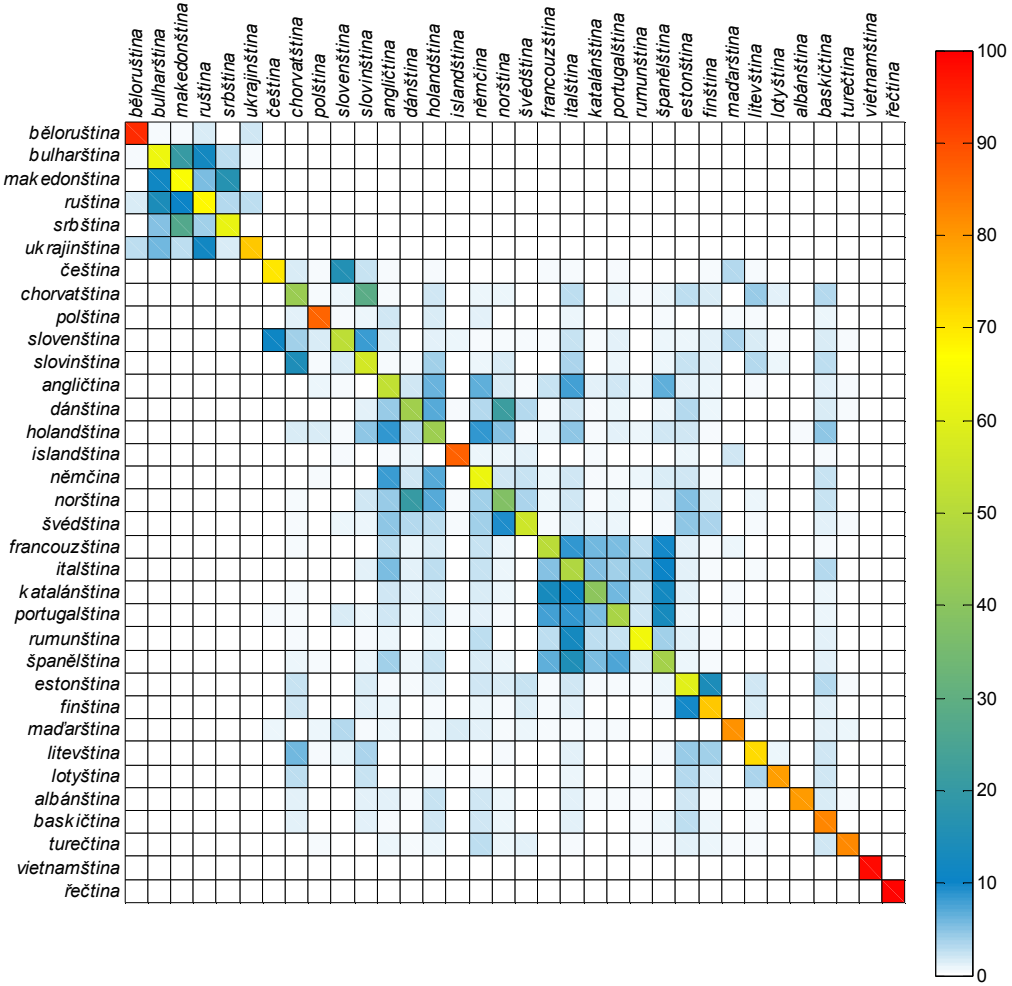
Tabulka D.7 Konfuzní matice 7-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.8 Konfuzní matice 8-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	88,76	1,53	1,20	2,73	1,30	4,01	0,03	0,02	0,06	0	0,05	0,02	0,01	0,02	0	0,02	0	0	0,03	0,01	0,01	0,01	0,02	0,03	0,01	0,02	0,04	0,01	0,02	0,01	0,01	0	0	0
bulharština	1,31	63,33	11,26	11,56	7,51	4,75	0,02	0	0	0,05	0,03	0,01	0	0,01	0	0	0,01	0,01	0	0	0,01	0	0,03	0,01	0,02	0	0,03	0,02	0	0,01	0	0	0	0
makedonština	1,37	11,75	59,25	7,88	15,49	3,46	0,02	0,11	0,02	0,04	0,10	0,02	0,03	0,03	0,01	0,05	0,01	0,02	0,02	0,08	0,01	0,04	0,06	0,01	0,01	0	0,01	0,02	0,01	0,03	0,01	0,02	0	0
ruština	2,82	9,51	5,34	67,93	5,35	8,75	0	0,01	0,01	0	0	0,03	0	0,06	0,03	0,05	0	0	0	0,04	0,01	0	0,01	0	0,01	0,01	0,01	0	0,01	0	0	0	0	0
srbština	1,39	7,84	13,66	6,95	64,23	4,74	0,05	0,44	0,03	0,09	0,08	0,01	0,01	0,01	0	0,01	0,01	0,01	0	0,04	0,03	0,01	0,05	0,02	0,04	0,01	0,01	0,02	0,06	0,06	0,04	0,02	0,01	0,01
ukrajinština	3,65	4,23	2,77	8,99	3,47	76,64	0	0	0,01	0	0,01	0,01	0	0,09	0	0,05	0	0	0	0,02	0	0	0,02	0	0	0	0,01	0	0,01	0	0	0	0,01	0
čeština	0	0,01	0,01	0,06	0,01	0,02	66,17	3,32	2,04	10,84	2,76	0,72	0,47	0,76	0,22	1,05	0,45	0,41	0,66	1,15	0,96	0,85	0,91	0,42	0,92	0,36	0,63	1,26	0,53	0,98	0,60	0,35	0,09	0
chorvatština	0,01	0	0	0,01	0,03	0	2,44	61,90	1,72	2,81	10,67	1,04	0,75	0,90	0,40	0,95	0,57	0,55	0,72	1,91	1,05	0,91	1,34	0,79	1,20	0,70	0,61	1,62	1,21	1,48	0,86	0,66	0,17	0,01
polština	0,01	0,01	0	0,01	0	0,02	2,34	2,15	74,66	2,53	1,99	0,90	0,72	0,50	0,34	1,11	0,56	0,49	0,60	1,59	0,90	0,73	1,19	0,55	0,84	0,45	0,42	1,15	0,62	1,29	0,78	0,46	0,06	0,02
slovenština	0,01	0,01	0,02	0,03	0,02	0,02	11,64	4,65	3,08	55,80	5,38	1,13	0,62	0,63	0,34	1,11	0,49	0,81	1,00	1,62	1,21	0,90	1,11	0,59	1,06	0,64	1,01	1,56	0,61	1,45	0,81	0,49	0,14	0
slovinština	0,01	0	0,02	0,05	0	0,03	2,92	11,75	1,68	4,09	59,64	0,84	0,66	0,73	0,57	1,27	0,76	0,60	0,78	1,75	1,20	0,67	1,25	0,52	1,17	0,97	0,55	1,35	0,81	1,08	1,22	0,86	0,18	0,01
angličtina	0	0	0,01	0,03	0	0,03	0,61	0,57	0,69	0,79	0,69	64,04	1,74	2,27	0,42	1,68	1,29	1,23	3,30	4,88	4,38	1,83	2,61	1,33	0,78	0,74	0,47	0,84	0,25	1,35	0,38	0,31	0,42	0,03
dánština	0	0,01	0	0,05	0	0,01	0,67	0,78	0,31	0,53	0,46	2,06	55,47	2,76	1,31	2,29	15,31	5,17	1,61	1,55	1,47	0,51	0,60	0,42	1,11	0,97	0,89	0,66	0,44	0,70	0,99	0,74	0,13	0,01
holandština	0	0	0	0,06	0,01	0,08	0,56	0,51	0,47	0,66	0,42	2,90	1,86	70,45	0,62	3,69	1,66	1,09	1,68	2,33	2,42	0,92	1,37	0,92	0,89	0,94	0,44	0,48	0,31	0,67	0,51	0,85	0,19	0,03
islandština	0	0,01	0	0,06	0	0,01	0,17	0,33	0,13	0,31	0,21	0,52	1,57	0,59	86,78	1,11	1,66	1,42	0,31	0,44	0,46	0,42	0,27	0,10	0,52	0,50	0,63	0,34	0,13	0,29	0,30	0,23	0,16	0,01
němčina	0,04	0	0,02	0,03	0,01	0,02	0,41	0,48	0,46	0,36	0,31	1,51	1,78	3,42	0,66	79,91	1,00	1,08	0,89	0,90	0,75	0,69	0,73	0,40	0,70	0,36	0,32	0,47	0,33	0,40	0,78	0,50	0,26	0,01
norština	0	0	0,02	0,02	0	0,02	0,64	0,77	0,50	0,60	0,93	2,48	18,05	2,78	1,91	2,81	47,42	6,41	1,41	1,88	1,48	1,08	1,01	0,46	1,00	1,09	0,71	0,95	0,54	0,93	1,11	0,75	0,23	0,01
švédština	0	0	0	0,03	0	0,01	0,39	0,65	0,28	0,29	0,34	1,19	5,72	1,47	1,51	2,26	8,06	67,71	0,75	0,99	1,30	0,56	0,79	0,36	0,66	1,60	0,39	0,51	0,52	0,58	0,62	0,30	0,15	0
francouzština	0	0	0	0,05	0	0,02	0,63	0,49	0,37	0,55	0,33	3,09	1,12	1,47	0,16	1,33	0,68	0,50	63,96	5,93	6,49	2,82	2,93	2,18	0,81	0,54	0,38	0,71	0,57	0,88	0,52	0,33	0,15	0
italština	0,02	0	0,02	0,08	0,01	0,06	0,42	0,57	0,38	0,66	1,18	2,97	0,76	1,15	0,50	1,69	1,01	0,65	3,04	60,21	8,04	3,51	3,41	2,95	0,91	1,05	0,64	0,89	0,51	0,93	0,74	0,76	0,24	0,03
katalánština	0	0,01	0	0,01	0	0,05	0,46	0,56	0,31	0,71	0,60	1,87	0,54	1,39	0,57	1,08	0,44	0,38	4,36	6,81	59,40	5,79	3,53	5,99	0,53	0,62	0,43	0,61	0,51	1,12	0,69	0,38	0,22	0,02
portugalština	0	0	0,01	0,04	0	0,01	0,68	0,70	0,41	0,81	0,95	2,65	0,76	0,98	0,76	0,90	0,87	0,41	3,19	5,30	10,67	52,47	3,31	7,70	0,90	0,81	0,67	0,98	0,55	0,99	0,84	0,49	0,17	0,01
rumunština	0,01	0	0	0,01	0	0,03	0,65	1,11	0,51	0,74	1,12	2,47	0,90	1,41	0,46	1,15	0,52	0,53	2,30	5,85	5,57	2,52	62,84	2,14	0,77	0,81	0,77	1,00	0,73	1,38	0,85	0,62	0,19	0,03
španělština	0	0	0	0	0	0	0,52	0,60	0,33	0,66	0,68	2,05	0,83	1,13	0,52	0,91	0,60	0,42	3,11	6,34	12,11	8,54	3,26	52,05	0,57	0,49	0,31	0,66	0,58	0,79	1,04	0,65	0,21	0,03
estonština	0,02	0	0	0,02	0	0,02	0,44	0,58	0,28	0,67	0,48	0,94	1,00	0,98	0,49	1,01	0,83	0,82	0,41	1,01	0,87	0,62	0,76	0,34	77,00	4,84	0,46	1,44	0,95	0,85	0,83	0,96	0,07	0
finština	0,03	0	0	0,02	0	0	0,52	0,43	0,36	0,39	0,34	0,87	0,52	0,57	0,40	0,76	0,75	0,80	0,47	0,80	0,73	0,44	0,65	0,31	5,10	80,65	0,62	1,17	0,53	0,58	0,61	0,46	0,10	0,01
maďarština	0	0	0	0,05	0,02	0,02	0,72	0,34	0,32	0,91	0,32	0,77	0,83	0,72	0,55	0,93	0,43	0,39	0,52	1,15	0,72	0,77	0,83	0,39	0,79	0,43	84,49	0,56	0,22	0,47	0,70	0,49	0,14	0
litevština	0	0	0	0,02	0	0	1,01	1,28	0,75	1,22	0,98	0,60	0,89	0,54	0,49	0,64	0,53	0,52	0,72	1,43	1,52	0,38	1,07	0,51	1,51	0,72	0,42	73,96	4,67	1,78	1,14	0,66	0,03	0
lotyština	0	0	0,01	0,01	0,02	0,01	0,94	1,32	0,57	0,85	1,01	0,46	0,41	0,61	0,44	0,79	0,62	0,58	0,54	1,30	1,13	0,69	0,99	0,62	1,52	0,89	0,36	5,21	75,08	1,48	0,87	0,59	0,06	0,01
albánština	0	0	0,01	0,01	0	0,05	0,80	1,18	0,40	0,60	0,66	0,75	1,14	0,71	0,45	0,82	0,53	0,42	0,87	2,27	1,79	1,11	1,45	0,70	1,01	0,42	0,27	1,37	0,74	77,55	0,97	0,79	0,14	0,01
baskičtina	0,01	0	0	0,01	0	0	0,50	0,50	0,59	0,45	0,41	0,77	1,08	0,67	0,47	1,16	0,91	0,26	0,85	1,51	1,94	1,15	1,02	1,40	0,89	0,50	0,49	0,79	0,38	1,21	79,08	0,92	0,04	0,03
turečtina	0,01	0	0	0,04	0,01	0,01	0,47	0,60	0,24	0,33	0,46	0,49	0,66	0,76	0,35	0,74	0,48	0,44	0,38	0,88	0,62	0,58	0,69	0,62	1,17	0,57	0,51	0,59	0,27	1,12	0,84	85,01	0,04	0,01
vietnamština	0	0	0	0,01	0	0,04	0,17	0,11	0,10	0,11	0,11	0,58	0,16	0,31	0,04	0,55	0,14	0,17	0,20	0,35	0,33	0,21	0,36	0,23	0,08	0,12	0,13	0,09	0,08	0,11	0,15	0,04	94,89	0,02
řečtina	0	0	0,01	0,02	0	0,05	0,02	0,04	0,01	0,03	0,01	0,20	0,08	0,23	0,03	0,15	0,04	0,04	0,08	0,09	0,07	0,07	0,08	0,04	0,04	0	0,03	0,01	0,02	0,09	0,02	0,03	0,03	98,33

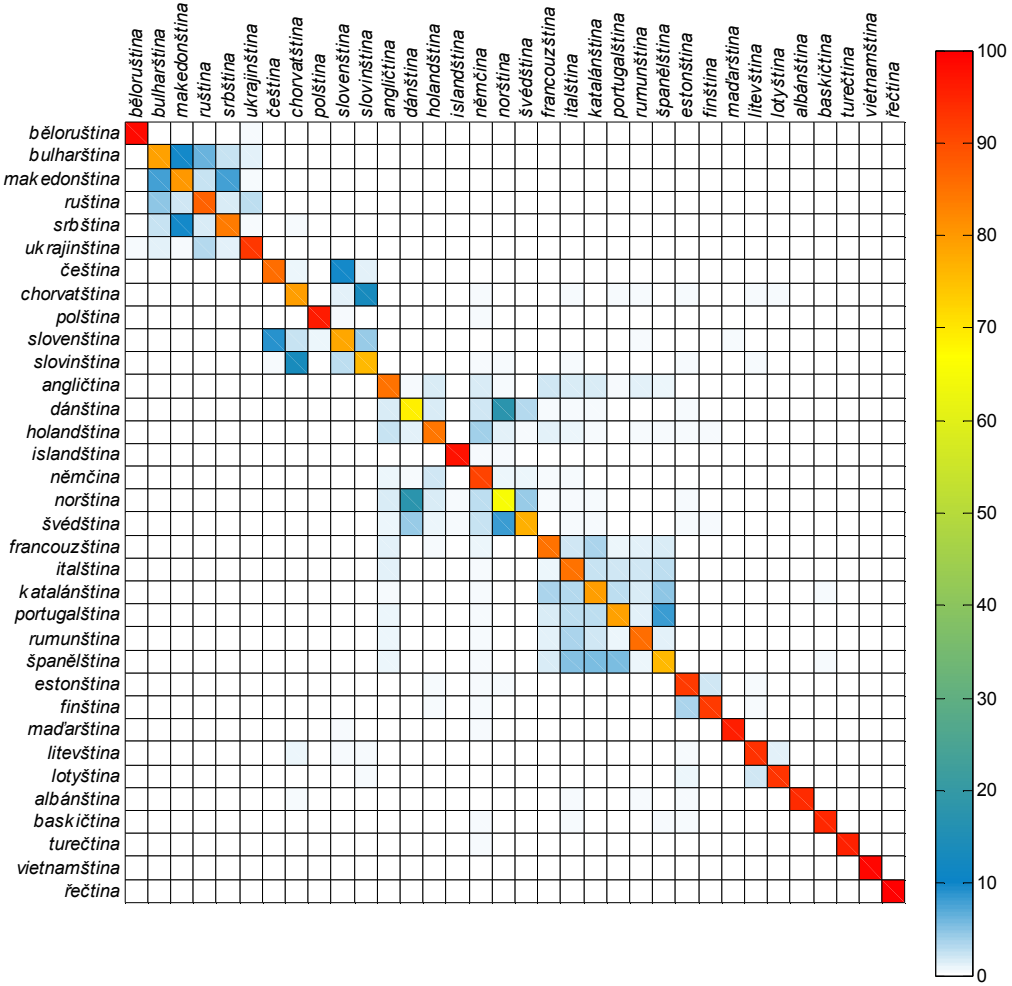
Tabulka D.8 Konfuzní matice 8-gramového modelu Witten-Bell při délce testovacích řetězců 5 znaků



Graf D.9 Konfuzní matice unigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	94,33	0,56	0,66	1,63	0,34	2,13	0	0,01	0,09	0,02	0	0,08	0	0,01	0	0,02	0	0	0,01	0,02	0	0	0,01	0	0	0	0,02	0,01	0,01	0	0,01	0,02	0	0
bulharština	0,40	63,03	20,49	12,16	3,05	0,69	0	0,03	0	0,01	0,01	0,02	0	0,01	0	0,01	0	0,01	0	0	0	0,03	0	0,01	0,01	0,01	0	0	0	0	0,01	0	0	0
makedonština	0,07	11,47	65,83	5,51	16,46	0,10	0	0,10	0	0,01	0,11	0,02	0	0,03	0	0,04	0	0	0	0,04	0	0,02	0,01	0,03	0,04	0,04	0	0	0,01	0	0,04	0,01	0	0
ruština	1,73	14,16	10,30	67,85	3,14	2,75	0	0	0	0	0	0,01	0	0	0	0,02	0	0	0	0,01	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,09	5,14	27,21	4,20	62,02	0,21	0	0,38	0	0,02	0,32	0,01	0	0,05	0	0,03	0,01	0	0	0,06	0	0,02	0,03	0,02	0,05	0,02	0	0,02	0,02	0	0,06	0	0	0
ukrajinština	3,01	5,86	3,08	12,07	1,88	74,06	0	0	0	0,01	0	0	0	0	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0	0	69,98	1,60	0,45	15,66	2,45	0,67	0,05	0,42	0,37	0,33	0,12	0,12	0,59	0,51	0,29	0,61	0,13	0,39	0,26	0,41	3,26	0,70	0,09	0,02	0,32	0,19	0	0
chorvatština	0	0	0	0	0	0	0	43,42	0,58	1,14	28,83	0,55	0,16	2,31	0,15	0,85	0,97	0,31	0,28	2,76	0,20	1,08	0,76	1,07	3,11	1,87	0,10	4,53	1,23	0,33	3,29	0,11	0,01	0
polština	0	0	0	0	0	0	0,02	1,25	87,04	0,42	0,91	2,05	0,03	1,61	0,04	1,21	0,20	0,23	0,04	0,81	0,02	0,17	0,22	0,40	0,17	0,35	0,60	0,75	0,09	0,08	1,01	0,25	0,02	0
slovenština	0	0	0	0	0	0	11,06	4,24	1,80	51,74	8,45	1,70	0,14	1,55	0,81	0,69	0,46	0,40	0,71	2,66	0,47	1,30	0,30	1,12	0,89	1,32	3,82	1,70	0,51	0,14	1,57	0,43	0,01	0
slovinština	0	0	0	0	0	0	0,07	14,49	0,58	1,94	56,59	0,70	0,63	4,05	0,10	0,87	1,79	0,33	0,28	3,54	0,21	0,98	0,36	1,17	2,42	1,26	0,21	3,20	0,85	0,35	2,91	0,11	0	0
angličtina	0	0	0	0	0	0	0	0,17	1,13	0,45	0,36	52,42	2,09	6,50	0,14	6,92	1,91	0,75	2,65	8,18	1,25	2,01	1,07	6,80	1,28	0,99	0,06	0,46	0,13	0,30	1,39	0,56	0,02	0
dánština	0	0	0	0	0	0	0,01	0,23	0,09	0,18	1,36	4,45	45,25	7,07	0,67	3,51	21,52	3,27	0,69	2,06	0,51	0,88	0,22	0,90	3,26	0,87	0,10	0,35	0,03	0,17	1,95	0,40	0	0
holandština	0	0	0	0	0	0	0	1,68	1,74	0,64	4,89	8,81	3,44	44,31	0,11	8,80	5,09	0,54	1,17	4,81	0,78	1,37	0,85	2,07	2,11	0,73	0,15	0,16	0,12	0,50	5,01	0,12	0	0
islandština	0	0	0	0	0	0	0,34	0,08	0,07	0,64	0,09	0,63	1,11	0,37	87,42	1,03	0,89	1,37	0,39	0,15	0,60	0,26	0,07	0,15	0,84	0,30	2,23	0,20	0	0,03	0,48	0,21	0,04	0
němčina	0	0	0	0	0	0	0	0,24	0,52	0,21	0,37	8,38	2,20	7,39	0,10	62,86	2,07	2,48	0,82	2,30	0,70	0,59	0,92	1,59	2,20	0,62	0,19	0,13	0,07	0,17	2,54	0,31	0,02	0
norština	0	0	0	0	0	0	0	0,55	0,12	0,26	2,22	4,64	20,60	7,36	0,74	4,04	38,20	3,52	0,97	1,97	0,52	1,05	0,42	1,23	5,26	1,86	0,32	0,82	0,02	0,32	2,71	0,26	0,02	0
švédština	0	0	0	0	0	0	0,01	0,64	0,16	1,00	0,94	4,73	3,39	3,01	0,67	4,09	9,68	55,28	0,47	1,55	0,81	0,79	0,34	0,75	4,78	3,59	0,23	0,64	0,17	0,21	1,49	0,56	0,01	0
francouzština	0	0	0	0	0	0	0,04	0,69	0,15	0,34	0,35	2,94	1,01	1,95	0,12	2,45	1,17	0,11	51,12	8,81	6,08	5,80	2,95	9,97	1,36	0,48	0,85	0,27	0,04	0,28	0,55	0,08	0,03	0
italština	0	0	0	0	0	0	0	0,36	0,37	0,25	1,29	5,56	1,44	3,12	0,13	2,35	1,12	0,10	5,27	48,79	5,35	4,00	3,99	10,35	1,44	0,73	0,07	0,42	0,12	0,02	3,21	0,15	0	0
katalánština	0	0	0	0	0	0	0,19	0,48	0,14	0,39	0,39	2,26	1,26	1,76	0,34	1,66	0,97	0,25	13,27	11,14	40,48	6,18	2,71	12,17	1,32	0,38	0,47	0,17	0,05	0,25	1,13	0,19	0	0
portugalština	0	0	0	0	0	0	0,41	0,71	0,15	1,68	0,87	2,18	0,83	1,99	0,43	1,44	0,58	0,05	8,05	8,85	5,74	47,26	2,03	12,92	1,13	0,35	0,69	0,25	0,06	0,17	1,06	0,10	0,02	0
rumunština	0	0	0	0	0	0	0	0,67	0,11	0,07	0,46	1,14	0,24	1,17	0,09	2,81	0,36	0,12	3,04	12,67	2,85	2,41	63,81	4,10	1,21	0,45	0,07	0,31	0,18	0,01	1,37	0,27	0	0
španělština	0	0	0	0	0	0	0	0,87	0,45	0,26	1,12	4,21	0,87	2,71	0,07	1,72	0,88	0,18	6,93	15,22	5,56	7,77	1,63	45,87	0,98	0,52	0,07	0,31	0,08	0,06	1,35	0,31	0	0
estonština	0	0	0	0	0	0	0	2,45	0,11	0,09	1,85	0,48	0,52	1,50	0,26	2,33	1,72	2,44	0,71	2,04	0,42	0,75	0,51	1,01	60,06	14,42	0,07	2,00	0,15	0,21	3,19	0,69	0,01	0
finština	0	0	0	0	0	0	0	2,33	0,11	0,18	1,54	0,93	0,01	0,72	0,08	1,05	0,41	1,64	0,53	1,48	0,20	0,32	0,36	0,52	10,12	73,83	0,04	1,79	0,07	0,10	1,40	0,23	0	0
maďarština	0	0	0	0	0	0	0,80	0,30	1,10	3,22	0,64	0,85	0,34	0,95	1,89	1,25	0,56	0,92	0,73	0,41	0,52	0,52	0,16	0,26	0,33	0,32	80,98	0,17	0,20	0,03	1,55	0,99	0	0
litevština	0	0	0	0	0	0	0,02	6,05	0,78	0,82	3,56	0,35	0,09	0,31	0,07	0,28	0,64	0,22	0,17	1,52	0,13	0,39	0,29	0,44	4,37	3,93	0,04	71,87	1,02	0,08	2,34	0,21	0	0
lotyština	0	0	0	0	0	0	0	3,12	0,08	0,11	2,43	0,15	0,07	0,44	0,04	0,56	0,24	0,11	0,21	0,92	0,20	0,25	0,51	0,26	3,25	1,33	0,04	3,86	79,32	0,10	2,33	0,06	0	0
albánština	0	0	0	0	0	0	0	1,26	0,12	0,32	1,35	1,26	0,54	2,36	0,19	2,14	1,10	0,24	0,28	1,42	0,43	0,48	0,62	0,28	1,96	0,75	0,06	0,72	0,06	79,70	1,73	0,62	0	0
baskičtina	0	0	0	0	0	0	0	1,22	0,12	0,06	1,21	0,76	0,35	2,28	0,12	1,97	0,81	0,08	0,21	1,36	0,37	0,38	0,46	0,80	2,76	1,06	0,14	0,48	0,10	0,05	82,70	0,14	0	0
turečtina	0	0	0	0	0	0	0,01	0,20	0,40	0,31	0,38	0,91	0,68	0,86	0,14	2,78	1,17	1,32	0,18	0,73	0,15	0,28	0,73	0,68	1,51	1,06	0,69	0,43	0,13	0,14	2,05	82,07	0	0
vietnamština	0	0	0	0	0	0	0,01	0,02	0,04	0,09	0	0,08	0	0,01	0,02	0,30	0	0,06	0,15	0,12	0,25	0,10	0,12	0,07	0,01	0,02	0,03	0	0	0	0,01	0	98,48	0
řečtina	0	0	0	0	0	0	0	0	0,03	0	0	0,08	0	0,01	0	0,16	0,01	0,02	0,02	0,07	0,02	0,01	0,03	0,03	0,02	0,04	0	0,02	0	0	0,02	0	0	99,40

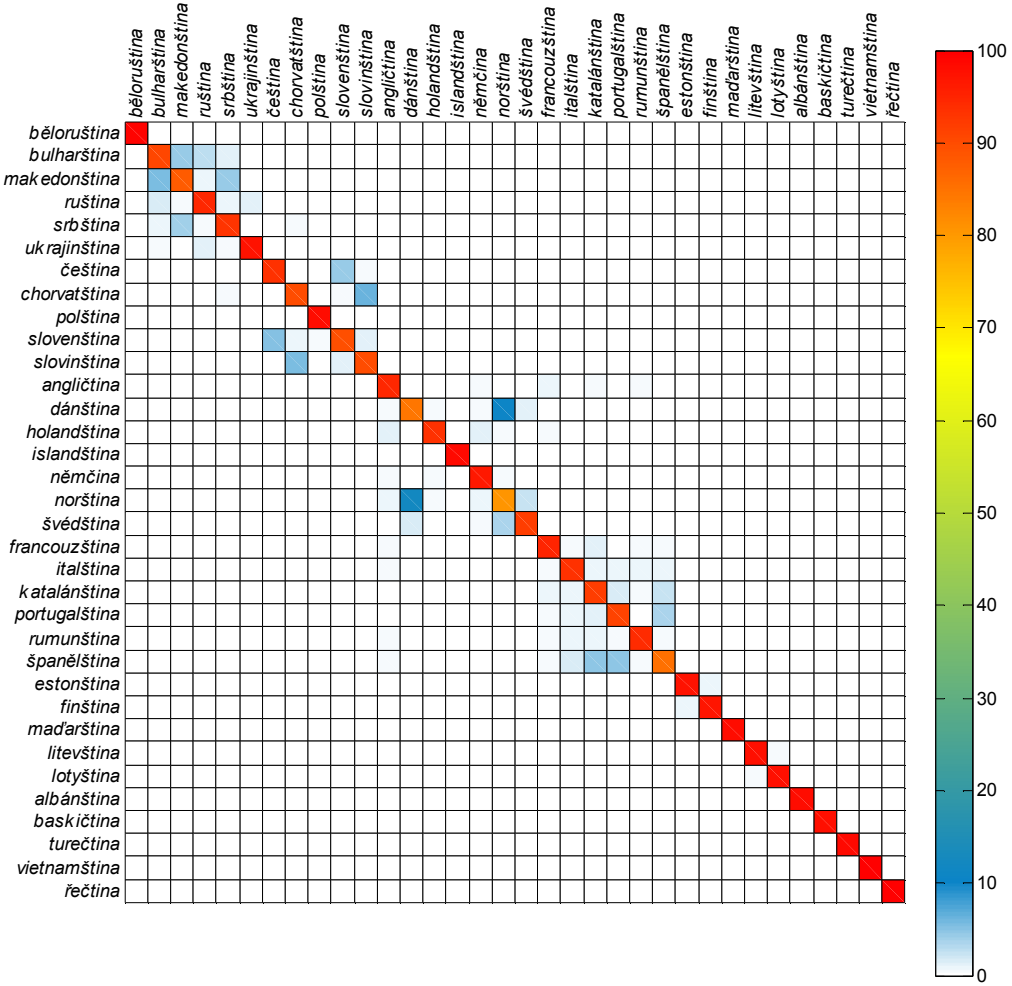
Tabulka D.9 Konfuzní matice unigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.10 Konfuzní matice bigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	98,69	0,07	0,07	0,38	0,02	0,47	0,02	0	0,08	0,08	0,01	0,03	0	0,01	0	0,01	0,02	0	0	0	0,01	0	0,01	0	0	0	0	0	0	0,01	0	0	0	0
bulharština	0,12	79,03	10,06	6,63	2,51	1,49	0,01	0,04	0	0,01	0,03	0	0,01	0,01	0	0	0	0	0	0,01	0	0	0,01	0	0	0	0	0	0	0	0,02	0	0	0
makedonština	0,10	8,05	80,35	2,54	7,93	0,50	0,01	0,23	0,01	0,01	0,14	0,02	0	0	0	0	0	0	0	0,02	0	0,04	0,01	0	0,01	0	0	0,02	0	0	0	0	0	0
ruština	0,32	5,04	2,31	87,42	1,82	3,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,03	2,66	9,79	1,78	83,87	0,74	0	0,75	0	0,03	0,31	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0,01	0	0	0	0	0
ukrajinština	0,71	1,34	0,73	3,39	1,18	92,62	0	0	0	0,01	0	0	0	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0,01	0	85,74	0,83	0,21	9,79	1,23	0,29	0,04	0,07	0,09	0,33	0,06	0,04	0,16	0,13	0,04	0,16	0,15	0,03	0,06	0,01	0,22	0,10	0,05	0,07	0,03	0,04	0	0,01
chorvatština	0	0	0	0	0,01	0	0,25	79,38	0,16	1,34	12,94	0,29	0,12	0,15	0,04	0,46	0,37	0,11	0,09	0,65	0,24	0,45	0,42	0,16	0,45	0,07	0,03	0,60	0,49	0,30	0,33	0,09	0	0
polština	0	0	0	0	0	0	0,12	0,38	96,16	0,49	0,34	0,34	0,06	0,10	0,01	0,55	0,05	0,10	0,12	0,21	0,09	0,04	0,15	0,02	0,11	0,05	0,08	0,12	0,08	0,07	0,13	0,02	0	0
slovenština	0	0	0	0	0	0	9,17	2,69	1,05	78,19	4,33	0,29	0,10	0,12	0,06	0,36	0,20	0,16	0,17	0,32	0,18	0,36	0,48	0,16	0,16	0,09	0,41	0,35	0,25	0,06	0,24	0,04	0	0
slovinština	0	0	0	0	0	0	0,60	13,49	0,27	3,08	76,03	0,33	0,21	0,24	0,03	0,51	0,55	0,25	0,14	0,51	0,09	0,35	0,25	0,36	0,52	0,07	0,12	0,75	0,35	0,31	0,37	0,20	0	0,01
angličtina	0	0	0	0	0	0	0,05	0,06	0,10	0,21	0,11	85,05	0,77	1,67	0,04	1,72	0,77	0,32	2,24	1,57	1,72	0,55	1,18	0,96	0,37	0,07	0,14	0,03	0,07	0,10	0,08	0,04	0	0
dánština	0,01	0	0	0	0,01	0	0,02	0,16	0,05	0,09	0,20	1,62	68,96	1,91	0,35	1,99	17,81	3,15	0,56	0,63	0,40	0,12	0,24	0,19	0,47	0,15	0,18	0,18	0,07	0,16	0,22	0,08	0	0,01
holandština	0	0	0	0	0	0	0,07	0,02	0,12	0,24	0,08	2,62	1,29	84,40	0,01	3,91	1,36	0,40	1,18	0,95	0,49	0,14	0,73	0,47	0,47	0,40	0,04	0,09	0,12	0,09	0,21	0,07	0	0,02
islandština	0	0	0	0	0	0	0,02	0,03	0,01	0,10	0,01	0,22	0,19	0,07	97,29	0,54	0,42	0,25	0,05	0,08	0,10	0,08	0,01	0,02	0,13	0,08	0,16	0	0,03	0,03	0,02	0,04	0	0,01
němčina	0	0	0	0	0	0	0,05	0,04	0,09	0,09	0,07	0,88	0,52	2,25	0,07	91,31	0,81	0,91	0,55	0,62	0,22	0,15	0,20	0,16	0,24	0,21	0,05	0,05	0,08	0,05	0,25	0,07	0	0
norština	0	0	0	0	0	0	0,01	0,25	0,08	0,17	0,29	1,75	17,68	1,83	0,48	2,80	65,49	4,54	0,63	0,63	0,52	0,14	0,21	0,24	0,75	0,34	0,32	0,12	0,07	0,17	0,29	0,19	0	0
švédština	0	0	0	0	0	0	0,02	0,23	0,03	0,22	0,34	1,09	4,38	1,06	0,54	2,64	8,53	76,98	0,19	0,55	0,45	0,14	0,13	0,13	0,72	0,58	0,23	0,19	0,09	0,11	0,15	0,26	0	0,01
francouzština	0	0	0	0	0	0	0,02	0	0,03	0,08	0,05	1,42	0,15	0,64	0,01	1,09	0,28	0,12	84,87	2,31	3,66	1,15	1,48	1,64	0,16	0,07	0,20	0,10	0,11	0,11	0,19	0,02	0,03	0
italština	0	0	0	0	0	0	0,01	0,22	0,07	0,09	0,22	1,34	0,20	0,26	0,02	0,49	0,32	0,20	0,88	84,81	2,71	2,10	2,24	2,84	0,23	0,11	0,06	0,13	0,05	0,08	0,20	0,08	0,03	0
katalánština	0	0	0	0	0	0	0,06	0,12	0,05	0,09	0,12	0,72	0,10	0,23	0,06	0,58	0,37	0,23	3,77	3,15	79,46	2,91	1,81	4,96	0,26	0,06	0,26	0,05	0,09	0,05	0,40	0,02	0	0,01
portugalština	0	0	0	0	0,01	0	0,11	0,17	0,05	0,22	0,26	0,84	0,12	0,07	0,08	0,48	0,09	0,05	1,67	2,87	2,97	79,20	1,27	8,44	0,20	0,05	0,13	0,27	0,10	0,04	0,18	0,05	0	0
rumunština	0	0	0	0	0,02	0	0,02	0,30	0,08	0,23	0,25	1,05	0,14	0,20	0,01	0,68	0,24	0,15	1,25	3,88	2,27	1,10	85,61	1,23	0,25	0,01	0,07	0,18	0,16	0,17	0,27	0,16	0	0,01
španělština	0	0	0	0	0,01	0	0	0,14	0,04	0,14	0,27	1,13	0,11	0,20	0,01	0,40	0,16	0,11	1,94	5,27	5,60	5,67	1,13	76,12	0,18	0,10	0,06	0,29	0,13	0,08	0,58	0,12	0	0
estonština	0	0	0	0	0	0	0,01	0,19	0,01	0,10	0,28	0,28	0,17	0,61	0,06	0,76	0,40	0,33	0,21	0,26	0,23	0,11	0,26	0,06	92,29	2,01	0,11	0,46	0,21	0,24	0,24	0,10	0	0
finština	0	0	0	0	0	0	0,01	0,04	0,05	0,06	0,11	0,27	0,07	0,48	0,07	0,40	0,32	0,26	0,17	0,38	0,04	0,06	0,14	0,08	3,77	92,36	0,06	0,42	0,13	0,01	0,19	0,04	0	0
maďarština	0	0	0	0	0	0	0,13	0,01	0,10	0,44	0,13	0,34	0,11	0,16	0,11	0,55	0,30	0,12	0,24	0,12	0,25	0,25	0,07	0,01	0,26	0,07	95,84	0,05	0,02	0,07	0,07	0,17	0	0
litevština	0	0	0	0	0	0	0,04	0,88	0,12	0,40	0,40	0,13	0,04	0,05	0,03	0,12	0,06	0,03	0,10	0,32	0,10	0,38	0,17	0,25	0,62	0,21	0,05	93,57	1,55	0,09	0,24	0,04	0	0
lotyština	0	0	0	0	0	0	0,01	0,39	0,08	0,26	0,47	0,09	0,02	0,24	0,02	0,31	0,10	0,07	0,15	0,10	0,11	0,14	0,21	0,13	1,11	0,31	0,03	2,16	93,28	0,07	0,12	0,01	0	0
albánština	0	0	0	0	0	0	0,03	0,44	0,03	0,16	0,39	0,21	0,16	0,26	0,02	0,34	0,39	0,06	0,19	0,62	0,22	0,17	0,42	0,16	0,57	0,10	0,03	0,20	0,06	94,35	0,25	0,16	0	0
baskičtina	0	0	0	0	0	0	0	0,25	0,06	0,06	0,13	0,22	0,06	0,28	0,03	0,52	0,24	0,09	0,15	0,58	0,33	0,26	0,32	0,40	0,41	0,19	0,14	0,20	0,06	0,14	94,66	0,21	0	0
turečtina	0	0	0	0	0	0	0	0,08	0	0,05	0,28	0,14	0,36	0,30	0,07	0,52	0,36	0,17	0,09	0,21	0,07	0,11	0,25	0,11	0,36	0,09	0,21	0,06	0,05	0,21	0,38	95,45	0	0,01
vietnamština	0	0	0	0	0	0	0	0,02	0,02	0,02	0	0,09	0	0,04	0	0,12	0	0,02	0,12	0,08	0,11	0,04	0,02	0,05	0	0	0,01	0	0,02	0,01	0	0,01	99,16	0,03
řečtina	0	0	0	0	0	0	0	0	0,01	0	0	0,06	0	0	0	0,08	0,01	0	0,02	0,02	0,01	0,02	0,02	0,01	0,01	0,01	0	0,01	0,03	0,03	0,01	0	0	99,63

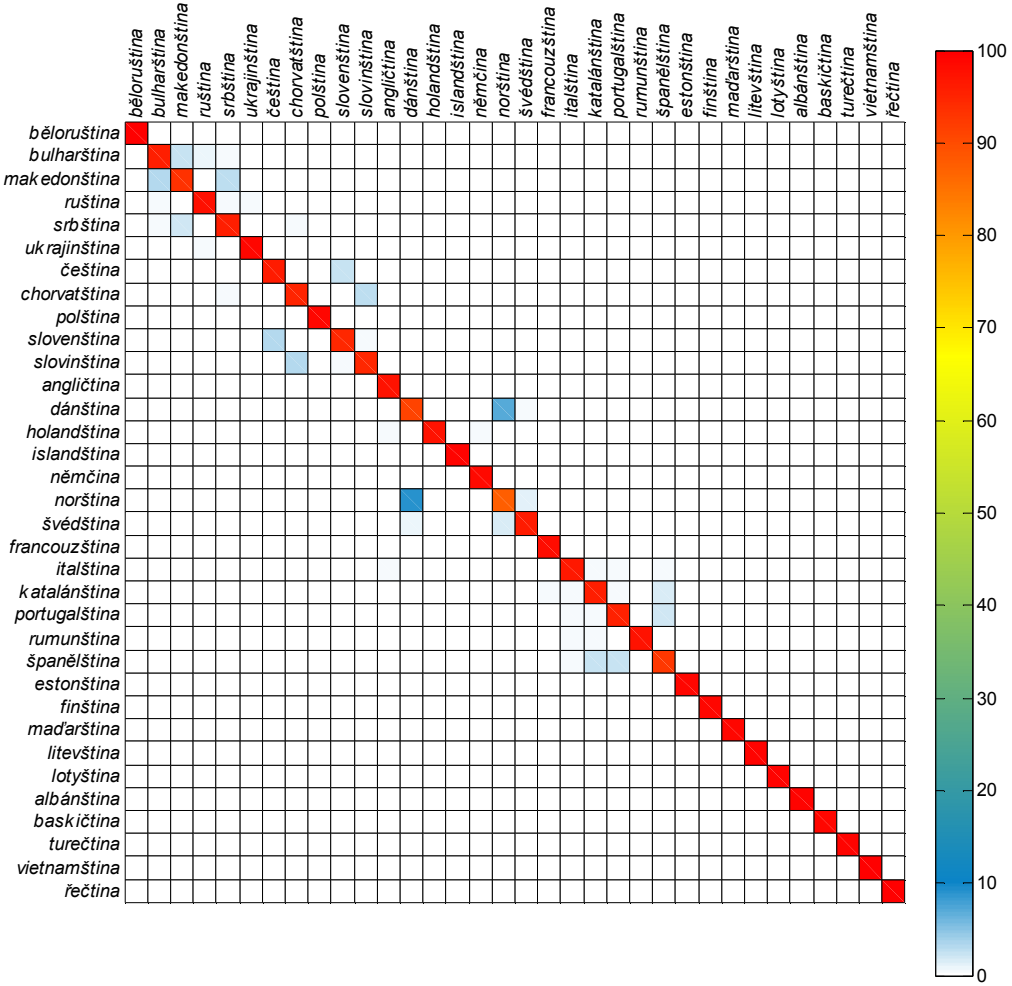
Tabulka D.10 Konfuzní matice bigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.11 Konfuzní matice trigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,59	0	0	0,10	0	0,13	0,01	0,03	0,05	0,04	0,01	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0,01
bulharština	0,02	90,68	4,53	2,96	1,44	0,26	0,01	0,04	0	0,03	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0,03	5,74	88,16	0,96	4,53	0,25	0,01	0,13	0,01	0,02	0,08	0,02	0	0	0	0	0	0	0	0	0	0,03	0,01	0,01	0	0	0	0	0	0	0	0	0	0
ruština	0,10	1,92	0,70	94,92	0,83	1,52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0	1,10	4,03	0,62	93,02	0,38	0,03	0,59	0	0,02	0,17	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0
ukrajinština	0,12	0,42	0,16	1,34	0,47	97,48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0,01	0	0,01	0	0,02	0	93,54	0,39	0,16	4,50	0,49	0,20	0,01	0,05	0	0,12	0,06	0,03	0,01	0,07	0,05	0,07	0,04	0,02	0,02	0	0,05	0,01	0,01	0,01	0,01	0,01	0,01	0,01
chorvatština	0,01	0	0,03	0,03	0,40	0	0,23	89,95	0,19	0,73	6,58	0,09	0,14	0,10	0,02	0,12	0,11	0,08	0,01	0,20	0,08	0,06	0,17	0,06	0,10	0,01	0,03	0,13	0,07	0,07	0,05	0,02	0,01	0,11
polština	0,01	0,01	0,01	0,02	0	0	0,15	0,14	98,14	0,30	0,11	0,15	0,08	0,04	0,01	0,21	0,11	0,01	0,07	0,05	0,03	0,06	0,05	0	0,04	0,01	0,02	0,01	0,01	0,02	0,05	0	0,01	0,06
slovenština	0,03	0,03	0	0	0,03	0	5,42	1,07	0,57	89,71	1,48	0,19	0,03	0,02	0	0,08	0,09	0,02	0,09	0,17	0,03	0,10	0,19	0,09	0,03	0,05	0,12	0,12	0,08	0,04	0,05	0,02	0	0,04
slovinština	0,02	0	0	0,02	0,06	0	0,37	5,84	0,25	1,42	90,11	0,18	0,04	0,10	0,02	0,20	0,19	0,06	0,05	0,13	0,05	0,09	0,11	0,08	0,03	0,02	0,02	0,07	0,10	0,15	0,07	0,05	0,05	0,04
angličtina	0	0,01	0,01	0,04	0	0,03	0,02	0,04	0,04	0,09	0,06	94,67	0,28	0,36	0,05	0,54	0,37	0,17	0,85	0,39	0,55	0,19	0,41	0,21	0,08	0,03	0,01	0	0	0,06	0	0,02	0,09	0,32
dánština	0,02	0	0,01	0,01	0	0	0,01	0,05	0,01	0,08	0,06	0,49	84,63	0,52	0,11	0,64	10,86	1,32	0,30	0,13	0,13	0,02	0,07	0,06	0,12	0,04	0,01	0	0	0,05	0,07	0,09	0	0,08
holandština	0,01	0,01	0	0,06	0	0,01	0,06	0,03	0,02	0,05	0,12	1,24	0,37	93,72	0,02	1,34	0,50	0,17	0,53	0,27	0,21	0,10	0,34	0,13	0,18	0,08	0,01	0,05	0,01	0,02	0,06	0,03	0,07	0,17
islandština	0	0	0,01	0	0	0	0,02	0,01	0,01	0,03	0,03	0,07	0,10	0,04	98,72	0,19	0,21	0,13	0,04	0,07	0,03	0,06	0	0,03	0,04	0	0,05	0,01	0,01	0,01	0,01	0,03	0	0,03
němčina	0,02	0	0	0,02	0	0,01	0,03	0,06	0,09	0,03	0,05	0,46	0,20	0,60	0	96,52	0,45	0,21	0,22	0,15	0,13	0,05	0,16	0,07	0,03	0,01	0,04	0,01	0,02	0,08	0,03	0,02	0,03	0,19
norština	0,01	0	0	0	0	0,01	0,03	0,12	0,01	0,06	0,13	0,79	12,36	0,49	0,19	0,79	80,68	2,55	0,21	0,27	0,28	0,05	0,10	0,12	0,17	0,09	0,03	0,03	0,04	0,13	0,07	0,02	0,02	0,14
švédština	0	0	0	0	0	0	0	0,04	0,01	0,03	0,09	0,33	1,74	0,29	0,13	0,56	3,87	91,88	0,07	0,08	0,16	0,05	0,05	0,03	0,09	0,08	0,01	0,06	0,04	0,05	0,08	0,05	0,02	0,10
francouzština	0,01	0	0,01	0,01	0	0	0,03	0,04	0,02	0,04	0,01	0,49	0,07	0,11	0	0,30	0,26	0,05	95,28	0,54	1,21	0,37	0,40	0,49	0,11	0	0,02	0,01	0	0,02	0	0,03	0,01	0,05
italština	0,01	0	0,02	0,04	0	0,01	0,01	0,19	0,03	0,06	0,16	0,66	0,02	0,10	0,01	0,19	0,18	0,09	0,51	93,59	0,82	0,94	0,87	0,91	0,05	0,04	0,04	0,05	0,04	0,06	0,04	0,04	0,06	0,15
katalánština	0	0	0	0	0,01	0,01	0,02	0,02	0,05	0,04	0,04	0,34	0,05	0,05	0,02	0,08	0,16	0,05	1,09	1,12	91,90	1,63	0,56	2,44	0,06	0,03	0,03	0,03	0	0,02	0,07	0,01	0,02	0,04
portugalština	0	0	0	0	0	0	0,06	0,04	0,05	0,05	0,06	0,26	0,03	0,03	0,06	0,12	0,11	0,05	0,51	1,13	1,49	91,35	0,38	3,78	0,06	0,02	0,03	0,08	0,01	0,05	0,06	0	0,04	0,08
rumunština	0,02	0	0,01	0,01	0	0	0,07	0,13	0,05	0,14	0,10	0,61	0,06	0,11	0,02	0,19	0,10	0	0,47	1,00	1,15	0,47	94,28	0,44	0,05	0,02	0,05	0,08	0,06	0,06	0,04	0,04	0,08	0,08
španělština	0	0,01	0,01	0,04	0,01	0,01	0	0,04	0,04	0,07	0,07	0,43	0,01	0,05	0,02	0,18	0,08	0,07	0,68	1,87	4,74	4,91	0,67	85,51	0,06	0,05	0,02	0,05	0,06	0,02	0,10	0,01	0,02	0,08
estonština	0,01	0	0	0	0	0,01	0,01	0,07	0,02	0,04	0,07	0,12	0,07	0,14	0,07	0,11	0,13	0,11	0,04	0,08	0,11	0,03	0,05	0,02	97,38	0,92	0,04	0,10	0,07	0,06	0,04	0,03	0,01	0,03
finština	0	0	0	0	0	0,01	0	0,07	0,02	0,06	0,07	0,17	0,06	0,09	0,02	0,16	0,13	0,13	0,05	0,10	0,06	0,06	0,05	0,05	1,14	97,19	0	0,05	0,06	0,03	0,03	0,02	0	0,11
maďarština	0,01	0	0	0,02	0	0	0,03	0,04	0,07	0,14	0,04	0,22	0,03	0,05	0,02	0,14	0,07	0,07	0,07	0,05	0,09	0,12	0,06	0,03	0,01	0,03	98,37	0	0,02	0,02	0,01	0,03	0,03	0,10
litevština	0,01	0	0,01	0	0	0	0,01	0,15	0,09	0,11	0,14	0,02	0,02	0	0	0,04	0,04	0,01	0	0,08	0,05	0,12	0,05	0,16	0,09	0,03	0,01	98,01	0,63	0,04	0,04	0	0	0,03
lotyština	0,01	0	0,01	0	0,01	0	0,03	0,11	0,04	0,11	0,13	0,03	0,02	0,02	0	0,09	0,06	0,07	0,05	0,06	0,04	0,04	0,06	0,07	0,29	0,07	0,01	0,62	97,81	0,02	0,06	0	0,02	0,03
albánština	0	0	0	0	0	0	0	0,14	0,04	0,05	0,09	0,06	0,02	0,08	0	0,07	0,08	0,02	0,04	0,17	0,15	0,07	0,10	0,03	0,07	0	0	0,06	0,05	98,42	0,08	0,03	0,02	0,05
baskičtina	0	0	0	0	0	0	0	0,10	0,03	0,02	0,09	0,07	0,03	0,04	0	0,19	0,13	0,07	0,07	0,22	0,12	0,10	0,15	0,18	0,12	0,06	0,01	0,04	0,03	0,05	97,98	0,05	0	0,04
turečtina	0	0	0,01	0,02	0,01	0,01	0,02	0,05	0,01	0,06	0,05	0,04	0,09	0,06	0,04	0,06	0,09	0,01	0,03	0,06	0,03	0,05	0,09	0,03	0,05	0,03	0,12	0,02	0	0,02	0,07	98,73	0	0,03
vietnamština	0	0	0	0	0	0	0	0	0	0	0	0,05	0	0,04	0	0,01	0,03	0	0,04	0,02	0,02	0,01	0	0,02	0	0	0,01	0	0	0	0	0	99,63	0,11
řečtina	0	0	0	0,01	0	0	0	0,01	0	0,01	0	0,06	0,01	0	0	0,02	0,01	0	0	0,01	0,01	0,01	0	0,01	0	0,01	0	0,02	0	0,02	0	0	0,01	99,76

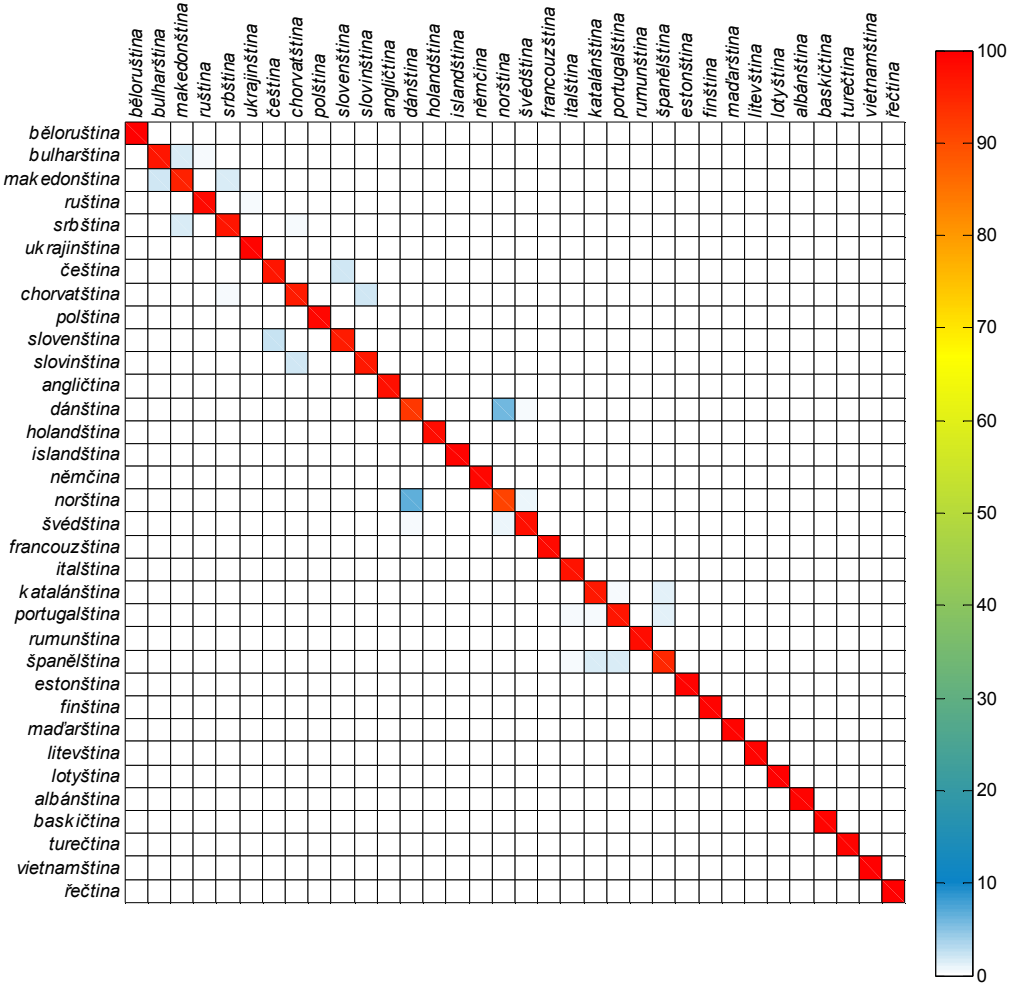
Tabulka D.11 Konfuzní matice trigramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.12 Konfuzní matice 4-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,80	0	0,01	0,04	0	0,02	0,01	0,01	0,03	0,01	0	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,03	0	0	0	0	0,01
bulharština	0	95,73	2,57	0,94	0,57	0,12	0	0,04	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0	3,33	93,42	0,20	2,79	0,05	0	0,07	0	0,02	0,07	0,02	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0
ruština	0,05	0,71	0,25	97,83	0,44	0,71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0	0,53	2,33	0,30	96,06	0,11	0	0,57	0	0,04	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0
ukrajinština	0,06	0,10	0,03	0,67	0,16	98,96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0	0
čeština	0,01	0	0	0	0,01	0	96,40	0,14	0,07	2,65	0,17	0,16	0,01	0,07	0	0,03	0,05	0,03	0,02	0,04	0	0,02	0,04	0	0	0	0,04	0	0	0,01	0	0,01	0	0,01
chorvatština	0,01	0	0,01	0,02	0,64	0	0,11	94,97	0,11	0,21	3,06	0,06	0,05	0,04	0,01	0,06	0,04	0,01	0,02	0,08	0,03	0,04	0,08	0	0,01	0,03	0,01	0,02	0,09	0,03	0	0,02	0,07	0,05
polština	0,03	0	0,02	0,03	0	0	0,17	0,07	98,83	0,16	0,13	0,11	0,05	0,01	0,01	0,04	0,06	0,01	0,06	0,03	0	0,01	0,03	0,01	0	0	0,03	0	0,01	0,01	0	0	0,02	0,05
slovenština	0,03	0,02	0,01	0	0,08	0	3,25	0,35	0,28	94,85	0,46	0,04	0,02	0,05	0,01	0,03	0,02	0,06	0,06	0	0,02	0,06	0,07	0,06	0	0,01	0,05	0,02	0,02	0	0,01	0	0,01	0,04
slovinština	0,02	0	0,03	0	0,10	0	0,18	3,19	0,09	0,70	94,78	0,12	0,01	0,05	0	0,10	0,07	0,04	0,04	0,08	0,04	0,02	0,04	0,05	0,02	0,02	0,03	0,03	0,04	0,01	0,02	0	0,06	0,01
angličtina	0	0	0,02	0,04	0	0,01	0,05	0,04	0,05	0,04	0,06	97,52	0,19	0,23	0,01	0,22	0,31	0,03	0,20	0,16	0,18	0,04	0,10	0,10	0,02	0	0,03	0	0,01	0,02	0	0	0,17	0,14
dánština	0,01	0	0	0	0	0	0,02	0,06	0,01	0,01	0,01	0,11	91,20	0,19	0,09	0,16	7,11	0,70	0,08	0,02	0,02	0,02	0,06	0	0	0,02	0	0	0	0,03	0,02	0	0,02	0,02
holandština	0	0	0,02	0,04	0	0,01	0,02	0,07	0,03	0,01	0,02	0,50	0,16	97,48	0,03	0,41	0,27	0,09	0,22	0,07	0,11	0,01	0,06	0,05	0,03	0	0,02	0	0,02	0,01	0	0	0,12	0,11
islandština	0	0	0,01	0	0	0	0	0,02	0	0,02	0	0,05	0,05	0,04	99,38	0,03	0,16	0,05	0,03	0,02	0	0,05	0,02	0,01	0,04	0	0	0	0	0	0,01	0	0	0
němčina	0,01	0	0	0	0	0	0,01	0,05	0,03	0,01	0,05	0,16	0,12	0,21	0,01	98,46	0,15	0,09	0,08	0,10	0,01	0,02	0,05	0,03	0,03	0	0,04	0	0,02	0,03	0,01	0,01	0,12	0,08
norština	0	0	0,01	0,04	0	0	0,03	0,09	0,02	0,04	0,05	0,28	9,06	0,13	0,13	0,23	87,79	1,33	0,09	0,09	0,10	0,05	0,03	0,04	0,07	0,03	0,03	0,01	0,01	0,02	0,02	0	0,05	0,12
švédština	0	0	0,01	0	0,01	0	0,01	0,04	0,01	0,01	0,03	0,16	0,89	0,13	0,01	0,18	1,82	96,36	0,02	0,01	0,04	0,01	0,03	0,01	0,01	0,02	0	0	0,02	0,03	0,01	0	0,04	0,07
francouzština	0	0	0	0	0	0	0	0,02	0	0,02	0,04	0,31	0,05	0,12	0	0,08	0,09	0,01	98,14	0,27	0,34	0,16	0,15	0,06	0,02	0,01	0	0	0	0	0	0,01	0,05	0,04
italština	0,02	0	0	0,06	0	0,02	0,03	0,10	0,03	0,03	0,05	0,41	0,03	0,08	0,04	0,03	0,06	0,02	0,30	96,72	0,40	0,47	0,34	0,46	0,01	0,02	0,02	0,01	0	0,03	0,03	0	0,09	0,08
katalánština	0	0	0	0	0,01	0,01	0	0,02	0,02	0,04	0,05	0,18	0,03	0,03	0,01	0,03	0,10	0,02	0,43	0,40	95,90	0,75	0,16	1,65	0	0	0,03	0	0	0,01	0,05	0	0,03	0,03
portugalština	0	0	0,01	0	0	0	0,01	0,03	0,03	0,03	0,02	0,14	0,05	0,02	0,02	0,06	0,05	0,03	0,23	0,67	0,73	95,41	0,09	2,20	0,02	0	0,03	0	0,01	0,01	0,02	0	0,03	0,04
rumunština	0,01	0,01	0,01	0,01	0,01	0	0,05	0,08	0,06	0,05	0,08	0,23	0,03	0,05	0,02	0,06	0,03	0,02	0,23	0,40	0,42	0,21	97,31	0,24	0,01	0,02	0	0	0,05	0,04	0,02	0,02	0,15	0,06
španělština	0	0	0	0,03	0	0	0,03	0,04	0,03	0,03	0,01	0,22	0	0,08	0,01	0,08	0,03	0,02	0,25	0,76	2,57	2,43	0,28	92,83	0	0,01	0,01	0,05	0,02	0,01	0,06	0	0,07	0,03
estonština	0	0	0	0	0	0,01	0,01	0	0,01	0,01	0,03	0,11	0,01	0,03	0,02	0,03	0,03	0,05	0,04	0,02	0,05	0,02	0,01	0	99,13	0,27	0	0	0,03	0,01	0	0,01	0,01	0,04
finština	0	0	0	0	0,01	0,01	0,01	0,04	0,02	0,01	0,02	0,07	0,07	0,04	0,02	0,06	0,03	0,07	0,03	0,07	0	0,02	0,02	0,02	0,27	98,90	0,01	0	0,03	0,03	0,03	0,01	0,03	0,04
maďarština	0,01	0	0	0	0,01	0	0,04	0,03	0,02	0,03	0,05	0,17	0,04	0,03	0	0,10	0,07	0,01	0,03	0,03	0,09	0,04	0	0,02	0	0,01	99,00	0,01	0,02	0,01	0,01	0	0,07	0,04
litevština	0	0	0	0	0	0	0,01	0,07	0,04	0,02	0,08	0	0,02	0	0	0,02	0,02	0,02	0,01	0,01	0,02	0,04	0,03	0,04	0,04	0	0,02	99,23	0,24	0	0,01	0	0	0
lotyština	0,01	0	0,01	0	0	0	0,04	0,07	0	0,03	0,05	0,04	0	0	0	0,05	0,01	0	0,01	0,04	0,01	0,03	0,04	0,02	0,04	0,01	0,01	0,13	99,29	0,02	0	0	0,01	0,02
albánština	0	0	0	0	0	0	0	0,06	0,01	0,02	0,04	0,02	0,01	0,04	0	0,01	0,04	0	0,03	0,05	0,09	0,04	0,03	0	0,02	0,02	0,02	0	0,02	99,31	0,01	0,02	0,04	0,04
baskičtina	0	0	0,01	0	0	0	0,04	0,01	0,03	0	0,02	0,02	0	0,01	0,01	0,05	0,08	0,05	0,06	0,05	0,08	0,08	0,05	0,11	0	0,07	0,01	0,01	0	0,01	99,02	0,03	0,03	0,05
turečtina	0	0	0	0,01	0,01	0,02	0,03	0,01	0,01	0,01	0,04	0,01	0,03	0,04	0	0,01	0,03	0,02	0,01	0,02	0	0,03	0,04	0,01	0,01	0,03	0,03	0,01	0	0,01	0,01	99,42	0,07	0,01
vietnamština	0	0	0	0	0	0	0	0	0	0	0,01	0,03	0,04	0,01	0	0,02	0	0	0,01	0	0,02	0,01	0,02	0	0	0	0	0	0	0	0	0	99,77	0,05
řečtina	0	0	0	0	0	0	0	0	0	0	0,01	0,05	0	0,01	0	0	0	0	0	0,02	0,02	0	0	0,01	0	0	0,01	0,01	0	0	0	0	0,03	99,82

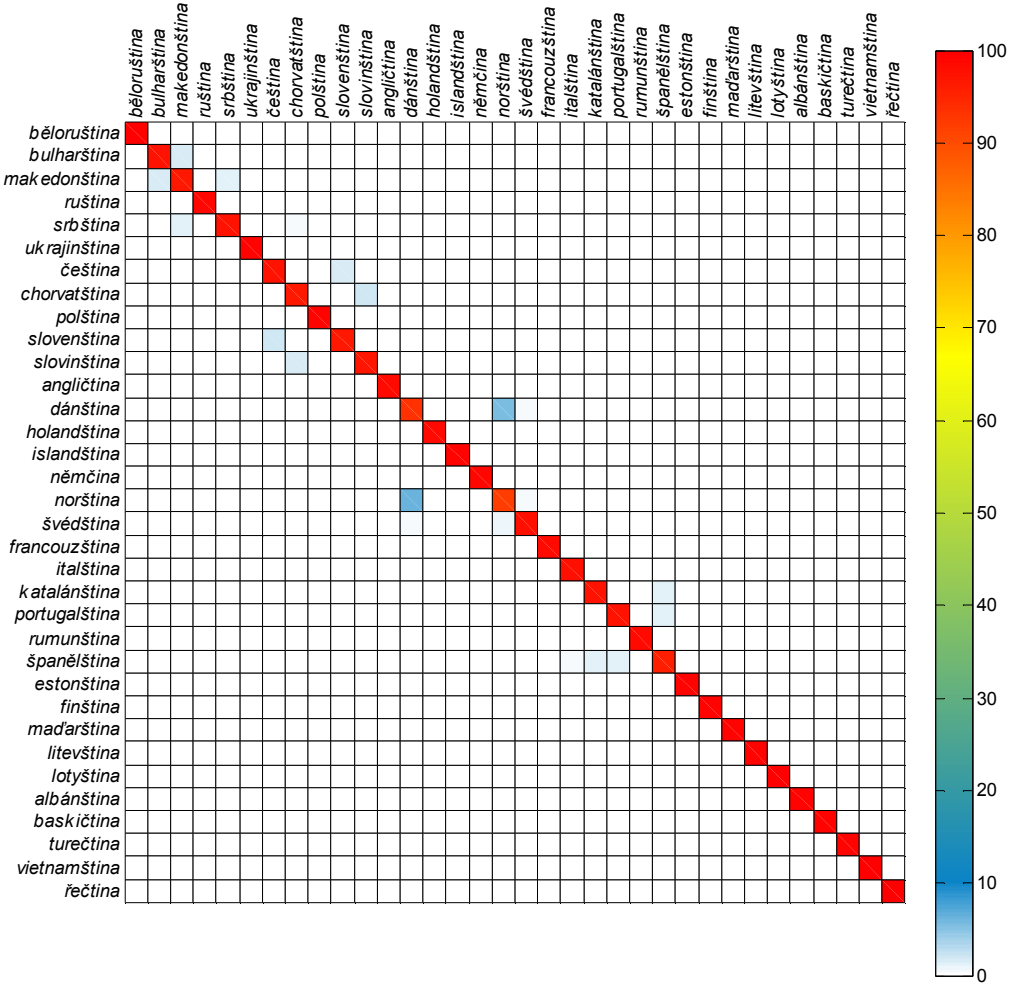
Tabulka D.12 Konfuzní matice 4-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.13 Konfuzní matice 5-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,80	0	0	0,06	0	0,03	0,01	0,01	0,02	0,01	0,01	0	0	0	0	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0,02	0	0	0	0	0,01
bulharština	0	97,23	1,85	0,42	0,30	0,13	0	0,03	0	0,02	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0,01	2,25	95,55	0,07	1,89	0,02	0	0,10	0	0	0,06	0,02	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0
ruština	0,04	0,39	0,14	98,72	0,27	0,43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,02	0,29	1,80	0,14	97,09	0,07	0	0,53	0	0	0,04	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0
ukrajinština	0,04	0,05	0,04	0,33	0,10	99,43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0,01	0	97,23	0,12	0,12	2,06	0,10	0,10	0	0,01	0	0,04	0,04	0,03	0,01	0,01	0	0,01	0,05	0	0,01	0	0,02	0	0,01	0,01	0	0	0	0
chorvatština	0,01	0,01	0,02	0,01	0,44	0	0,07	96,35	0,09	0,15	2,27	0,09	0,02	0,01	0,01	0,05	0,05	0,02	0,03	0,03	0,01	0,04	0,05	0,02	0,01	0,01	0,02	0,01	0,03	0,02	0	0,01	0,02	0,01
polština	0,05	0	0	0	0	0	0,15	0,06	99,19	0,08	0,07	0,05	0,02	0,03	0,01	0,06	0,03	0	0,06	0,02	0,02	0,01	0,01	0	0	0	0,03	0	0	0	0	0	0,01	0,03
slovenština	0,04	0,02	0,04	0	0,02	0	2,45	0,21	0,16	96,33	0,33	0,03	0,01	0,03	0	0,05	0,02	0,01	0,03	0,01	0,01	0,03	0,01	0,05	0	0	0,05	0,02	0,01	0	0	0	0	0,02
slovinština	0	0,01	0,04	0	0,05	0	0,12	2,11	0,09	0,35	96,72	0,06	0	0,04	0	0,04	0,09	0	0,03	0,03	0	0,02	0,05	0,02	0	0,02	0,02	0,02	0,02	0	0,01	0	0,03	0
angličtina	0	0	0,01	0,02	0	0,01	0,04	0,02	0,07	0,02	0,09	98,04	0,15	0,24	0,02	0,12	0,12	0,04	0,18	0,15	0,12	0,04	0,08	0,06	0,01	0	0,08	0,01	0,01	0	0,01	0,01	0,07	0,15
dánština	0,01	0	0	0	0	0	0	0,04	0	0,01	0	0,06	92,82	0,09	0,07	0,06	6,20	0,47	0,02	0,02	0,02	0,01	0	0	0,01	0,01	0,02	0	0	0	0,01	0,01	0,03	0
holandština	0	0	0,01	0,01	0	0	0	0,05	0,02	0	0,03	0,34	0,11	98,41	0,03	0,22	0,14	0,02	0,14	0,03	0,04	0,04	0,04	0,07	0	0,02	0,05	0	0,01	0	0	0	0,09	0,07
islandština	0	0	0	0	0	0	0,01	0	0,01	0	0,01	0,05	0,05	0,04	99,44	0,03	0,14	0,02	0,02	0,03	0,01	0,05	0,01	0,02	0,01	0	0,01	0	0	0	0,01	0,01	0	0,01
němčina	0	0	0	0	0	0	0,03	0,05	0,01	0,04	0,04	0,15	0,06	0,17	0	98,89	0,11	0,03	0,05	0,06	0,02	0,02	0,03	0,02	0,04	0	0,06	0	0	0,03	0	0,01	0,04	0,03
norština	0,02	0	0,02	0	0,01	0	0,01	0,05	0,03	0,01	0,05	0,26	6,76	0,13	0,10	0,12	91,04	0,90	0,04	0,06	0,03	0,03	0,06	0,04	0,06	0,02	0,05	0	0,01	0,01	0	0	0,04	0,03
švédština	0	0	0,01	0	0	0	0	0,02	0,02	0,03	0,02	0,13	0,58	0,07	0,04	0,09	1,10	97,70	0,01	0,01	0,02	0	0,01	0	0	0	0,01	0	0,02	0,02	0,01	0	0,03	0,04
francouzština	0	0,01	0	0,01	0	0	0	0,02	0,01	0,01	0,04	0,22	0,03	0,08	0,02	0,07	0,02	0	98,67	0,14	0,18	0,09	0,17	0,08	0,01	0,01	0,01	0	0,01	0,03	0,02	0	0,01	0,02
italština	0,02	0	0	0,01	0	0,02	0,03	0,04	0,04	0,02	0,06	0,36	0	0,04	0,04	0,02	0,05	0,01	0,20	97,60	0,30	0,35	0,17	0,32	0,03	0,04	0,04	0	0	0,05	0,03	0	0,04	0,06
katalánština	0	0	0	0	0,01	0,01	0	0,02	0,01	0,05	0,03	0,09	0,02	0,01	0,02	0,01	0,08	0,02	0,25	0,22	96,85	0,49	0,19	1,50	0,01	0	0,02	0	0,01	0,01	0,03	0	0,01	0,02
portugalština	0,01	0	0,01	0	0	0	0	0,03	0,02	0,02	0	0,10	0,03	0,05	0,03	0,06	0,04	0	0,15	0,42	0,43	96,90	0,06	1,45	0,03	0,01	0,03	0	0	0,03	0,01	0	0,02	0,05
rumunština	0	0	0	0,01	0	0	0,05	0,03	0,04	0,02	0,01	0,17	0,01	0,04	0,03	0,06	0,05	0	0,15	0,23	0,25	0,13	98,26	0,18	0,02	0	0,04	0,01	0,05	0,02	0,01	0,02	0,06	0,04
španělština	0	0	0	0,01	0	0	0,01	0,04	0,05	0,01	0,02	0,11	0	0,03	0,01	0,09	0,04	0	0,21	0,57	1,58	1,91	0,17	94,91	0	0,01	0,01	0,03	0,02	0,01	0,09	0	0,02	0,03
estonština	0	0	0,01	0	0	0,01	0	0,01	0,01	0	0,01	0,02	0,02	0,01	0,02	0,03	0,03	0,03	0,01	0,02	0,03	0	0	0	99,47	0,16	0,01	0	0,02	0	0,01	0,01	0,01	0,03
finština	0	0	0	0	0,01	0	0,01	0,02	0,01	0	0,02	0,04	0,02	0,04	0,04	0,04	0,06	0,04	0	0,01	0	0,01	0,03	0,02	0,10	99,34	0,02	0	0,01	0,02	0,02	0,01	0,02	0,03
maďarština	0	0	0	0	0	0	0,05	0,02	0,03	0,04	0,03	0,07	0,04	0,02	0,04	0,06	0,02	0	0,05	0,02	0,06	0,02	0,01	0,02	0	0,01	99,28	0	0,01	0,01	0	0,01	0,01	0,06
litevština	0,01	0	0	0	0	0	0	0,04	0,01	0,04	0,04	0	0,01	0	0	0,03	0,01	0	0,01	0	0,02	0,04	0,01	0,01	0,03	0	0	99,54	0,13	0	0,01	0	0	0
lotyština	0,01	0	0,02	0	0	0	0	0,05	0,03	0	0,02	0,02	0	0	0	0,06	0,01	0	0,02	0,01	0	0	0,01	0,01	0,01	0,01	0	0,05	99,62	0	0,01	0	0,01	0,01
albánština	0	0	0	0	0	0	0	0,04	0,01	0,03	0,05	0,01	0,02	0,02	0,03	0,02	0,02	0	0,01	0,05	0,02	0,01	0,04	0	0,01	0,02	0,03	0	0,01	99,50	0	0	0,03	0,01
baskičtina	0	0	0,01	0	0	0	0	0,03	0,01	0,02	0,02	0,02	0,03	0,02	0,01	0,04	0	0	0,06	0,02	0,03	0,02	0,05	0,14	0	0,02	0,02	0	0,01	0,02	99,32	0,02	0,02	0,03
turečtina	0,01	0	0	0,01	0,01	0,02	0,03	0,03	0	0,02	0,02	0,01	0,02	0,02	0	0,03	0,03	0,01	0	0,02	0,01	0,03	0,02	0	0	0,03	0,01	0	0,02	0	0	99,56	0,02	0
vietnamština	0	0	0	0	0	0	0	0	0	0,01	0	0,05	0,05	0,03	0	0,01	0	0	0,01	0,01	0,01	0,02	0,01	0,02	0	0	0	0	0	0	0	0	99,72	0,04
řečtina	0	0	0	0	0	0	0	0	0	0	0	0,08	0	0,01	0	0,01	0	0	0	0,01	0,01	0	0	0,02	0	0	0	0	0	0	0	0	0,02	99,83

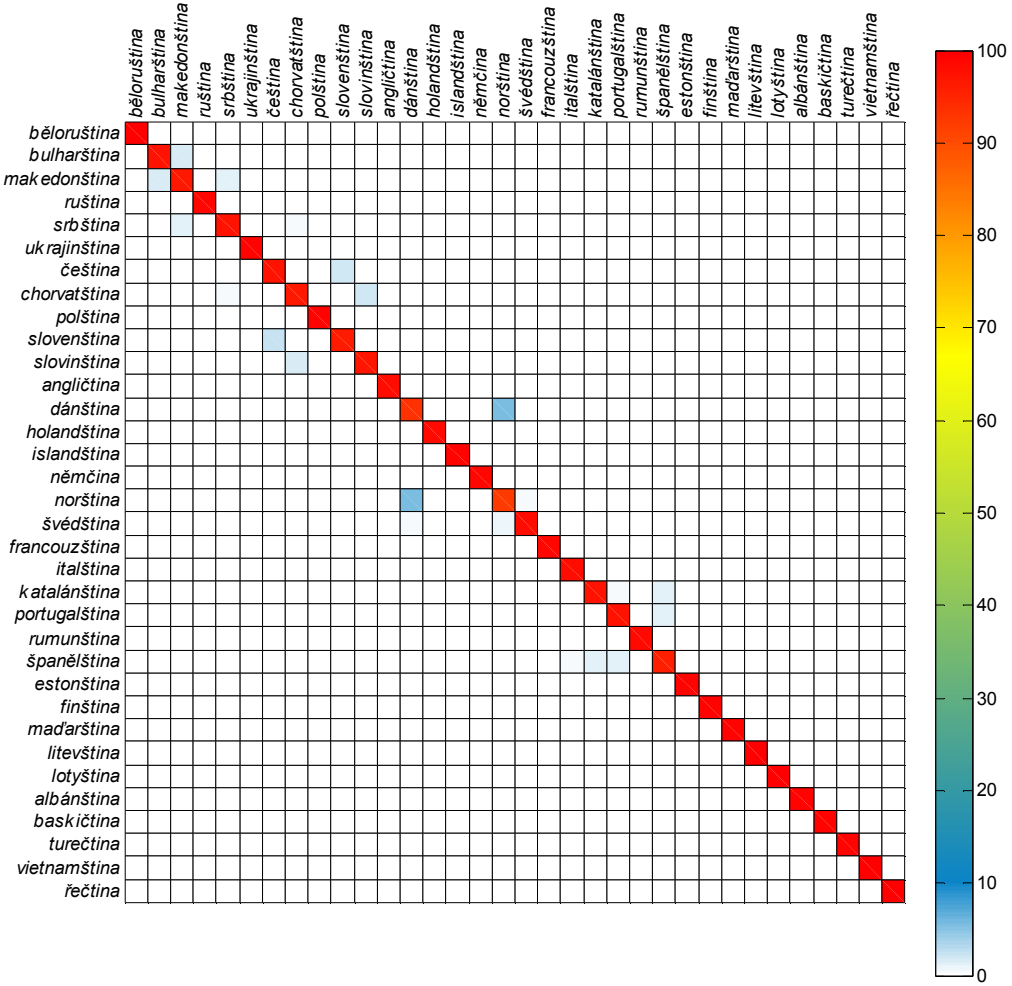
Tabulka D.13 Konfuzní matice 5-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.14 Konfuzní matice 6-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,85	0	0	0,04	0	0,03	0,01	0	0,01	0	0,01	0,01	0	0	0,01	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01
bulharština	0	97,40	1,90	0,38	0,19	0,08	0	0,01	0	0,02	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0,01	1,75	96,61	0,03	1,40	0,01	0	0,10	0	0,01	0,03	0,02	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0
ruština	0,04	0,32	0,11	98,98	0,22	0,32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,01	0,25	1,52	0,12	97,41	0,07	0	0,50	0	0	0,10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0
ukrajinština	0,04	0,06	0,03	0,35	0,06	99,45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0,01	0	97,45	0,11	0,13	1,88	0,11	0,07	0	0	0	0,04	0,02	0,02	0,01	0,01	0	0,01	0,04	0,03	0	0	0,02	0	0,02	0	0	0	0,01	0
chorvatština	0	0,01	0,03	0,01	0,32	0	0,07	96,78	0,09	0,11	2,08	0,08	0,03	0	0,01	0,04	0,04	0	0,03	0,01	0,01	0,05	0,02	0,02	0,02	0,02	0,03	0,02	0,01	0,01	0	0,01	0,02	0,01
polština	0,05	0	0	0	0	0	0,08	0,07	99,24	0,12	0,05	0,06	0,02	0,04	0,01	0,06	0,02	0	0,05	0,02	0,02	0,01	0,02	0	0	0	0,02	0	0	0,01	0	0	0	0,02
slovenština	0,03	0,05	0,01	0	0,02	0	2,26	0,15	0,18	96,54	0,35	0,04	0,01	0,02	0	0,07	0,02	0	0,04	0,01	0	0,04	0,01	0,05	0	0	0,03	0,02	0,02	0	0,01	0	0	0,01
slovinština	0	0	0,02	0	0,02	0	0,09	1,76	0,11	0,29	97,25	0,05	0	0,06	0	0,04	0,02	0,01	0,03	0,03	0,01	0,02	0,03	0,02	0	0,01	0,02	0,02	0,04	0	0,01	0	0,02	0,01
angličtina	0	0	0,01	0	0	0,01	0,03	0,04	0,07	0,01	0,08	98,44	0,09	0,22	0,02	0,09	0,12	0,06	0,12	0,13	0,07	0,04	0,08	0,06	0,01	0	0,06	0,01	0,01	0	0,01	0	0,03	0,07
dánština	0,01	0	0	0	0	0	0	0,01	0	0,01	0	0,05	93,37	0,08	0,06	0,04	5,77	0,44	0,01	0,01	0,03	0	0,01	0,01	0,01	0	0,02	0	0	0	0,01	0,02	0,02	0
holandština	0	0	0	0,01	0	0	0,02	0,03	0	0,01	0,03	0,31	0,12	98,73	0,01	0,21	0,13	0	0,11	0,02	0,03	0	0,02	0,04	0	0,01	0,01	0	0,01	0	0	0	0,07	0,06
islandština	0	0	0	0	0	0	0,01	0	0,01	0,01	0,01	0,04	0,05	0,04	99,45	0,03	0,12	0,02	0,03	0,03	0,02	0,04	0,03	0,01	0,01	0	0,01	0	0	0	0,01	0	0	0,01
němčina	0	0	0	0	0	0	0,04	0,05	0,01	0,04	0,03	0,11	0,02	0,15	0,01	99,00	0,08	0,04	0,07	0,07	0,01	0,01	0,04	0,01	0,03	0	0,08	0	0,01	0,03	0	0,01	0,02	0,02
norština	0,01	0	0,02	0	0,01	0	0,01	0,05	0,03	0,02	0,03	0,18	6,28	0,08	0,12	0,12	91,86	0,72	0,02	0,07	0,03	0,03	0,05	0,04	0,06	0,05	0,05	0	0	0,01	0	0	0,01	0,03
švédština	0	0	0,01	0	0	0	0,03	0,02	0,01	0,04	0,03	0,08	0,48	0,03	0,04	0,06	0,97	98,02	0,01	0,01	0,03	0	0,01	0	0,01	0,02	0	0	0,03	0	0,01	0	0,01	0,03
francouzština	0	0,01	0	0,01	0	0	0,01	0,02	0,01	0,01	0,02	0,22	0,03	0,04	0,03	0,06	0,05	0	98,77	0,11	0,20	0,11	0,15	0,04	0,02	0,01	0	0	0,01	0,02	0	0	0,02	0,01
italština	0,01	0	0,02	0	0	0,02	0,01	0,04	0,04	0,03	0,06	0,28	0,01	0,04	0,02	0,01	0,09	0,01	0,16	98,01	0,18	0,27	0,14	0,26	0,02	0,03	0,07	0	0	0,07	0,01	0	0,04	0,04
katalánština	0	0	0	0	0,01	0,01	0	0,06	0,01	0,04	0,05	0,07	0,02	0,01	0,02	0,01	0,05	0,02	0,20	0,20	97,27	0,39	0,12	1,31	0,01	0	0,02	0,01	0,01	0,01	0,03	0	0	0,03
portugalština	0,01	0	0,01	0	0	0	0,02	0,03	0,02	0,02	0	0,07	0,03	0,03	0,03	0,04	0,04	0	0,11	0,32	0,33	97,30	0,10	1,29	0,04	0	0,04	0	0	0,04	0	0	0,03	0,04
rumunština	0	0	0	0,01	0	0	0,06	0,03	0,04	0,06	0,02	0,12	0,01	0,04	0,02	0,05	0,03	0	0,12	0,20	0,22	0,13	98,48	0,15	0,02	0	0	0,01	0,05	0,01	0	0,01	0,06	0,04
španělština	0	0	0	0,02	0	0	0,03	0,06	0,04	0	0,01	0,07	0	0,03	0	0,06	0,01	0	0,15	0,48	1,35	1,49	0,17	95,86	0	0	0,02	0,01	0,02	0	0,07	0,01	0	0,03
estonština	0	0	0	0,01	0	0	0,01	0,02	0	0	0,02	0,06	0,01	0,03	0,03	0,03	0,06	0,02	0,02	0,03	0,01	0	0	0	99,44	0,11	0	0	0,03	0	0,01	0,01	0	0,03
finština	0,01	0	0	0	0,01	0	0,01	0,01	0,02	0	0,01	0,04	0,01	0,02	0,03	0,01	0,02	0,04	0,01	0,02	0	0,02	0	0	0,11	99,48	0,01	0	0,01	0	0,02	0,01	0,01	0,05
maďarština	0	0	0,01	0	0	0	0,04	0,02	0,05	0,03	0,02	0,09	0,02	0,02	0,03	0,06	0,03	0	0,05	0,02	0,03	0,02	0,01	0,01	0,01	0,02	99,35	0	0,01	0,02	0	0	0,01	0,01
litevština	0,02	0	0	0	0	0	0,02	0,03	0,02	0,06	0,03	0	0,01	0	0	0,01	0	0	0	0	0,01	0,03	0,01	0,03	0,02	0,01	0	99,57	0,10	0	0,01	0	0	0
lotyština	0,01	0,01	0,02	0	0	0	0	0,02	0,03	0	0,01	0,02	0	0,01	0	0,04	0,01	0	0,02	0	0	0	0,01	0	0,01	0	0,01	0,03	99,69	0	0,01	0	0,02	0,01
albánština	0	0	0	0	0	0	0	0,07	0,01	0,02	0,03	0,01	0,02	0,01	0,03	0,02	0,02	0	0,01	0,05	0,02	0,02	0,05	0	0,02	0,02	0,03	0	0	99,48	0,01	0	0,02	0,02
baskičtina	0	0	0,01	0	0	0	0,02	0	0,01	0,03	0,03	0,03	0,02	0,02	0	0,04	0,03	0	0,05	0,02	0,04	0,01	0,03	0,07	0	0,02	0,01	0	0,01	0,02	99,42	0,01	0,02	0,02
turečtina	0,01	0	0	0	0,01	0,02	0,03	0,02	0,02	0,01	0,02	0,01	0,02	0,04	0	0,01	0,02	0	0,02	0,03	0,01	0,03	0,02	0,01	0,02	0,01	0,01	0	0	0	0	99,57	0,02	0
vietnamština	0	0	0	0,01	0	0	0	0	0	0,01	0	0,06	0,06	0,01	0	0,01	0,01	0	0,01	0	0,02	0,03	0	0,02	0	0	0	0	0	0	0	0	99,69	0,05
řečtina	0	0	0	0	0	0	0	0	0	0	0	0,07	0	0,01	0	0,02	0	0	0,01	0,01	0,02	0	0	0,02	0	0	0	0	0	0	0	0	0,02	99,81

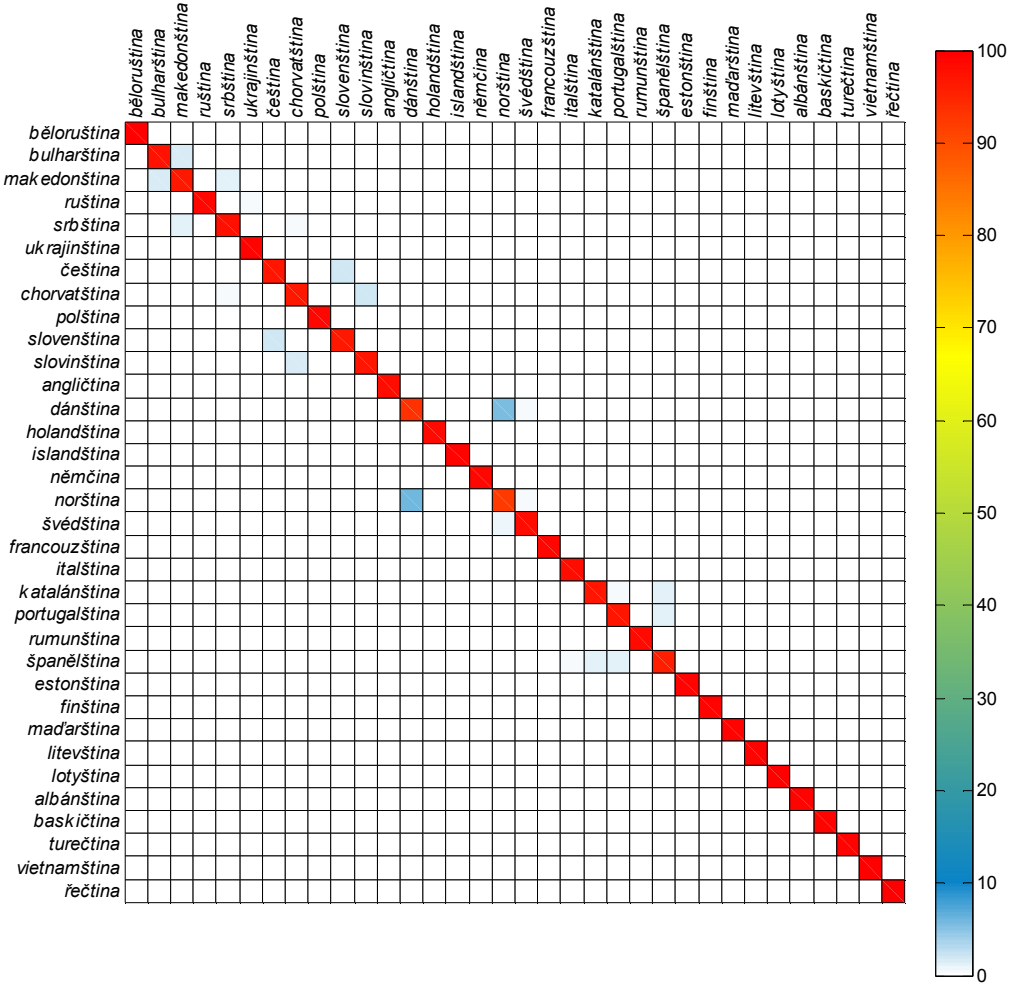
Tabulka D.14 Konfuzní matice 6-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.15 Konfuzní matice 7-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,85	0	0	0,04	0	0,02	0,01	0	0,02	0	0,01	0,01	0	0	0,01	0	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01
bulharština	0	97,52	1,80	0,37	0,17	0,11	0	0	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0,01	1,59	96,79	0,03	1,40	0,01	0	0,09	0	0	0,03	0,02	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0
ruština	0,03	0,36	0,08	98,94	0,20	0,38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,02	0,24	1,35	0,11	97,65	0,04	0	0,50	0	0	0,07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0
ukrajinština	0,05	0,06	0,04	0,35	0,07	99,42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0,01	0	97,27	0,09	0,11	2,16	0,07	0,08	0	0,01	0	0,02	0,02	0,02	0,02	0,02	0	0,01	0,03	0,02	0	0	0,02	0	0,01	0	0	0	0	0
chorvatština	0	0,01	0,03	0,01	0,44	0	0,05	96,68	0,09	0,13	2,09	0,08	0	0	0,01	0,04	0,05	0,01	0,02	0,02	0,02	0,03	0,02	0,02	0,02	0,01	0,03	0,02	0,01	0,02	0	0,01	0,02	0
polština	0,05	0	0	0	0	0	0,08	0,06	99,22	0,14	0,05	0,07	0,02	0,04	0,01	0,06	0,02	0	0,05	0,03	0,02	0	0,01	0	0	0	0,02	0	0	0,01	0	0	0,01	0,02
slovenština	0,02	0,04	0,03	0	0,03	0	2,36	0,17	0,19	96,40	0,35	0,04	0,01	0,02	0,01	0,07	0,02	0	0,03	0	0	0,04	0,01	0,06	0	0,01	0,03	0,02	0,02	0	0,01	0	0	0
slovinština	0	0	0,03	0	0,06	0	0,09	1,75	0,11	0,23	97,22	0,05	0	0,05	0	0,05	0,06	0,01	0,04	0,03	0,01	0,02	0,03	0,01	0	0,01	0,02	0,01	0,04	0	0,02	0,01	0,02	0,01
angličtina	0	0	0,01	0,01	0	0	0,02	0,02	0,05	0,03	0,09	98,34	0,09	0,20	0,01	0,13	0,16	0,05	0,12	0,17	0,06	0,06	0,10	0,08	0,01	0	0,04	0	0	0,01	0,01	0	0,02	0,10
dánština	0	0	0	0	0	0	0	0,01	0	0,01	0	0,04	93,49	0,06	0,04	0,05	5,80	0,35	0,01	0,01	0,02	0	0,04	0,01	0,01	0	0,01	0	0	0	0,01	0,01	0,01	0
holandština	0	0	0	0	0	0	0,01	0,04	0	0,01	0,03	0,28	0,12	98,81	0,01	0,16	0,14	0,01	0,11	0,02	0,02	0	0,02	0,04	0,01	0,02	0,02	0	0,01	0	0	0	0,06	0,04
islandština	0	0	0	0	0	0	0,01	0	0,01	0,02	0	0,04	0,05	0,05	99,46	0,03	0,13	0,02	0,03	0,03	0,01	0,02	0,02	0,01	0,01	0	0,02	0	0	0	0	0,01	0	0,01
němčina	0	0	0	0	0	0	0,03	0,03	0,01	0,02	0,03	0,07	0,04	0,15	0	99,20	0,04	0,02	0,05	0,05	0,01	0,02	0,03	0,01	0,02	0	0,05	0	0,01	0,03	0	0,02	0,02	0,03
norština	0,01	0	0,02	0	0,01	0	0,03	0,04	0,03	0,01	0,01	0,18	5,81	0,09	0,10	0,13	92,52	0,63	0,01	0,08	0,03	0,01	0,03	0,03	0,06	0,04	0,04	0	0,01	0	0	0	0,01	0,02
švédština	0	0	0,01	0	0	0	0,03	0,01	0,01	0,01	0,04	0,11	0,44	0,03	0,02	0,07	0,88	98,18	0,01	0,02	0,03	0	0,01	0,01	0	0,02	0	0	0,03	0	0,01	0	0	0,01
francouzština	0	0,01	0	0,01	0	0	0,01	0,03	0,01	0,01	0,02	0,23	0,01	0,04	0,03	0,04	0,06	0,01	98,87	0,08	0,22	0,05	0,12	0,07	0,02	0	0	0	0	0,01	0	0,01	0,01	0,01
italština	0,01	0	0,02	0	0	0,01	0,03	0,03	0,04	0,03	0,07	0,25	0	0,05	0,02	0,01	0,09	0	0,17	98,10	0,17	0,29	0,13	0,20	0,01	0,04	0,05	0	0	0,08	0,01	0	0,04	0,04
katalánština	0	0	0	0	0	0	0,01	0,04	0,01	0,04	0,05	0,06	0,03	0,01	0,02	0,03	0,04	0	0,15	0,17	97,27	0,40	0,17	1,37	0,01	0	0,03	0	0,01	0,02	0,03	0	0	0,02
portugalština	0,01	0	0	0	0,01	0	0,02	0,03	0,01	0,03	0,01	0,04	0,04	0,03	0	0,04	0,02	0	0,10	0,37	0,34	97,34	0,12	1,27	0,03	0	0,06	0	0	0,02	0	0,01	0,01	0,03
rumunština	0	0	0	0,01	0	0	0,06	0,03	0,05	0,04	0,04	0,13	0,02	0,04	0,01	0,06	0,04	0	0,10	0,25	0,20	0,11	98,46	0,12	0,01	0	0,01	0,01	0,06	0,02	0,01	0,01	0,07	0,02
španělština	0	0	0	0,02	0	0	0,04	0,05	0,04	0	0,02	0,07	0	0,03	0	0,04	0,02	0	0,13	0,49	1,23	1,46	0,14	96,05	0	0	0,02	0,01	0,02	0	0,08	0,01	0	0,02
estonština	0	0	0,01	0,01	0	0	0,02	0,02	0,02	0,01	0,02	0,02	0,01	0,04	0,02	0,04	0,08	0,01	0,03	0,03	0,01	0	0	0	99,40	0,11	0	0	0,03	0	0,01	0,01	0,01	0,02
finština	0,01	0	0	0	0,01	0	0,01	0,01	0,02	0	0,01	0,03	0	0,03	0,04	0,01	0,01	0,03	0,01	0,01	0	0,01	0,01	0	0,10	99,51	0,01	0	0,01	0,01	0,02	0,01	0,01	0,05
maďarština	0	0	0,01	0	0,01	0	0,04	0,02	0,05	0,03	0,04	0,09	0,01	0,01	0,02	0,06	0,02	0,01	0,05	0,01	0,03	0,01	0,02	0,02	0,02	0,01	99,32	0	0,01	0,02	0,01	0,01	0,01	0,02
litevština	0,01	0	0	0	0	0	0,02	0,02	0,01	0,05	0,04	0	0,01	0	0	0,01	0	0	0	0	0,01	0,03	0,01	0,03	0,01	0,01	0	99,62	0,09	0	0,01	0	0	0
lotyština	0,02	0,01	0,02	0	0	0	0,01	0,03	0,03	0	0,01	0,03	0	0,01	0	0,03	0,01	0	0,01	0	0	0,01	0	0	0,02	0	0,01	0,06	99,64	0	0	0	0,02	0,01
albánština	0	0	0	0	0	0	0	0,06	0	0,03	0,03	0,02	0,02	0,01	0,02	0,02	0,03	0	0,02	0,06	0,02	0,02	0,03	0	0,02	0,01	0,03	0	0	99,51	0	0	0,01	0,02
baskičtina	0	0	0,01	0	0	0	0,03	0	0,02	0,02	0,02	0,03	0,02	0,01	0	0,04	0,02	0	0,04	0,02	0,04	0,03	0,05	0,07	0	0,02	0,01	0	0,01	0,01	99,42	0,01	0,01	0,03
turečtina	0,01	0	0	0	0,01	0,01	0,03	0,01	0,02	0,02	0,03	0,01	0,02	0,03	0	0,02	0,01	0	0,02	0,02	0	0,04	0,02	0,01	0,01	0,01	0,02	0	0	0	0,01	99,57	0,02	0,01
vietnamština	0	0	0	0,01	0	0	0	0	0	0,01	0	0,06	0,05	0,02	0,01	0,01	0,01	0	0,01	0	0,03	0,01	0	0,02	0	0	0	0	0	0	0	0	99,69	0,05
řečtina	0	0	0	0	0	0	0	0	0	0	0,01	0,08	0,01	0,01	0	0,01	0	0	0,01	0,01	0,01	0	0	0,02	0	0	0	0	0	0	0	0	0,01	99,81

Tabulka D.15 Konfuzní matice 7-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků



Graf D.16 Konfuzní matice 8-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

	běloruština	bulharština	makedonština	ruština	srbština	ukrajinština	čeština	chorvatština	polština	slovenština	slovinština	angličtina	dánština	holandština	islandština	němčina	norština	švédština	francouzština	italština	katalánština	portugalština	rumunština	španělština	estonština	finština	maďarština	litevština	lotyština	albánština	baskičtina	turečtina	vietnamština	řečtina
běloruština	99,84	0	0	0,04	0	0,02	0,01	0	0,02	0	0,01	0,01	0	0	0,01	0	0,01	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0	0,01
bulharština	0	97,53	1,68	0,37	0,27	0,12	0	0	0	0,02	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
makedonština	0,01	1,64	96,76	0,03	1,38	0,01	0	0,08	0	0	0,04	0,01	0	0	0	0	0	0	0	0,01	0,01	0	0	0	0	0	0	0	0	0	0	0	0	0,01
ruština	0,04	0,33	0,07	98,92	0,17	0,46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
srbština	0,02	0,18	1,31	0,12	97,77	0,05	0	0,45	0	0	0,08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01	0	0	0	0	0	0
ukrajinština	0,04	0,06	0,03	0,32	0,06	99,48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
čeština	0	0	0	0	0,01	0	97,09	0,10	0,11	2,30	0,07	0,09	0,01	0,01	0	0,02	0,03	0,02	0,02	0,02	0	0,01	0,03	0,02	0	0	0,02	0	0,01	0	0	0	0	0
chorvatština	0	0,01	0,03	0,01	0,58	0	0,05	96,58	0,07	0,12	2,08	0,08	0,01	0	0,01	0,04	0,06	0,01	0,02	0,02	0,02	0,04	0,02	0,01	0,02	0,01	0,03	0,01	0,01	0,01	0	0,01	0,02	0
polština	0,06	0	0	0	0	0	0,11	0,06	99,19	0,12	0,06	0,07	0,02	0,04	0,01	0,06	0,02	0	0,05	0,03	0,02	0	0,01	0	0	0	0,02	0	0	0,01	0	0,01	0	0,02
slovenština	0,01	0,02	0,03	0	0,04	0	2,21	0,17	0,17	96,56	0,35	0,04	0	0,01	0,01	0,08	0,02	0	0,03	0,01	0	0,04	0,01	0,07	0	0,01	0,04	0,03	0,03	0	0	0	0	0
slovinština	0	0	0,03	0	0,10	0	0,09	1,93	0,11	0,24	96,95	0,06	0	0,04	0	0,05	0,07	0,01	0,04	0,04	0,01	0,02	0,04	0,01	0	0,01	0,02	0,02	0,03	0,01	0,02	0,01	0,02	0,01
angličtina	0	0	0,01	0,01	0	0	0,02	0,01	0,05	0,04	0,12	98,24	0,07	0,18	0,02	0,14	0,16	0,05	0,15	0,17	0,07	0,06	0,10	0,09	0	0	0,04	0	0	0,01	0,02	0	0,03	0,13
dánština	0	0	0	0	0	0	0,02	0	0	0,01	0	0,06	93,44	0,06	0,05	0,06	5,78	0,40	0,02	0,01	0,01	0,01	0,01	0,01	0	0	0,01	0	0	0	0,01	0,01	0,01	0
holandština	0	0	0,01	0	0	0	0,01	0,04	0	0	0,03	0,30	0,10	98,80	0,02	0,15	0,13	0,02	0,10	0,02	0,03	0	0,03	0,04	0,01	0,01	0,03	0	0,01	0	0	0	0,06	0,04
islandština	0	0	0	0	0	0	0,01	0	0,01	0,03	0	0,05	0,06	0,05	99,44	0,03	0,12	0,02	0,04	0,03	0,01	0,01	0,03	0,01	0,01	0	0,02	0	0	0	0	0	0	0,01
němčina	0	0	0	0	0	0	0,03	0,02	0,02	0,02	0,03	0,08	0,07	0,15	0	99,14	0,07	0,02	0,04	0,05	0,01	0,04	0,02	0,02	0,02	0	0,02	0	0	0,04	0	0,02	0,03	0,03
norština	0,01	0	0,02	0	0,01	0	0,02	0,03	0,01	0,02	0,03	0,17	5,99	0,07	0,09	0,13	92,40	0,60	0,04	0,07	0,05	0,01	0,02	0,04	0,03	0,04	0,04	0	0,01	0	0,01	0	0,01	0,02
švédština	0	0	0,01	0	0	0	0,03	0	0,01	0,02	0,02	0,09	0,39	0,02	0,02	0,06	0,83	98,32	0,01	0,02	0,03	0	0,03	0,01	0	0,02	0	0	0,03	0	0,01	0	0	0,01
francouzština	0	0,01	0	0,01	0	0	0,01	0,02	0,01	0,01	0,02	0,22	0,01	0,05	0,03	0,03	0,07	0,01	98,90	0,06	0,20	0,06	0,12	0,09	0,01	0	0	0	0	0,01	0	0,01	0,01	0,01
italština	0,01	0	0,04	0	0	0,01	0,03	0,03	0,05	0,03	0,06	0,25	0	0,04	0,02	0,01	0,11	0	0,15	98,09	0,15	0,32	0,11	0,24	0,01	0,05	0,04	0	0	0,05	0,01	0	0,04	0,04
katalánština	0	0	0	0	0	0	0	0,04	0,01	0,04	0,04	0,07	0,03	0,02	0,02	0,03	0,04	0	0,10	0,16	97,23	0,41	0,15	1,47	0,01	0	0,04	0	0,01	0	0,05	0	0	0,02
portugalština	0,01	0	0	0	0,01	0	0,02	0,03	0,01	0,03	0,01	0,07	0,05	0,03	0	0,03	0,01	0	0,11	0,35	0,35	97,22	0,12	1,36	0,02	0	0,06	0	0	0,01	0	0	0,02	0,06
rumunština	0	0	0	0,01	0	0	0,09	0,03	0,04	0,06	0,04	0,13	0,02	0,03	0,01	0,04	0,05	0	0,12	0,21	0,16	0,06	98,53	0,12	0,01	0	0,02	0,01	0,06	0,01	0,01	0,02	0,07	0,03
španělština	0	0	0	0,02	0,01	0	0,03	0,06	0,03	0	0,01	0,07	0	0,03	0	0,07	0,02	0	0,13	0,48	1,33	1,44	0,10	95,97	0	0	0,03	0,02	0,02	0,01	0,09	0,01	0	0,01
estonština	0	0	0,01	0	0	0	0,02	0,02	0,02	0	0,03	0,03	0,02	0,05	0,02	0,03	0,07	0,01	0,03	0,04	0,01	0	0	0	99,41	0,10	0	0	0,03	0	0,01	0,01	0	0,02
finština	0,01	0	0	0	0,01	0	0,01	0,01	0,02	0	0	0,04	0	0,04	0,01	0,03	0,02	0,03	0,01	0,01	0	0,01	0,01	0	0,08	99,52	0,01	0	0,01	0,01	0,02	0,01	0,01	0,05
maďarština	0	0	0,01	0	0,01	0	0,04	0,02	0,06	0,04	0,04	0,08	0,01	0,01	0,02	0,09	0,02	0	0,04	0,01	0,02	0,03	0,02	0,03	0,01	0,01	99,30	0	0,01	0,01	0,01	0,01	0,01	0,02
litevština	0,01	0	0	0	0,01	0	0,02	0,02	0,01	0,04	0,07	0	0,01	0	0	0,02	0	0	0	0	0,01	0,03	0,01	0,03	0,01	0,01	0	99,53	0,14	0	0,01	0	0	0
lotyština	0,02	0,01	0,02	0	0	0	0,01	0,03	0,03	0	0,01	0,01	0	0,02	0	0,03	0,02	0	0,01	0	0	0,01	0,01	0	0,01	0	0,01	0,06	99,64	0	0	0	0,02	0,01
albánština	0	0	0	0	0	0	0	0,06	0	0,03	0,03	0,03	0,01	0,01	0,01	0,02	0,04	0	0,01	0,04	0,03	0,03	0,03	0	0,02	0,01	0,03	0	0	99,52	0	0	0,01	0,02
baskičtina	0	0	0,01	0	0	0	0,03	0	0,02	0,02	0,02	0,02	0,03	0	0	0,04	0,02	0	0,04	0,03	0,05	0,03	0,04	0,08	0	0,02	0,01	0	0	0	99,43	0,01	0,01	0,03
turečtina	0,01	0	0	0	0,01	0,01	0,03	0,01	0,02	0,01	0,03	0,01	0,02	0,04	0	0,02	0,01	0	0,01	0,02	0	0,04	0,02	0,02	0,01	0,01	0,02	0	0	0	0,01	99,57	0,02	0,01
vietnamština	0	0	0	0,01	0	0	0	0	0	0,01	0	0,06	0,06	0,01	0,01	0,01	0,01	0	0,01	0	0,03	0,01	0	0,02	0	0	0	0	0	0,01	0	0	99,68	0,05
řečtina	0	0	0	0	0	0	0	0	0	0	0,01	0,08	0,01	0,01	0	0	0	0	0,01	0,01	0,01	0	0	0,02	0	0	0	0	0	0	0	0	0,01	99,82

Tabulka D.16 Konfuzní matice 8-gramového modelu Witten-Bell při délce testovacích řetězců 20 znaků

Příloha E – Manuály k aplikacím

Model Creator

Aplikace slouží k vytvoření jazykového modelu z trénovacího korpusu. Využívá část ngram-count projektu SRILM, která slouží k vytváření modelů v rámci projektu SRILM. Znaky trénovacího korpusu rozdělí mezerami, tak jak to vyžaduje projekt SRILM, aby pracoval se znaky. Původní mezery nahradí za podtržítko „_“, proto by se v trénovacím korpusu nemělo podtržítko objevovat.

Povinné parametry

Parametr	Popis
input	Soubor s texty daného jazyka (trénovací korpus).
output	Název výstupního souboru (model).
vocabulary	Soubor slovníku, kde každý znak je na jednom řádku. Soubor musí být v kódování UTF-8 bez BOM.

Volitelné parametry

Parametr	Popis	Defaultní hodnota
order	Stupeň modelu, který se má vytvořit.	5
encoding	Kódování vstupního souboru.	UTF-8
discounting	Použitá vyhlazovací technika. Hodnoty tohoto paremetru jsou shodné jako ve SRILM.	-wbdiscount -interpolate

Příklad

ModelCreator input Czech.txt output Czech.lm vocabulary vocabulary.txt order 3 encoding windows-1250

Language Recognizer

Aplikace slouží pro třídění textů podle jazyka. Aplikaci v parametru předáte soubor s texty, který chcete roztřídit, a řeknete, které modely má použít při identifikaci jazyka. Dále aplikaci předáte seznam oddělovačů, podle kterých má rozdělovat text na části, na kterých se bude jazyk identifikovat. Implicitně se oddělují pouze odstavce. Aplikaci lze předat ještě mnohem více parametrů, viz níže. Výsledkem jsou roztříděné texty, umístěné v jednotlivých textových souborech a označené identifikovaným modelem.

Povinné parametry

Parametr	Popis
file	Soubor s texty. Je možné zadat více souborů.
model	Model, který se použije při identifikaci jazyka. Je potřeba zadat minimálně 2 různé modely. Modely musí být ve složce „Models“, která je v hlavní složce aplikace.

Volitelné parametry

Parametr	Popis	Možné hodnoty	Defaultní hodnota
models all	Použije všechny modely ze složky „Models“, která je v hlavní složce aplikace.		
order	Stupeň n-gramů, který se použije pro identifikaci jazyka. Pokud se zvolí vyšší stupeň než mají modely, použije se nejvyšší stupeň modelů.	Celé kladné číslo	5
difference	Určuje, o kolik musí mít identifikovaný model vyšší log(p) proti modelu s druhým nejvyšším log(p) pro danou větu. Pokud je rozdíl menší než tento parametr, věta se zařadí mezi nejisté výsledky, jinak se zařadí mezi jisté výsledky.	Celé kladné číslo	0
write_incertain_results	Určuje, zda se na výstup mají zapisovat nejisté výsledky.	true/false	true
folder	Jako vstupní textové soubory použije všechny soubory ze zadaného adresáře a všech jeho podadresářů.	Textový řetězec	
encoding	Kódování vstupního textového souboru. Stejné kódování se použije pro výstupní soubory.	Textový řetězec	UTF-8
separators	Oddělovače vět v odstavci. Oddělování odstavců je automatické. Pokud se nezadají žádné oddělovače, bude text rozdělen jen podle odstavců.	Textový řetězec	
min_length	Minimální délka vět včetně oddělovače. [znaky]	Celé kladné číslo	1
ignore_next_separator	Pokud je délka věty kratší než parametr „min_lenght“, ignoruje následující oddělovač, tak aby získal delší větu.	true/false	false
split_text	Určuje, zda se má výstup rozdělit podle zadaných oddělovačů (každá věta na samostatný řádek).	true/false	false
lower_case	Převede všechny znaky rozpoznávaného textu na malá písmena. Při použití tohoto parametru by i modely měly být natrénovány na textech s malými písmeny.	true/false	false
remove_names	Určuje, zda se z rozpoznávaných vět odstraní „jména“ (slova ve větě začínající velkým písmenem, kromě prvního slova). Tento parametr má vliv pouze na texty, podle kterých se identifikuje jazyk, výstupní texty budou beze změny (se „jmény“).	true/false	false

Výstup

Výstupní soubory jsou ve stejné složce jako vstupní textový soubor. Kódování těchto souborů je podle parametru „encoding“, defaultně UTF-8. Výsledky jsou rozděleny podle parametru „difference“ na jisté a nejisté. Značení názvů souborů je následující:

Jisté výsledky:	[název souboru s texty]-[název modelu]
Nejisté výsledky:	[název souboru s texty]-[název modelu]-uncertain

Příklad

LanguageRecognizer file data.txt folder C:\Texty encoding windows-1250 model Czech model Slovak order 6 separators .?! ignore_next_separator true difference 30

Language Recognizer View

Aplikace slouží k zobrazení výsledků identifikace jazyka zvoleného textu nebo souboru. Na konsoly, případně do souboru, vypíše věty a hodnoty jejich pravděpodobností v jednotlivých modelech. Hodnota s největší pravděpodobností je pak barevně zvýrazněna (identifikovaný model jazyka).

Povinné parametry

Parametr	Popis
file	Soubor s texty, je možné zadat pouze jeden soubor.
text	Je možné zadat pouze text (v uvozovkách), na kterém se má identifikovat jazyk, pak není povinný parametr „file“
model	Model, který se použije při identifikaci jazyka a je umístěn ve složce „Models“

Volitelné parametry

Parametr	Popis	Možné hodnoty	Defaultní hodnota
models all	Použije všechny modely ze složky „Models“, která je v hlavní složce aplikace.		
order	Stupeň n-gramů, který se použije pro identifikaci jazyka. Pokud se zvolí vyšší stupeň než mají modely, použije se nejvyšší stupeň modelů.	Celé kladné číslo	5
encoding	Kódování vstupního textového souboru. Stejné kódování se použije pro výstupní soubory.	Textový řetězec	UTF-8
separators	Oddělovače vět v odstavci. Oddělování odstavců je automatické. Pokud se nezadají žádné oddělovače, bude text rozdělen jen podle odstavců.	Textový řetězec	
min_length	Minimální délka vět včetně oddělovače. [znaky]	Celé kladné číslo	1
ignore_next_separator	Pokud je délka věty kratší než parametr „min_lenght“, ignoruje následující oddělovač, tak aby získal delší větu.	true/false	False
lower_case	Převede všechny znaky rozpoznávaného textu na malá písmena. Při použití tohoto parametru by i modely měly být natrénovány na textech s malými písmeny.	true/false	False
file_out	Při nastavení se výsledky nezapisují na konsoly, ale do zadaného souboru.	Textový řetězec	

Příklad

LanguageRecognizerView file data.txt encoding windows-1250 model Czech model Slovak order 6 separators .?! ignore_next_separator true

Výstup

Zobrazení výsledků na konsoly. Vlevo jsou hodnoty log(p) zvolených modelů pro rozpoznávanou větu, která je vpravo. Nejvyšší hodnota log(p) pro danou větu je pak zvýrazněna barevně.