Jezikovno neodvisno modeliranje pregibnega jezika

Mirjam Sepesy Maučec, Tomaž Rotovnik, Bogomir Horvat

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Smetanova 17, 2000 Maribor
{mirjam.sepesy, tomaz.rotovnik, bogo.horvat}@uni-mb.si

Abstract

This article concerns statistical language modelling of Slovenian language for automatic speech recognition. We investigate various techniques for overcoming the difficulties in modelling highly inflected languages. Slavic languages are particularly challenging languages and Slovenian language is one of them. Two main problems arise when modelling Slovenian language in comparison to English. The first problem is high OOV rate, caused by high number of inflected word forms. The second problem is a relatively relaxed word order. Slovenian language uses inflections to do what English does by word order. This presentation discusses how to alleviate the problems by choosing different basic units. We discuss the ability of two different models. The first model uses sub-word units and the second classes of words. Sub-word units are chosen to be stems and endings. Class model groups together inflections and derivations of the same root by using approximate string matching and hierarchical clustering. Only data-driven methods are employed and no linguistic knowledge is used. We show that new models improve OOV (Out-Of-Vocabulary) rate, perplexity and speech recognition performance.

1. Uvod

Vsi sistemi, ki vključujejo uporabo jezika, imajo znanje o jeziku predstavljeno v obliki modela jezika. Čeprav se v pričujočem članku omejujemo le na uporabo jezikovnega modela v razpoznavalniku govora, so modeli splošni in uporabni tudi v drugih sistemih, ki uporabljajo naravni jezik.

Z modelom jezika želimo opisati značilnosti jezika. Opis jezika naj bi bil čim bolj popoln in splošen. Predvsem pri aplikacijah, ki ne določajo posebnih lastnosti uporabljenega jezika, so se uveljavili statistični modeli, ki temeljijo na statistični analizi obsežnih besedilnih zbirk. V preteklosti je bilo opravljenih veliko raziskav na področju modeliranja angleškega jezika. Slovanski jeziki so bili dolgo zapostavljeni in so postali zanimivi šele v zadnjih nekaj letih (Whittaker, 2000; Ircing, Psutka, 2001; Byrne et al., 2001; Kneissler, Klakow, 2001). Raziskave so pokazale, da bogata slovnica slovanskih jezikov poraja številne nove probleme, ki jih pri modeliranju angleškega jezika ne zasledimo (Kaiser et al., 2000; Maučec, Kačič, 2001). Čeprav bomo v nadaljevanju obravnavali slovenski jezik, se ne bomo poglabljali v njegove posebnosti. Predstavili bomo le značilnosti, ki so bolj ali manj skupne vsem slovanskim jezikom.

Pri modeliranju povzročata največ težav:

- pregibanje besed in
- svoboda pri izbiri vrstnega reda besed.

1.1. Pregibanje

Pregibno oblikoslovje slovenskega jezika je razmeroma zapleteno. Slovenski jezik ima bogat nabor končnic. Določanje končnice izbrane besedne oblike ni enolično. Odvisno je od oblikoslovnih značilnosti osnovne oblike besede in od oblikoslovnih značilnosti besed v soseščini (tj. besed pred in za izbrano besedo v stavku) (Mladenić, 2001). Besedne oblike se, razen v končnicah, lahko razlikujejo tudi v zapisu osnove besede (npr. premene morfemov).

Kot osnovna enota jezikovnih modelov se je uveljavila beseda (Jelinek, 1997), kar pomeni, da je vsaka besedna oblika samostojna enota v slovarju. Posledica so veliki slovarji besed, potrebni za enako pokritost besedišča, kot jo na primer dosežemo pri angleškem jeziku. Eksperimenti so pokazali, da bi slovanski jeziki za enako pokritost potrebovali desetkrat večji slovar (Whittaker, 2000). Žal velikosti slovarja ne moremo poljubno izbirati, saj jo običajno omejuje zasnova sistema, v našem primeru razpoznavalnika. Danes lahko razpoznavalniki uporabljajo slovarje velikosti od 20,000 do 64,000 besed. Posledica te omejitve je velik delež neznanih besed v novih besedilih (ang. OOV rate) (Whittaker, 2000; Kaiser et al., 2000).

1.2. Vrstni red besed

Domnevamo, da vrstni red besed v slovenskem jeziku ni tako pomemben kot pri angleškem jeziku. Kot primer navedimo vseh šest permutacij besed preprostega stavka: *Maja študira angleščino* (tabela 1). Prvi stavek vsebuje nevtralen besedni vrstni red. Drugi stavek je primer poudarjanja jedra povedi. Tretji primer je običajni vprašalni stavek. Četrti stavek kaže uporabo dostavka. Peti stavek je primer preskoka ali anakoluta. Šesti stavek je s stališča zbornega jezika nepravilen, vendar se v pogovornem jeziku lahko pojavi. S spreminjanjem vrstega reda besed v slovenskem jeziku lahko spreminjamo pomen povedi.

V nasprotju s slovenščino angleščina ne dopušča tolikšne svobode pri izbiri vrstnega reda besed. Upravičeno lahko trdimo, da angleške besede vsebujejo manj slovnične informacije. Le-ta je skrita v vrstnem redu besed, ki je v angleškem jeziku bolj določen.

Statistični modeli določajo ocene verjetnosti nizov besed. V splošnem jih lahko uporabimo za modeliranje nizov poljubnih enot, vendar se zaenkrat omejimo le na besede. Izhodišče so relativne frekvence besed in besednih nizov v besedilih. Relativne frekvence so tem bolj zanesljive, čim večja je besedilna zbirka.

Danes obstajajo relativno velike besedilne zbirke za pisani jezik. Ker želimo jezikovni model vključiti v razpoznavalnik govora, bi potrebovali besedilno zbirko govorjenega jezika, ki pa v dovolj velikem obsegu žal ne obstaja. Zadovoljiti se moramo s kompromisom, da bo jezikovni

Slovenski jezik		Angleški jezik	
Stavek	Pravilnost	Stavek	Pravilnost
Maja študira angleščino.	$\surd$	Maja studies English.	$\surd$
Angleščino študira Maja.	$\surd$	English studies Maja.	$\times$
Študira Maja angleščino?	$\surd$	Studies Maja English?	$\times$
Študira angleščino, Maja ?	$\surd$	Studies English Maja.	$\times$
Maja, angleščino študira.	$\surd$	Maja English studies.	$\times$
Angleščino Maja študira.	$\surd$	English Maja studies.	$\times$

Tabela 1: Permutacije besed v slovenščini in angleščini.

model opisoval pisani jezik, čeprav ga bomo uporabili pri razpoznavanju govora.

Navedli smo dve osnovni značilnosti slovanskih jezikov, po katerih se ti razlikujejo od angleškega jezika in ki povzročata največ težav pri statističnem modeliranju. V nadaljevanju bomo predstavili metode, s katerimi skušamo reševati problem pregibanja besed. Nedoločenost vrstnega reda besed pa žal še vedno predstavlja pereč problem. Za reševanje tega problema bi potrebovali korpus besedil govorjenega jezika.

2. Model osnova – končnica

Iz oblikoslovja slovenskega jezika je razvidno, da prevladujejo pregibne besedne vrste, pri katerih ločimo dva dela: osnovo in končnico. Osnova je tisti del besede, ki ima predmetni pomen, končnica pa tisti, ki zaznamuje slovnične lastnosti besede. V zameno za besedo predlagamo dve manjši osnovni enoti modeliranja, ki izhajata iz razčlenitve besede w na osnovo s in končnico e:

$$w = s\,e \qquad s \in \mathcal{S} \ \wedge \ e \in \mathcal{E}. \tag{1}$$

Osnove s pripadajo slovarju osnov $\mathcal{S}$, končnice e pa slovarju končnic $\mathcal{E}$. Slovar končnic $\mathcal{E}$ vsebuje tudi prazno končnico (na primer: lep-$\varnothing$, jaz-$\varnothing$).

Verjetnost niza besed $W = w_1 \ldots w_n$ ocenimo s trigramskim modelom, imenujmo ga model osnovam – končnica:

$$\begin{aligned} P(W) &= \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1}) = \\ &= \prod_{i=1}^{n} P(s_i e_i|s_{i-2}e_{i-2}, s_{i-1}e_{i-1}) = \\ &= \prod_{i=1}^{n} P_S(s_i) \cdot (\lambda P_E(e_i) + (1-\lambda) P_{ES}(e_i)). \end{aligned} \tag{2}$$

$$\begin{aligned} P_S(s_i) &= P(s_i|s_{i-2}, s_{i-1}), \\ P_E(e_i) &= P(e_i|e_{i-2}, e_{i-1}), \\ P_{ES}(e_i) &= P(e_i|s_i). \end{aligned} \tag{3}$$

$P_S(s_i)$ označuje trigramski model osnov. $P_E(e_i)$ označuje trigramski model končnic. Ker ima vsaka osnova besede omejen nabor končnic, smo dodali tudi bigramski model osnova – končnica, ki smo ga označili s $P_{ES}(e_i)$. λ je interpolacijska utež in določa razmerje med trigramskim modelom končnic in bigramskim modelom osnova – končnica.

Največja prednost modela osnova – končnica je modeliranje novih besednih oblik. Z modelom lahko ocenimo verjetnost nove besedne oblike ($w_i = s_i\,e_i$), če je njena osnova v slovarju osnov ($s_i \in \mathcal{S}$) in njena končnica v slovarju končnic ($e_i \in \mathcal{E}$). Če bi razčlenjali besede na poljubno število podenot in ne le na osnove in končnice, bi teoretično odprli možnost modeliranja vseh besednih oblik. Slovarju osnovnih enot bi bilo potrebno dodati le vse črke abecede. Čeprav je ideja na prvi pogled privlačna, poruši osnovni model odvisnosti, kjer so osnove besed nosilci informacije o pomenu besede, končnice besed pa o pregibanju.

Model osnova – končnica predpostavlja, da poznamo pravila razčlenjanja besed na osnove in končnice. Za pravilno razčlenjanje bi potrebovali morfološki leksikon (Mladenić, 2001). Da ohranimo jezikovno neodvisnost predlaganih metod, predlagamo podatkovno vodeno razčlenjanje, ki sestoji iz dveh korakov. V prvem koraku tvorimo seznam vseh možnih končnic. Pri tem smo uporabili metodo krnjenja, ki smo jo povzeli po (Dimec et al., 1998; Popovič, 1992). V drugem koraku razčlenimo besede učnega besedilnega korpusa. Poudarimo, da smo razčlenjanje dodatno poenostavili tako, da smo izločili pravilo za selektivno pretvorbo soglasniških parov v posamezne soglasnike. S tem smo zagotovili, da lahko z enostavnim zlepljanjem osnov (tj. krnov) in končnic tvorimo izvorne besedne oblike. Kvaliteta krnjenja se je s tem žal zmanjšala, saj besednih oblik, pri katerih je na primer prišlo do premen morfemov, ne poenotimo na skupni krn. V postopku razčlenjanja smo uporabljali dva kriterija, najmanjšo dovoljeno dolžino krna in največjo dovoljeno dolžino končnice. Kriterij najmanjše dovoljene dolžine krna je pomemben predvsem v drugem koraku razčlenjanja besed, ker vsaj delno preprečuje premočno krnjenje. S kriterijem največje dovoljene dolžine končnice smo izločili končnice, ki bi nastale kot rezultat primerjav sestavljenih besed.

3. Model z razredi besed

Osnovna enota modeliranja je lahko tudi razred besed. V razrede združujemo besede, ki imajo podobne lastnosti (npr. POS razredi, semantični razredi),

Predpostavimo, da poznamo funkcijo, ki vsako besedo slovarja preslika v razred:

$$C : w_i \longrightarrow c_i = C(w_i). \tag{4}$$

Verjetnost niza besed $W = w_1 \ldots w_n$ ocenimo s trigram-

skim modelom z razredi:

$$P(W) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1}) = \\ = \prod_{i=1}^{n} P_0(w_i) \cdot P_C(c_i). \tag{5}$$

$$P_0(w_i) = P(w_i|C(w_i)) \\ P_C(c_i) = P(C(w_i)|C(w_{i-2}), C(w_{i-1})). \tag{6}$$

$P_C(c_i))$ označuje trigramski model razredov in $P_0(w_i)$ model pripadnosti besed razredom.

Razrede tvorimo s podatkovno vodenim združevanjem besed, ki sestoji iz dveh korakov. Najprej definiramo metriko, ki vsakemu paru besed priredi mero podobnosti. V drugem koraku besede korakoma združujemo v razrede, upoštevajoč vrednosti metrike podobnosti.

3.1. Določanje podobnosti med besedami

V razrede želimo združevati besede, ki imajo isti koren. Definirajmo metriko podobnosti sim:

$$sim : (w_1, w_2) \longrightarrow [0, 1] \quad \forall w_1, w_2 \in \mathcal{V}. \tag{7}$$

Naj velja, da je $sim(w_1, w_2) = 1$ le, če je $w_1 = w_2$.

Definirajmo dve pravili. Prvo pravilo pravi, da sta besedi zelo podobni, če se razlikujeta le v končnici. Drugo pravilo pa pravi, da sta besedi tem bolj podobni, čim manjša je Levenshteinova razdalja med njunima osnovama. Z drugim pravilom rešujemo napake krnjenja. Če pravili združimo, dobimo

$$sim(w_1, w_2) = \begin{cases} 0.99; & \text{če } (s_1 = s_2) \wedge (e_1, e_2 \in \mathcal{E}) \\ 1 - \frac{lev(s_1, s_2)}{\max(|s_1|, |s_2|)}; & \text{sicer.} \end{cases} \tag{8}$$

$lev(s_1, s_2)$ označuje Levenshteinovo razdaljo med s_1 in s_2. $|s_1|$ je dolžina osnove s_1.

3.2. Združevanje besed

Besede združujemo v razrede s hierarhičnim združevanjem. V začetku definira vsaka beseda svoj razred. Razrede nato korakoma združujemo. V vsakem koraku združimo tista dva razreda, ki sta najbolj podobna. Podobnost med razredoma R_1 in R_2 definirajmo kot:

$$sim(R_1, R_2) = \frac{1}{|R_1||R_2|} \sum_{w_1 \in R_1} \sum_{w_2 \in R_2} sim(w_1, w_2). \tag{9}$$

$|R_1|$ označuje število besed v razredu R_1. Združevanje prekinemo, ko podobnost med poljubnima razredoma ne preseže vnaprej določene pragovne vrednosti.

Števila razredov ne moremo poljubno izbirati. Nanj lahko le posredno vplivamo z izbiro praga.

4. Eksperimenti

Pri eksperimentih smo uporabili besedilni korpus *Večer*. To je zbirka člankov dnevnega časopisa Večer iz obdobja od leta 1998 do 2000. Korpus obsega 60 M besed. V njem se pojavlja približno 600,000 različnih besed. V slovar smo vključili 20,000 najpogostejših besed v korpusu.

Jezikovni model smo preizkušali na testni množici stavkov govornega korpusa SNABI (Kačič et al., 2000), saj je naš osnovni namen uporabiti jezikovni model pri razpoznavanju tekočega slovenskega govora. Testna množice vsebuje 780 stavkov, ki imajo skupno 8,000 besed.

Najprej smo zgradili referenčna jezikovna modela (bigramski in trigramski model), ki kot osnovno enoto modeliranja uporabljata besedo. Uporabili smo odprti slovar, kar pomeni, da smo vse neznane besede preslikali v simbol <OOV> in ga modelirali enako kot ostale besede v slovarju. Uporabili smo glajenje frekvenc po S. Katz (Katz, 1987).

Rezultati so zbrani v tabeli 2.

	Bigramski model	Trigramski model
Perpleksnost	788	652
OOV	18.21%	18.21%
Zadetki 3-gramov	-	65.58%
Zadetki 2-gramov	91.27%	25.78%
Natančnost razpoznavanja	41.62%	44.16%
Čas razpoznavanja	9.66×RT	15.18×RT

Tabela 2: Rezultati uporabe besednih modelov

Razloga za veliki perpleksnosti modelov sta dva. Prvi je slovnica jezika, drugi pa (ne)podobnost učnega in testnega korpusa. Učni korpus predstavlja vzorec publicističnega jezika, testni korpus pa vsebuje stavke različnih funkcijskih zvrsti jezika: praktično sporazumevalni jezik, strokovni, umetnostni jezik ipd. Perpleksnost trigramskega modela je za 17% boljša od bigramskega, a je žal trigramski model prostorsko bistveno potratnejši od bigramskega. Le dve tretjini trigramov testnega vzorca se pojavita že v učnem korpusu. Verjetnosti 9% besed testnega korpusa so ocenjene z uteženimi verjetnostmi unigramov. Iz deleža OOV razberemo, da sta v povprečju med 10 besedami 2 neznani.

Jezikovna modela smo preizkusili v enoprehodnem in dvoprehodnem razpoznavalniku (Rotovnik et al., 2002). Referenčni sistem z bigramskim modelom uporablja en prehod. Dekoder referenčnega sistema s trigramskim modelom je dvoprehoden. V prvem prehodu je uporabljen bigramski model, v drugem pa trigramski. Drugi referenčni sistem je bistveno kompleksnejši, kar se odraža v času razpoznavanja, ki smo ga zapisali v odvisnosti od realnega časa (ang. R*eal* T*ime*). Z dvoprehodnim razpoznavalnikom dosežemo 2.54% (absolutno) izboljšano natančnost razpoznavanja, vendar za 5.52×RT (absolutno) daljši čas razpoznavanja. Na natančnost razpoznavanja močno vplivata velika perpleksnost in velik delež neznanih besed (OOV).

V drugem eksperimentu smo zgradili model osnova – končnica. Najprej smo določili seznam končnic. Generirali smo več seznamov, tako da smo spreminjali kriterija

najmanjše dovoljene dolžine osnove in največje dovoljene dolžine končnice. Najboljšo pokritost učnega korpusa smo dosegli z razčlenjanjem s seznamom 3,225 končnic in vrednostjo najmanjše dovoljene dolžine osnove 3 in največje dovoljene dolžine končnice 8. Zanimiv je podatek, da se le 40% končnic ujema s končnicani iz (Dimec et al., 1998).

Slovar 20,000 besed smo razčlenili v 8497 različnih manjših enot (1661 končnic in 6836 osnov) in jih uporabili kot slovar. Velikost slovarja se je v primerjavi s slovarjem besed zmanjšala za 57%.

Interpolacijsko utež $\lambda = 0.1$, ki določa razmerje med trigramskim modelom končnic in bigramskim modelom osnova – končnica, smo določili z EM (ang. *Expectation-Maximisation*) algoritmom. Rezultati modela osnova – končnica v dvoprehodnem razpoznavalniku so zbrani v tabeli 3.

	Model osnova-končnica
Perpleksnost	112
OOV	6.52 %
Zadetki 3-gramov	82.48%
Natančnost razpoznavanja	48.50 %
Čas razpoznavanja	10.40×RT

Tabela 3: Rezultati uporabe modela osnova-končnica

Perpleksnosti modela osnova – končnica ne moremo primerjati z besednim modelom, saj uporabljamo različne osnovne enote modeliranja. Največja odlika modela je za 64% izboljšani delež OOV. Jezikovni model smo vključili v dvoprehodni razpoznavalnik. Dosegli smo 4.34% (absolutno) izboljšanje razpoznavanja primerjavi z razpoznavalnikom, ki uporablja besedne modele. Tudi čas razpoznavanja je za 4.78×RT (absolutno) krajši.

V tretjem eksperimentu smo zgradili model z razredi besed. V razrede smo razvrščali le besede slovarja (tj. 20,000 besed). Najprej smo določili podobnosti med besedami. Vsako besedo smo primerjali z vsemi ostalimi. Ohranili smo le tiste pare besed, katerih podobnost (glej enačbo (8)) je bila večja kot 0.5. Podobnostim ostalih parov smo priredili vrednost 0. V naslednjem koraku smo tvorili razrede s hierarhičnim združevanjem. Združevanje smo prekinili, ko podobnosti med razredi niso presegale pragovne vrednosti 0.7. Kot rezultat smo dobili $7,200$ razredov. Nekaj naključno izbranih razredov je v tabeli 4. Iz primerov je razvidna osnovna pomanjkljivost primerjanja le na osnovi ujemanja podnizov (npr. besedi vzor in vzorci sta v istem razredu). Rezultati modela z razredi so podani v tabeli 5. Perpleksnosti modela nismo bistveno izboljšali (če primerjamo z besednim trigramskim modelom). Opazno pa je izboljšanje zadetkov trigramov (za 12.94% absolutno). Modeli z razredi so v fazi razvoja, zato rezultatov razpoznavanja še nimamo.

zakon zakonu zakonov zakonom zakonodajo zakonodaji zakonodaje zakonodaja zakonitosti zakonito zakonih zakoni zakone zakona
nima nimajo nimam nimamo nimaš nimata nimate nisem niso nisi nista niste
vzgoja vzgojno vzgoji vzgoje vzgojiteljice vzgojo
vzor vzoru vzorno vzorce vzorcev vzorci vzorcu
čuden čudež čudi čudovito
vodnar vodne vodni vodnih vodnik vodnika vodniki vodnikov vodno
cen cena cenah cenami cene ceneje cenejša cenejše cenejši ceni cenijo cenikov ceno cenovno

Tabela 4: Naključno izbrani razredi.

	Model z razredi besed
Perpleksnost	645
OOV	18.21%
Zadetki 3-gramov	78.52%

Tabela 5: Rezultati uporabe modela z razredi besed

5. Zaključek

Eksperimenti so pokazali, da standardni besedni modeli pri modeliranju pregibnih jezikov ne dajejo zadovoljivih rezultatov. Izpostavili smo bistvene razlike med angleškim in slovenskim jezikom, predvsem tiste, ki povzročajo največ težav pri verjetnostnem modeliranju jezika. Predlagali smo jezikovno neodvisne pristope reševanja problema pregibnosti jezika, ki so se pokazali kot uspešni pri modeliranju slovenskega jezika. Zanimiva bi bila primerjava z jezikovno odvisnimi metodami, ki zaradi (zaenkrat še) nepopolne infrastrukture ni mogoča.

Z modelom osnova – končnica smo rešili problem velikega deleža neznanih besed v novih besedilih. Model z razredi ne zmanjša deleža neznanih besed. Z njim omilimo problem razpršenosti podatkov.

V prihodnosti bomo vse tri jezikovne modele vključili v razpoznavalnik tekočega govora, še prej pa bomo skušali izboljšati razrede tretjega modela. Razen ujemanja v podnizih črk bomo pri ugotavljanju podobnosti med besedami upoštevali tudi daljnosežne korelacije v besedilih.

6. Literatura

W. Byrne, J. Hajič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, J. Psutka. 2001. On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech. In *In Proceedings of the Eurospeech*.

J. Dimec, S. Džeroski, L. Todorovski, D. Hristovski. 1998. Iskalnik za slovenske in angleške dokumente na svetovnem spletu. *Zbornik konference Jezikovne tehnologije za slovenski jezik*.

P. Ircing, J. Psutka. 2001. Two-Pass Recognition of Czech Speech Using Adaptive Vocabulary. In V. Matoušek, editor, *In Proceedings of the Text, Speech and Dialogue Workshop*, pages 273–277. Springer.

F. Jelinek. 1997. *Statistical Methods for Speech Recognition*. The MIT Press.

Z. Kačič, B. Horvat, A. Zogling. 2000. Issues in design and

collection of large telephone speech corpus for Slovenian language. *In Proceedings of the LREC*.

J. Kaiser, M. Sepesy Maučec, Z. Kačič, B. Horvat. 2000. Razpoznavanje tekočega slovenskega govora z velikim slovarjem. *Zbornik konference Jezikovne tehnologije*, Ljubljana. Insitut Jožef Stefan.

S. Katz. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

J. Kneissler, D. Klakow. 2001. Speech recognition for huge vocabularies by using sub-word units. In *In Proceedings of the Eurospeech*.

M. S. Maučec, Z. Kačič. 2001. Topic Detection for Language Model Adaptation of Highly-Inflected Languages by Using a Fuzzy Comparison Function. In *In Proceedings of the Eurospeech*.

D. Mladenić. 2001. Modeling natural language information in textual data using machine and text learning methods. *Zbornik konference ERK*.

M. Popovič. 1992. The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science*.

T. Rotovnik, M. S. Maučec, B. Horvat, Z. Kačič. 2002. Large Vocabulary Speech Recognition of Slovenian Language Using Data-Driven Morphological Models. In *In Proceedings of the Text, Speech and Dialogue Workshop*.

E. W. D. Whittaker. 2000. *Statistical Language Modelling for Automatic Speech Recognition of Russian and English*. Ph.D. thesis, Cambridge University.