ŠTATISTICKÝ PRÍSTUP PRI URČOVANÍ SLABIČNÝCH HRANÍC

Jozef Ivanecký

IBM Voice Systems, European Speech Research,
Gottlieb-Deimlerstr. 12, D-69115 Mannheim
ivanecky@de.ibm.com

Technická Univerzita Košice
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie

ABSTRAKT

One of the requirements for the rule based phonetic transcription in Slovak language is knowledge about syllabic boundaries. The rule based approach in this case is not very effective. The paper describe our effort to apply a language model theory to the syllabic segmentation. We give an theoretical overview as well as description of the real experiment together with achieved results for Slovak. The results lead to conclusion, that proposed approach can overpass the limitation of the rule based systems for syllabic segmentation.

1. ÚVOD

Určovanie hraníc slabík komplikuje skutočnosť, že už samotná definícia slabiky nie je jednoznačná [12]. Na definíciu slabiky existuje niekoľko rôznych pohľadov. Ako východisko bola zvolená charakteristika E. Paulinyho vychádzajúca z jednoznačného fonologického hľadiska [9]. Presná definícia slabiky pre účely určovania hraníc pomocou navrhnutého riešenia nie je nevyhnutne dôležitá, keďže navrhnuté postupy takúto definíciu nevyžadujú.

Tak ako neexistuje presná definícia slabiky, neexistujú ani presné pravidla na určovanie hraníc slabík. Je možné zadefinovať sadu jednoduchých pravidiel na delenie slova na slabiky, avšak takéto pravidlá dosiahnu maximálne 80% úspešnosť. Pri použití zložitejších pravidiel je možné dosiahnuť vyše 85% úspešnosť. Ďalšie zvyšovanie presnosti vedie hlavne k nárastu množstva výnimiek.

Druhým problémom je skutočnosť, že pre niektoré slová existuje viacero možných delení na slabiky. To vyplýva aj z vyššie spomínanej absencie presnej definície slabiky v slovenčine. Napríklad slovo *bystrý* možno rozdeliť ako *by–strý*, *bys–trý*, alebo aj *byst–rý*. Vo všetkých týchto troch prípadoch zostáva počet slabík nemenný a vo všetkých troch prípadoch je delenie na slabiky správne.

Pri návrhu spôsobu určovania hraníc slabík bolo nutné vychádzať z nasledujúcich skutočností:

- V slovenčine neexistujú presné pravidlá na určenie slabičných hraníc.
- V niektorých prípadoch je možné delenie viacerými spôsobmi, pričom všetky sú správne.

V navrhnutom riešení je kombinované jednoduché delenie založené na pravidlách a novonavrhnutý prístup aplikujúci teóriu jazykových modelov na slabičnú segmentáciu.

2. URČOVANIE HRANÍC SLABÍK POMOCOU JEDNODUCHÝCH PRAVIDIEL

Pri hľadaní jednoduchých pravidiel na určovanie hraníc slabík sme vychádzali hlavne z [9], [12], [13]. Tieto pravidlá boli použité v nasledujúcej forme:

- Ak medzi dvoma samohláskami (dvojhláskami, samohláskou a dvojhláskou, slabičnými *r, l, ŕ, Ĺ* a samohláskou alebo dvojhláskou) je jedna spoluhláska, slabičná hranica je pred spoluhláskou, napr. *že–na, pra–co–vať, bie–ly, vl–na, vr–tieť, vŕ–ba, Sĺ–ňa–va, ka–me–nár, pa–ra–bo–la, čia–ra, bie–lia–reň, zna–me–niu.*
- Ak medzi dvoma samohláskami (dvojhláskami, samohláskou a dvojhláskou, slabičnými *r, l, ŕ, Ĺ* a samohláskou alebo dvojhláskou) je skupina dvoch spoluhlások, slabičná hranica je na rozhraní medzi obidvoma spoluhláskami, napr. *všet–ci, všet–ky, žat–va, mas–lo, lás–ka, prch–ký, mĺk–vý, maš–lič–ka, ot–cami, chlap–cami, kviet–kami, chrb–tami, pas–ca, Poprad.*

Formálne je ich možné zapísať nasledujúcim spôsobom:

$$h_{sp} \in \{Množina\ všetkých\ spoluhlások\} \quad (1)$$
$$h_{sam} \in \{Množina\ všetkých\ samohlások\} \quad (2)$$
$$h_{dif} \in \{Množina\ všetkých\ dvojhlások\} \quad (3)$$
$$h_{slsp} \in \{\acute{r}, \acute{l}\} \quad (4)$$
$$h_{slspf} \in \{r, l, \acute{r}, \acute{l}\} \quad (5)$$

$$h_{sam}h_{sp}h_{sam} \rightarrow h_{sam} - h_{sp}h_{sam} \quad (6)$$
$$h_{dif}h_{sp}h_{dif} \rightarrow h_{dif} - h_{sp}h_{dif} \quad (7)$$
$$h_{sam}h_{sp}h_{dif} \rightarrow h_{sam} - h_{sp}h_{dif} \quad (8)$$
$$h_{slsp}h_{sp}h_{sam} \rightarrow h_{slsp} - h_{sp}h_{sam} \quad (9)$$
$$h_{slsp}h_{sp}h_{dif} \rightarrow h_{slsp} - h_{sp}h_{dif} \quad (10)$$

$$h_{sam}h_{sp}h_{sp}h_{sam} \rightarrow h_{sam}h_{sp} - h_{sp}h_{sam} \quad (11)$$

$$h_{dif}h_{sp}h_{sp}h_{dif} \rightarrow h_{dif}h_{sp} - h_{sp}h_{dif} \quad (12)$$

$$h_{sam}h_{sp}h_{sp}h_{dif} \rightarrow h_{sam}h_{sp} - h_{sp}h_{dif} \quad (13)$$

$$h_{slspf}h_{sp}h_{sp}h_{sam} \rightarrow h_{slspf}h_{sp} - h_{sp}h_{sam} \quad (14)$$

$$h_{slspf}h_{sp}h_{sp}h_{dif} \rightarrow h_{slspf}h_{sp} - h_{sp}h_{dif} \quad (15)$$

kde − označuje slabičnú hranicu. V (9) a (10) sa neuvažuje slabičné *r, l* z dôvodu vyhnutia sa zložitým pravidlám. V prípade *r, l* by bolo nutné určiť, či ide o slabičné, alebo neslabičné. V (14) a (15) tento problém nevzniká, keďže v slovenčine sa nevyskytuje spoluhlásková postupnosť pozostávajúca z troch spoluhlások začínajúca neslabičným *r*. V (3) je potrebné rozlíšiť medzi dvojhláskou a samohláskovou skupinou, no keďže tento systém je súčasťou väčšieho systému na automatickú transkripciu a problém samohláskových skupín je v ňom riešený, implementácia (3) bola jednoduchá.

Slabičná segmentácia podľa týchto pravidiel je schopná správne rozdeliť približne 78% vstupných slov. Samozrejme, aj v prípade, že získane delenie je správne, nemusí to byť jediné správne riešenie. Ak sa napríklad (11) aplikuje na slovo *hospodár*, výsledok bude *hos–podár*. Je to síce správny výsledok, avšak správne delenie je aj *ho–spodár*. Pre slovo *verbovať* je však segmentácia *ver–bovať* jediná správna.

Pravidlá (6) až (15) je možné rozpísať pre dvoj až päťslabičné slova a získať tak možné kombinácie pre jednoduché postupnosti [12], avšak na naše účely je táto jednoduchá slabičná segmentácia dostačujúca.

3. ŠTATISTICKÝ PRÍSTUP PRI URČOVANÍ HRANÍC SLABÍK

Keďže z vyššie uvedeného vyplýva, že delenie slov na slabiky pomocou pravidiel je pre slovenčinu problematické, pokúsili sme sa o štatistický prístup. Všeobecne známe postupy, ktoré sa používajú pri jazykových modeloch boli modifikované pre použitie na slabičnú segmentáciu.

V prípade jazykových modelov je základnou jednotkou slovo. V tomto prípade to bude slabika. Každé slovo, ktoré bude delené na slabiky, sa najprv rozdelí na postupnosti všetkých možných slabík. Pre každú takúto postupnosť slabík **S**, kde

$$\mathbf{S} = s_1, s_2, \ldots, s_n \qquad s_i \in \xi \quad (16)$$

a ξ je množina všetkých možných slabík, je možné na základe Baysovho kritéria podmienených pravdepodobností definovať pravdepodobnosť pre každú postupnosť slabík **S** ako

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_1, \ldots, s_{i-1}) \quad (17)$$

kde $P(s_i \mid s_1, \ldots, s_{i-1})$ je pravdepodobnosť, že slabika s_i bude nasledovať po slabikách $s_1, \ldots, s_{i-1}$. Postupnosť $s_1, \ldots, s_{i-1}$ je možné označiť ako históriu h_i

a (17) zapísať do tvaru

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid h_i) \tag{18}$$

Rovnica (17) hovorí, že pravdepodobnosť danej postupnosti slabík $\mathbf{S}$, je rovná pravdepodobnosti prvej slabiky krát pravdepodobnosť druhej slabiky, za podmienky, že sa pred ňou nachádza prvá slabika, atď. krát pravdepodobnosť poslednej slabiky v prípade, že ju predchádzajú predchádzajúce slabiky.

V prípade, že sa uvažujú len dve predchádzajúce slabiky, je možné (17) zapísať v tvare

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid s_{i-2}, s_{i-1}) \tag{19}$$

Postupnosť $s_i \mid s_1, \ldots, s_{i-1}$ možno považovať za konečný automat, ktorý je v čase $i-1$ v stave Φ_{i-1} a ďalšia slabika ho zmení do stavu Φ_i. Rovnicu (17) potom možno zapísať ako

$$P(\mathbf{S}) = \prod_{i=1}^{n} P(s_i \mid \Phi_{i-1}) \tag{20}$$

Otázka znie, ako sa odhadne pravdepodobnosť $P(s_i \mid \Phi_{i-1})$. Na odhad pravdepodobnosti bol použitý "slabičný korpus" vytvorený z 10000 náhodne vybraných slovenských slov, ktoré boli rozdelené na slabiky. Každé slovo rozdelené na slabiky bolo spracované konečným automatom, ktorý kumuloval početnosť $C(s, \Phi)$, kde $C(s, \Phi)$ označuje, koľkokrát slabika s nasledovala potom, čo automat bol v stave Φ. Ak $C(\Phi)$ označuje, koľkokrát bol automat v stave Φ,

$$C(\Phi) = \sum_{s} C(s, \Phi) \tag{21}$$

potom odhad požadovanej pravdepodobnosti bude

$$P(s_i \mid \Phi_i = \Phi) = \frac{C(s_i, \Phi)}{C(\Phi)} \tag{22}$$

V prípade výpočtu pravdepodobností podľa (22) je dôležitá správne definovaná trénovacia množina, ktorá bude viesť k spoľahlivému odhadu požadovaných pravdepodobnosti.

Ako bolo uvedené vyššie, navrhnuté riešenie uvažuje len dve predchádzajúce slabiky, a preto z (19) pomocou (21) a (22) vyplýva

$$P(s_3 \mid s_1, s_2) = f(s_3 \mid s_1, s_2) \doteq \frac{C(s_1, s_2, s_3)}{C(s_1, s_2)} \tag{23}$$

kde $f(\ \mid\ )$ je funkciou početnosti výskytu.

Bohužiaľ rovnica (23) nie je vhodná na výpočet pravdepodobnosti pre danú postupnosť slabík, a to hneď z dvoch dôvodov:

- v slovenčine je bežný výskyt aj jednoslabičných a dvojslabičných slov
- nie všetky možné postupnosti slabík s_1, s_2, s_3 sa musia vyskytnúť v trénovacej množine

Na základe vyššie uvedeného je potrebné pravdepodobnosť $P(s_3 \mid s_1, s_2)$ uvažovať ako interpoláciu početnosti výskytov pre postupnosť troch, dvoch slabík a pre jednu slabiku:

$$P(s_3 \mid s_1, s_2) = \lambda_3 f(s_3 \mid s_1, s_2) + \lambda_2 f(s_3 \mid s_2) + \lambda_1 f(s_3) \tag{24}$$

Nezáporné váhy musia spĺňať podmienku $\lambda_1 + \lambda_2 + \lambda_2 = 1$. [1] Ostáva určiť optimálne hodnoty pre λ_i.

Slabičný model správajúci sa podľa (24) môže byť považovaný za skrytý Markovov model (HMM). Z počiatočného stavu $\tau_0(s_1, s_2)$ môže prejsť do jedného z troch nasledujúcich stavov $\tau_1(s_1, s_2), \tau_2(s_1, s_2), \tau_3(s_1, s_2)$ s prechodovými pravdepodobnosťami λ_1, λ_2, resp. λ_3. Z každého z týchto troch stavov je možný prechod $\mid \gamma \mid$. Každý produkuje rozdielny výstup $v \in \gamma$ a na základe neho vedie do stavu $\tau_0(s_2, v)$. v-prechody zo stavov $\tau_1(s_1, s_2)$, $\tau_2(s_1, s_2)$, $\tau_3(s_1, s_2)$ majú pravdepodobnosti $f(v), f(v \mid s_2), f(v \mid s_1, s_2)$.

Výstupné pravdepodobnosti sú v tomto prípade známe. Prechodové pravdepodobnosti je potrebné určiť. Celkový HMM pre túto situáciu je však pre slabičnú segmentáciu neakceptovateľné veľký, keďže obsahuje $4 \times \mid \gamma \mid^2$ stavov. Našťastie podľa (24) pre $i \in \{1, 2, 3\}$ sú všetky prechodové pravdepodobnosti λ_i vedúce z $\tau_0(s_1, s_2)$ do $\tau_i(s_2, v)$ rovnaké, bez ohľadu na aktuálnu kombináciu s_1, s_2, v.

Keďže (24) je HMM, odhad optimálnych hodnôt pre λ_i môžme uskutočniť pomocou Baum–Welchovho (forward–backward) algoritmu [2], [3]. [2] Predtým, než bude popísaný odhad požadovaných parametrov, je potrebné zodpovedať otázku, aké trénovacie dáta sú vhodné na odhad váh.

Z (24) vyplýva, že na odhad váh nie je možné použiť tie isté dáta, ktoré sú použité na výpočet početnosti výskytu $f(\ \mid\ )$, pretože v tomto prípade by bolo $\lambda_3 = 1$ a $\lambda_1 = \lambda_2 = 0$. Z toho vyplýva, že trénovacia množina musí byť rozdelená na dve časti. Prvá, väčšia časť dát, bude slúžiť na výpočet početnosti výskytov $f(\ \mid\ )$ a druhá, menšia časť, bude použitá na odhad váh λ_i. Samozrejme, po tomto kroku môže byť slabičný model vylepšený použitím všetkých dát na opätovný výpočet $f(\ \mid\ )$.

Táto technika je niekedy nazývaná "deleted interpolation". Kôli zjednodušeniu bola λ_i považovaná za konštantu. Je však jasné, že z $f(s_3 \mid s_1, s_2)$ je možné zrátať $P(s_3 \mid s_1, s_2)$ lepšie použitím väčšej početnosti $C(s_1, s_2)$, λ_i potom bude závisieť od podmienených početností $C(s_1, s_2)$ a $C(s_2)$. Jednoduchý spôsob, ako to dosiahnuť, je zmena štruktúry HMM, a to tak, že pred stavy τ_1 a τ_2 – nie

[1] Na určenie pravdepodobnosti nemusí byť použitá početnosť výskytu v trénovacej množine. Je to len jeden z možných prístupov. Ďalšie metódy, ktoré sa používajú pre jazykové modely možno nájsť napríklad v [4], [11].

[2] Tento postup je však v tomto prípade nepraktický a jednoduchší je priamy odhad napríklad pomocou "deleted interpolation" [1].

však pred τ_3 je vložený pomocný stav τ_4. Tým pádom prechody t_3 a t_4 vedú zo stavu $\tau_0(s_1, s_2)$ do stavov $\tau_3(s_1, s_2)$ a $\tau_4(s_1, s_2)$, a prechody t_1 a t_2 vychádzajú zo stavu $\tau_4(s_1, s_2)$ a vedú do stavov $\tau_1(s_1, s_2)$ a $\tau_2(s_1, s_2)$. Keďže bola zmenená štruktúru HMM pridaním nových prechodových pravdepodobnosti λ'_i, pre λ'_i platí

$$\lambda_1 = \lambda'_4 \times \lambda'_1\lambda_2 = \lambda'_4 \times \lambda'_2\lambda_3 = \lambda'_3 \tag{25}$$

Samozrejme platí že $\lambda'_2 = 1 - \lambda'_1$ a $\lambda'_4 = 1 - \lambda'_3$. Výhoda tohto postupu je, že $P(s_3 \mid s_1, s_2)$ sa vyváži v dvoch krokoch. Najprv sa získa

$$P^*(s_3 \mid s_2) = \lambda'_1 f(s_3) + \lambda'_2 f(s_3 \mid s_2) \tag{26}$$

a potom

$$P(s_3 \mid s_1, s_2) = \lambda'_4 P^*(s_3 \mid s_2) + \lambda'_3 f(s_3 \mid s_1, s_2) \tag{27}$$

Hodnoty λ'_i možno odhadnúť pomocou Baum–Welchovho algoritmu, avšak ako už bolo uvedené vyššie, takýto postup je v tomto prípade zbytočne komplikovaný.

Nech váhy λ sú funkciami početnosti výskytov $C(s_1, s_2)$ a $C(s_2)$. Na základe tohto (26) môže byť zmenená na

$$P^*(s_3 \mid s_2) = \xi(C(s_2)) \times f(s_3) + (1 - \xi(C(s_2))) \times f(s_3 \mid s_2) \tag{28}$$

a (27) na

$$P(s_3 \mid s_1, s_2) = \theta(C(s_1, s_2)) \times P^*(s_3 \mid s_2) + (1 - \theta(C(s_1, s_2))) \times f(s_3 \mid s_1, s_2) \tag{29}$$

kde koeficient ξ z rovnice (28) môže byť odhadnutý nezávisle na začiatku pre všetky rozdielne hodnoty $C(s_2)$. Hodnoty ξ by mali závisieť len na rozsahu, v ktorom sa vyskytuje $C(s_2)$, pretože len málo slabík s_2 bude mať vysoké hodnoty $C(s_2)$. Za týmto účelom je možné zadefinovať $\psi(s_2)$, ktoré bude označovať rozsah, do ktorého patrí početnosť $C(s_2)$. Rozsahy sú určené experimentálne, aby sa zabezpečilo, že pokrývajú dostatočné množstvo dát. Samozrejme, pre malé početnosti môže daný rozsah obsahovať len jeden výskyt a opačne.

Koeficienty ξ pre (28) boli zrátané nasledujúcim spôsobom:

1. Trénovacia množina bola rozdelená na dve časti, tak ako je opísané vyššie.
2. Zrátali sa relatívne početnosti $f(s_3 \mid s_2)$ a $f(s_3)$ z prvej časti dát.
3. Zrátal sa výskyt $N(s_2, s_3)$ bigramov s_2, s_3 v druhej časti trénovacej množiny. [3]
4. Zrátala sa hodnota ξ nájdením maxima pre

$$\sum_{N(v) \in \psi} \sum_{s_3} N(v, s_3) log[\xi \times f(s_3) + (1 - \xi) \times f(s_3 \mid v)] \tag{30}$$

[3] Početnosť v prvej časti trénovacej množiny bola označená ako $C()$ a v druhej časti ako $N()$

Hodnota $\xi = \xi(\psi)$ môže byť samozrejme určená pomocou reestimačného procesu z Baum–Welchovho algoritmu. Na určenie maximálnej hodnoty ξ však stačí zrátať prvú deriváciu z (30)

$$\sum_{N(v)\in\psi} \sum_{s_3} N(v, s_3) \left[\xi + \frac{f(s_3 \mid v)}{f(s_3) - f(s_3 \mid v)}\right]^{-1} = 0 \tag{31}$$

a zrátať hodnotu ξ. Rovnica (31) má jedno riešenie, pretože druhá derivácia z (30) je

$$- \sum_{N(v)\in\psi} \sum_{s_3} N(v, s_3) \left[\xi + \frac{f(s_3 \mid v)}{f(s_3) - f(s_3 \mid v)}\right]^{-2} \tag{32}$$

čo je záporné pre všetky hodnoty ξ. Riešenie rovnice 31 môže byť nájdené ľubovoľným vhodným prehľadávaním intervalu.

Ako bolo uvedené vyššie, tento aplikovaný postup je len jeden z mnohých možných postupov. Ďalšie postupy možno nájsť napríklad v [4], [5]. Ako však vyplýva zo získaných výsledkov, presnosť odhadu váh λ_i nie je taká kritická, ako v prípade jazykových modelov. Vyplýva to hlavne zo skutočnosti, že trénovacia množina (zastúpenie slabík) pokrýva väčšie množstvo možných vstupov[4], a z toho dôvodu je možný lepší odhad pravdepodobností $P(s_3 \mid s_2, s_1), P(s_3 \mid s_2)$ a $P(s_3)$.

4. DOSIAHNUTÉ VÝSLEDKY

Pri aplikácii navrhnutých postupov bolo na natrénovanie systému použitých 11000 náhodne zvolených slovenských slov. 10000 slov bolo použitých na výpočet pravdepodobností $P(s_3 \mid s_2, s_1), P(s_3 \mid s_2)$ a $P(s_3)$. Zvyšných 1000 bolo použitých pre výpočet váh λ zo vzťahu (24).

Pred samotným výpočtom štatistík bolo potrebné urobiť slabičnú segmentáciu trénovacej množiny. Táto segmentácia bola v prvom kole uskutočnená pomocou pravidiel uvedených v časti o delení pomocou pravidiel, čím bolo získaných približne 78% správne nasegmentovaných slabík. Ostatne problematické delenia boli urobené ručne súčasne s kontrolou správnosti automatického delenia, čo potvrdilo správnosť segmentačných pravidiel.

Priemerná dĺžka slova v trénovacej množine bola 3.374 slabiky a rozloženie monogramov, bigramov a trigramov bolo nasledujúce:

Monogramy :	3009
Bigramy :	11258
Trigramy :	9126

Počet monogramov je vlastne počet rôznych slabík v trénovacej vzorke.

Samotný proces segmentácie ľubovoľného slová na slabiky prebieha v dvoch nasledujúcich krokoch:

[4]Tu máme na mysli počet rôznych slabík, ktorý je oproti počtu rôznych slov oveľa nižší.

- Dané slovo je rozdelené na všetky možné (nemožné) postupnosti slabík (monogramov).

- Pre každú postupnosť slabík je zrátaná pravdepodobnosť podľa vzťahu (17) upraveného pre trigramový model.

Ako demonštračný príklad bolo zvolené slovo *programové*. V prvom kroku je slovo rozdelené na všetky hypotetické postupnosti slabík[5]: *prog-ra-mo-vé, prog-ra-mov-é, pro-gra-mo-vé, prog-ram-o-vé, pro-gra-mov-é, p-ro-gra-mo-vé, prog-ram-ov-é, p-ro-gra-mov-é*. Pre každé slovo je zrátaná pravdepodobnosť (17) a výsledky sú zoradené zostupne. Pre zvolený príklad je generovaný nasledujúci výstup:

prog–ra–mo–vé	9.71614053467775	e-14
prog–ra–mov–é	6.8819726337521	e-15
pro–gra–mo–vé	1.23501771678877	e-15
prog–ram–o–vé	2.39035687120406	e-17
pro–gra–mov–é	1.45480521187513	e-18
p–ro–gra–mo–vé	3.34387569322274	e-20
prog–ram–ov–é	2.98421581923104	e-21
p–ro–gra–mov–é	3.93896194381078	e-23

Pri pohľade na uvedený výstup je zrejmé, že systém generuje aj úplne nereálne postupnosti, ktoré je možné jednoducho vylúčiť pravidlami navrhnutými pravidlami. Pri aplikovaní pravidiel by v tomto prípade z ôsmich vygenerovaných hypotéz bolo 5 vylúčených a zostali by:

prog–ra–mo–vé	9.71614053467775	e-14
pro–gra–mo–vé	1.23501771678877	e-15
p–ro–gra–mo–vé	3.34387569322274	e-20

Kombinácia obidvoch postupov sa javí výhodná, čo vyplýva aj zo skutočnosti, že pravidlá pre segmentáciu síce nedokážu rozdeliť každé slovo na správnu postupnosť slabík, avšak tam, kde sa slabičná hranica pomocou týchto pravidiel určí, je hranica 100% správna[6].

V prípade skombinovania oboch navrnutých postupov je možné postupovať dvoma spôsobmi:

- Aplikovať pravidlá v prvom kroku, t. j. počas generovania všetkých možných postupností slabík.

- Aplikovať pravidla v druhom kroku, t. j. ako filter na vygenerovaný výstup.

V princípe nezáleží na tom, v ktorom kroku budú pravidlá aplikované. V prípade, že štatistická segmentácia je dostatočne spoľahlivá a pravidlá nemajú

[5] Všetky možné postupnosti slabík obsiahnutých v trénovacej množine.

[6] Samozrejme to platí len v prípade, ak nebudú použité pravidlá, ktoré môžu v určitých prípadoch generovať nesprávny výstup.

za úlohu korigovať prípadný nesprávny výstup, nie je potrebné ich uvažovať vôbec.

Z hľadiska výpočtovej náročnosti je vhodné použiť ich v prvom kroku a zredukovať tak prehľadávaný priestor pri generovaní všetkých možných slabík, čo je výpočtovo náročný krok, keďže sa generujú všetky možné postupnosti slabík pre dané slovo.

Na otestovanie spoľahlivosti celého systému bolo zvolených 203 náhodne vybratých slov, ktoré neboli súčasťou trénovacej množiny. Nad touto množinou testovacích dát bola urobená segmentácia a následne filtrácia vygenerovaných postupností slabík. Na výstupe bola sledovaná úspešnosť správne vykonanej segmentácie pre najpravdepodobnejšiu postupnosť bez filtrácie a s filtráciou. Takisto bola sledovaná početnosť výskytu správne generovanej postupnosti na druhom mieste v prípade, že prvá postupnosť nebola správna. Ďalej bola sledovaná početnosť správnej postupnosti slabík pre prípad, že prvé aj druhé delenie je správne. Nakoniec bol sledovaný počet slov, kde správne delenie nebolo na prvom alebo druhom mieste, a počet slov, kde správne delenie nemohlo byť generované v dôsledku absencie slabiky v trénovacej množine. Dosiahnuté výsledky sú uvedené v nasledujúcej tabuľke:

	Delenie bez filtrácie	**Delenie s filtráciou**
1. delenie správne	66.99 %	71.92 %
2. delenie správne	7.88 %	2.95 %
1. aj 2. delenie správne	12.80 %	12.80 %
Nesprávne delenie	1.47 %	1.47 %
Neexistujúce slabiky	8.37 %	8.37 %

Ako príklad na nesprávne delenie na prvom mieste možno uviesť slovo *paneurópske* a na slovo, z ktorého slabika sa nenachádzala v trénovacej množine možno uviesť slovo *lúčoch*. Pre dané slová boli generované nasledujúce postupnosti:

pa–neu–róp–ske	lúč–och
pa–ne–u–róp–ske	lú–čo–ch
pan–eu–róp–ske	lúč–o–ch
pan–e–u–róp–ske	
pa–ne–u–ró–pske	
pa–ne–ur–ó–pske	
pa–neu–ró–pske	
pan–e–u–ró–pske	
pan–e–ur–ó–pske	
pan–eu–ró–pske	

Z dosiahnutých výsledkov vyplýva, že počet skutočne nesprávnych delení je relatívne malý. V prípade, že za nesprávne delenia považujeme tie, pri ktorých sa správna postupnosť slabík nachádza na druhom alebo niektorom ďalšom mieste,

získame hodnoty **9.35 %** resp. **4.42 %** slov, kde sa správne delenie nenachádzalo na prvom mieste. **8.37 %** chýb tvoria v oboch prípadoch slová, z ktorých slabíky neboli súčasťou trénovacej množiny. V prípade, že takéto slová budú z testu vylúčené, po prerátaní výsledkov dospejeme k nasledujúcej tabuľke:

	Delenie bez filtrácie	**Delenie s filtráciou**
1. delenie správne	87.09 %	92.47 %

V oboch prípadoch je tu výsledok lepší ako výsledky dosiahnuté len použitím pravidiel. Ako vyplýnulo z analýzy nesprávne generovaných výsledkov, veľký rozdiel medzi filtrovaným a nefiltrovaným delením je spôsobený taktiež nedostatočnou trénovacou množinou. V prípade nesprávnych delení prevládali slová začínajúce predponou *vy-*, na ktorú sa vzťahujú pravidlá pre filtráciu. Pri zväčšovaní trénovacej množiny by mal rozdiel medzi filtrovaným a nefiltrovaným delením postupne zaniknúť.

Z výsledkov taktiež vyplýva, že 3009 rôznych slabík obsiahnutých v trenovacej množine je stále nedostatočný počet, keďže slová, ktoré neboli rozdelené z dôvodu absencie slabík, tvorili 8.37 % nesprávne nasegmentovaných slov z testovacej množiny.

Priemerný počet trénovacích slov na slabiku sú približne 3 slová. Vzhľadom k celkovému počtu slabík v sytéme — vyše 3000 — je toto číslo veľmi malé, čo oprávňuje k predpokladu, že rozšírenie trénovacej množiny povedie k ďalšiemu vylepšeniu celého systému.

5. ZÁVER

Cieľom tohto príspevku bolo overiť možnosť aplikácie teórie jazykových modelov na slabičné delenie v slovenčine. Ako ukázali dosiahnuté výsledky, tento prístup sa javí pre slovenčinu vhodný. Otvorenou otázkou zostáva veľkosť trénovacej množiny, ktorá bude pokrývať všetky slovenské slabiky a bude dostatočne veľká na odhad parametrov generujúcich správne delenie pre čo najväčší počet prípadov.

LITERATÚRA

[1] Bahl L. R., Brown P. F., de Souza P. V., Mercer R. L., Nahamoo D.: A fast algorithm for deleted interpolation. *Proceedings of Eurospeech 91*, pp. 1209–12, Genova, Italy, September 1991.

[2] Baum L.: An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, vol. 3, pp. 1–8, 1972.

[3] Baum L. E., Petrie T.: Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, vol. 37, pp. 1559–63, 1966.

[4] Jelinek F.: *Statistical Methods for Speech Recognition*, The MIT Press, Cambridge, Massachusetts, London 1998, 283 s.

[5] Katz S.: Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE TRansactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 400–01, March 1987.

[6] Kráľ A.: *Pravidlá slovenskej výslovnosti*, Slovenské pedagogické nakladateľstvo, Bratislava 1983, 632 s.

[7] Kráľ A., Sabol J.: *Fonetika a fonológia*, Slovenské pedagogické nakladateľstvo, Bratislava 1987, 392 s.

[8] Mistrík J.: *Frekvencia tvarov a konštrukcií v slovenčine*, Vydavateľstvo VEDA, Bratislava 1985, 320 s.

[9] Pauliny E.: *Fonológia spisovnej slovenčiny*, Slovenské pedagogické nakladateľstvo, Bratislava 1968

[10] Pauliny E.: *Slovenská fonológia*, Slovenské pedagogické nakladateľstvo, Bratislava 1979, 213 s.

[11] Rabiner L., Juang B.-H.:*Fundamental of Speech Recognition*, Prentice Hall, New Jersey, 1993, 507 s.

[12] Sabol J.: Slovenská slabika (Náčrt problematiky). *Studia Academica Slovaca 23, Prednášky XXX. letného seminára slovenského jazyka a kultúry*, STIMUL, Bratislava 1994.

[13] *Pravidlá slovenského pravopisu*, VEDA, Vydavateľstvo Slovenskej akadémie vied, Bratislava 1998.