

AVTOMATIČNO PREPOZNAVANJE LASTNIH IMEN

Mihael Arčan*, Špela Vintar‡

*mihael_arcan@yahoo.de

‡spela.vintar@guest.arnes.si

Filozofska fakulteta, Univerza v Ljubljani
Arškerčeva 2, 1000 Ljubljana

Named Entity Recognition

The paper deals with Named Entities in German and Slovene texts. We first describe Named Entities from a linguistic viewpoint, where a typology of Named Entities and a theoretical framework is given. The second part of the paper describes different methods of Named Entity Recognition (NER). Since names, like all nouns, are capitalized in German, the task of automatic recognition is not trivial. In Slovene, capitalization is a useful indicator for names. Two main methods for NER are employed: The grammatical rule method utilizes both the internal structure of names as well as their context. The second method is statistical, where frequencies of Named Entities are compared to other parts-of-speech. In the described experiment both methods, but mainly the first, were put to use. While simple or single word names, especially geographical names, are conventionally recognized via lists, multi-word names are more difficult to recognize. The system was evaluated with F-Measure with the result 0,64 for German and 0,74 for Slovene.

Povzetek

Prispevek se ukvarja z lastnimi imeni v nemščini in slovenščini. Na začetku bodo lastna imena obravnavana z lingvističnega vidika, kjer v uvodnih razdelkih podamo razčlenitev lastnih imen in podroben opis. Drugi del prispevka se posveti različnim metodam prepoznavanja lastnih imen. Kot podskupino samostalnikov jih je v nemškem jeziku vizualno težko prepoznati, ker se tudi samostalniki pišejo z veliko začetnico. V slovenskem jeziku pa se pišejo z veliko začetnico le lastna imena, kar olajša njihovo prepoznavanje. Za prepoznavanje lastnih imen se uporabljata dve metodi. Metoda s slovničnimi pravili se poslužuje tako notranje strukture kot tudi okolice lastnih imen. Pri drugi, statistični metodi pride do uporabe statističnih računov, in sicer pogostost nastopa lastnih imen v relaciji z ostalimi besednimi vrstami. V pričujoči raziskavi sta bili uporabljeni obe metodi, poudarek pa je bil na metodi pravil. Izkazalo se je, da največ težav povzročajo večbesedna lastna imena. Prepoznavanje preprostih ali enobesednih lastnih imen, posebno zemljepisnih, se pri vseh sistemih rešuje s pomočjo seznamov. Na koncu je bil izračunan uspeh raziskave z vrednostjo F, ki je za nemško besedilo znašala 0,64 in za slovensko besedilo 0,74 točke.

1. Uvod

Dandanes si sveta ne moremo več predstavljati brez računalnikov. Veliko ljudi se niti ne zaveda, da so jezikovne tehnologije del vsakdanjega življenja. Pri pisanju besedila nam pri prepoznavanju napak pomaga črkovalnik, ki napake tudi avtomatično popravi. Uporaba elektronskih slovarjev nam lajša iskanje neznanih besed. Strojni prevajalniki, kot sta BabelFish¹ in Presis², prevedejo želeno besedilo v različne jezike. Vsem tem pripomočkom je skupno to, da uporabljajo izčrpane baze podatkov.

Iskanje lastnih imen v nemškem jeziku predstavlja večji izziv, ker se v nemščini vsi samostalniki pišejo z veliko začetnico. Zato smo se v tej raziskavi osredotočili na prepoznavanje lastnih imen v nemškem jeziku. Čeprav se v slovenskem jeziku pišejo z veliko začetnico le lastna imena, ta lastnost za prepoznavnost lastnih imen velikokrat ne zadošča.

2. Klasifikacija lastnih imen

V slovenskem in nemškem jeziku se lastna imena delijo na dve glavni skupini:

- osebna lastna imena,
- zemljepisna lastna imena.

Med lastna imena uvrščamo tudi imena stvaritev, organizacij, delovnih skupnosti ipd. (Slovenski pravopis 1994, 22-24.)

Posebna skupina lastnih imen so imena proizvodov. Ker so jim imena podeljena, imajo lastnosti lastnih imen, ker pa hkrati razvrščajo proizvode v različne kategorije, jih lahko uvrščamo tudi med občna imena (Shippan 1992, 65).

Ker so se prvi programi za prepoznavanje lastnih imen razvijali za prepoznavanje imen v angleškem jeziku, je pomembno tudi dejstvo, da v angleškem jeziku v skupino lastnih imen sodijo tudi imena dni (Monday), mesecev (May) in praznikov (Christmas). Zanimivo je, da v angleščini med lastna imena sodijo tudi poimenovanja valut, datumi in odstotki (Roth 2002, 5; Bekavac 2005, 19).

¹ Babelfish Altavista, <http://babelfish.altavista.com>.

² Presis, <http://presis.amebis.si>.

3. Prepoznavanje lastnih imen

Na mednarodni konferenci o razumevanju sporočil MUC-7³ leta 1998 so se med sabo pomerili programi, ki delujejo na osnovi slovničnih pravil (PNF, BSEE, LaSie, LT TTT), in programi, ki za prepoznavanje lastnih imen uporabljajo statistične metode (IdentiFinder, MENE) (Roth 2002, 25-82).

Za prepoznavanje lastnih imen lahko torej uporabljamo dve osnovni metodi.

Prva temelji na slovničnih pravilih, vanjo pa sodita interna in eksterna evidenca. Pri interni evidenci gre za tipične lastnosti in zgradbo imen (*d. o. o., d. d., Inštitut X, X vrh, X jezero*). K njej pripisujemo tudi začetnice imen (*George W. Bush*). Eksterna evidenca pa uporablja sobesedila oz. okolice lastnih imen (*Prof. X Y, Gospa X Y ali X Y ml.*). Pri eksterni evidenci se lahko uporabljajo tudi sezname indikatorjev, ki napovedujejo lastna imena.

Druga metoda je statistična in uporablja izključno pojavitve besed ali besednih vrst v besedilu.

Na prvi pogled se zdi, da bi bil idealen sistem sinteza obeh prijemov, vendar se na konferenci MUC-7 ni izkazalo tako. Pokazalo se je, da so za hitro vzpostavitev sistema primernejše statistične metode. Velika prednost tovrstnih sistemov je tudi ta, da so jezikovno neodvisni. V primerjavi z metodo pravil, pri kateri moramo napisati vsa slovnična pravila in seznam indikatorjev za vsak jezik posebej, se zdi statističen pristop ustrežnejši. Njegova slabost pa je velika odvisnost od označenega korpusa, zaradi česar se prednosti in slabost obeh metod izenačita.

4. Uporabljene metode za prepoznavanje lastnih imen

Za raziskavo smo izdelali manjši slovenski in nemški korpus turističnih besedil s številnimi zemljepisnimi imeni. Izbrali smo turistični vodnik *Pozdrav iz Slovenije* in njegov prevod v nemški jezik *Griße aus Slowenien*. S pomočjo korpusa smo iskali lastna imena v slovenskem in nemškem jeziku, stavčna poravnava besedil s programom Déjà Vu⁴ pa je omogočila še iskanje imen po vzporednem korpusu.

<p><stavek><sl>"Čez Julijske Alpe teče razvodje med Jadranskim in Črnim morjem."</sl><ge>"Über die Julischen Alpen verläuft auch die Wasserscheide zwischen der Adria und dem Schwarzen Meer."</ge></stavek></p>
--

Tabela 1: Primer vzporednega korpusa.

Korpusu smo dodali morfološko-sintaktične oznake. Za označevanje slovenskega dela je poskrbelo podjetje Amebis, nemški del pa so označili na univerzi v Stuttgartu. Za izdelavo sistema za avtomatično prepoznavanje lastnih imen smo uporabili programski jezik Perl.

³ MUC - Message Understanding Conference, http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_pceedings/overview.html.

⁴ Déjà Vu: <http://www.atril.com>.

4.1. Velika začetnica

Glavni poudarek v raziskavi je bil na veliki začetnici. V slovenskem jeziku nam je ta v veliko pomoč pri iskanju lastnih imen, v nemškem jeziku pa nastanejo težave, ker se tudi ostali samostalniki pišejo z veliko začetnico. Tabela 2 prikazuje razmerje med obema jezikoma glede vseh besed v izdelanem korpusu in besed, zapisanih z veliko začetnico.

	slov.	nem.
pojavnica (ang. token)	14.975	13.874
velika začetnica	2.324	4.832
velika zač. v [%]	15,92	34,83

Tabela 2: Razmerje med pojavniciami in besedami z veliko začetnico v slovenščini in nemščini.

Iz tabele 2 je razvidno, kako samostalniki v nemškem jeziku povečujejo delež besed, zapisanih z veliko začetnico. Drugače je v slovenskem jeziku, kjer se z veliko pišejo le besede na začetku stavka in lastna imena.

V tabeli 3 je seznam vseh besed z veliko začetnico, najdenih v dveh vzporednih stavkih iz korpusa.

n	sl.	nem.	n	nem.
1	Povojni*	Die*	15	Einwohner*
2	Nagla*	Entwicklung*	16	<i>Italien</i>
3	<i>Kopra</i>	Kriegsende*	17	Gebiet*
4	<i>Izole</i>	Positivem*	18	Verlust*
5	<i>Piranu</i>	Folgen*	19	Identität*
6	Ker*	Das*	20	Mit*
7	<i>Istre</i>	Wachstum*	21	Willen*
8	Z*	Bebauung*	22	Staates*
9		Stadtbild*	23	Atmosphäre*
10		<i>Koper</i>	24	Toleranz*
11		<i>Izola</i>	25	Gemeinschaft*
12		<i>Piran</i>	26	Verbindungen*
13		Weil*	27	Staatsgrenze*
14		Nachkriegsjahren*		* - ni lastno ime

Tabela 3: Seznam vseh besed z veliko začetnico v dveh vzporednih stavkih iz korpusa.

4.2. Abeceda

Ena od možnosti za prepoznavanje imen je bilo iskanje po nestandardnih črkah (ang. diacritics) v posameznem jeziku. Ker je bilo slovensko besedilo prevedeno v nemški jezik in ker se lastna imena praviloma ne prevajajo (Grah 2000, 1), se v nemškem

besedilu pojavljajo tudi slovenski šumniki. Tako smo v nemškem besedilu hitro našli slovenska lastna imena.

Soča-Tal, Soča, Krško, Kočevje, Snežnik, Goričko, Vače, Hrušica, Mežica-Tal, Lož, Solčava, Boč, Lašče, Brežice, Ormož, Rateče, Stična, Divača, Veržej, Ajdovščina, Sečovelje, Konjiška, Otočec, Sežana, Donačka, Loški, Tržič, Višnja, Košenjak, Paški, Rogaška, Olševa, Mežica, Stržen, Križna, Dobriča, Komarča

Tabela 4: V nemškem besedilu najdena slovenska imena.

4.3. Rodilnik in lastna imena

V nemškem jeziku se samostalniku v rodilniku doda končnica -s (*des Tischeſ*). Lastnim imenom pa se doda končnica -s tudi, kadar ta stojijo brez člena (*die Landschaft Sloweniens*). Če pa pred imenom stoji določni člen, se sklanja samo člen in lastno ime ne dobi končnice.

*Südgrenze des Pohorje
Besiedlung des Prekmurje
Bergland des Snežnik
Bergrücken des Boč
Erlebnis des Kucelj*

Tabela 5: Rodilnik in lastna imena.

4.4. Nemški predlog von

Čprav je rodilnik v nemškem jeziku precej pogost, samostalnik v rodilniku nima vedno končnice -s. Če se samostalnik konča s sičnikom ali če se želimo izogniti hiatu, uporabimo nemški predlog *von* (Grah 2000: 9, 19-20).

Tudi s tem modulom smo našli kar nekaj lastnih imen.

*Bischöfe von Freising belehnt
Gebiet von Idrijsko und
Zeiten von Tavčar und
Ebenen von Radovljica und
Wasserscheide von Rateče ist
Talkessels von Ljubljana
Gegend von Bled übergehend
Grenzen von Slowenien
Eibenbaum von Solčava
Tal von Vitanje*

Tabela 6: Predlog *von*.

4.5. Seznam zemljepisnih imen

Kot je za sisteme za prepoznavanje lastnih imen običajno, smo tudi mi uporabili seznam zemljepisnih imen (ang. gazetteer). Ta so bila izluščena iz digitalne enciklopedije Wikipedia (različica na DVD-ju, 2005). Seznam je vseboval 59 imen v slovenskem jeziku in

njihove prevode v nemškem jeziku, k temu pa smo dodali še 193 lastnih imen v slovenščini.

4.6. Seznam osebnih imen

Osebna imena smo enako kot zemljepisna izluščili iz enciklopedije Wikipedia. Za razliko od zemljepisnih imen smo vključili vsa osebna imena, zgodovinska in sodobna, kar predstavlja približno 35.000 različnih osebnih imen.

S pomočjo izdelanega seznama je sistem pravilno prepoznal imena kot npr.: *H. Freyer*, (Kaiser) *Augustus*, (Kaiser) *Karl VI.*, (Erzherzog) *Rudolf IV.* ali (König) *Heinrich II.*

Kljub navedenim rezultatom se je ta možnost izkazala za zelo slabo. Ker so bila imena izluščena iz nemške enciklopedije in ker je bilo v seznam zajetih veliko imen, je prišlo do ogromnega prekrivanja med splošnimi samostalniki in lastnimi imeni. Ta prekrivanja so bila npr.: (George) Moor, (Simon) Koper, (Dieter) Jahr, (Peet) Stol, (Viktor) Klima oder (Friedrich) Wetter. Vsi ti primeri so v korpusu nastopali kot samostalniki, zaradi velike baze Wikipedie pa jih je ta modul (napačno) označil kot lastna imena.

4.7. Iztočnice iz sobesedila

Druga možnost prepoznavanja lastnih imen je že omenjena eksterna evidenca. S to metodo analiziramo okolice lastnih imen in najdene vzorce uporabimo za napovedovanje imen. Čeprav lastnih imen načeloma ne prevajamo, jih prevajalci pogosto prevedejo ali razložijo v oklepajih ali v narekovajih.

Capris (Koper), Vrhnika (Nauportus), Kočevje (Gotschee), Otok (Gutenwerth), Unterkrain (Dolenjsko), Igla (Nadel), Savario (Szombathely), Nauportus (Vrhnika), Poetovio (Ptuj), Siscia (Sisak), Rog (Hornwald)

Tabela 7: Prepoznavanje imen z opisom.

Druga možnost prepoznavanja imen s pomočjo okolice je opisovanje gora z nadmorsko višino. Tako je v besedilu nekaj primerov, kjer za imenom nastopa oklepaj z navedbo nadmorske višine.

<i>Križ (2429 m)</i>	<i>Brezje (538 m)</i>
<i>Slavnik (1028 m)</i>	<i>Macelj (718 m)</i>
<i>Grmada (887 m)</i>	<i>Peca-Berg (2126 m)</i>
<i>Mrzlica (1122 m)</i>	<i>Stenica (1091 m)</i>
<i>Gozdnik (1090 m)</i>	<i>Boč (979 m)</i>
<i>Rogla (1517 m)</i>	<i>Ratitovec (1678 m)</i>

Tabela 8: Prepoznavanje z višinsko oznako.

4.8. Zaporedje črk

Drug način prepoznavanja lastnih imen je različno zaporedje črk v besedi v posameznem jeziku.

Za ugotavljanje zaporedja črk v nemškem jeziku je bil uporabljen seznam osnovnega besedišča nemškega jezika,

za slovensko zaporedje črk pa smo uporabili izvorno besedilo.

sch che hen ich gen ten ste ter ein ver cht cke ken ung ach sen rei lic ren ben nde lle len sse ern den tte aus fen nge der rop gpl ägl rnr nap önl röh rok önn rog ukt uku rrü nbe nfä	pre sta ske ega ski sko ove nsk ost nje dol rav lin let jsk red kra lov anj sto eni oli pri ven ran slo lam bes tah bev gič lar a-f iol ion lii naz lif zcv mda lau žer hol lib nbe lay lfs
--	--

Tabela 9: Zaporedje črk v posameznem jeziku.

Modul je primerjal nemško besedilo s tipičnim slovenskim zaporedjem črk in če je imela beseda v prevodu netipično zaporedje črk, jo je označil kot slovensko lastno ime.

<i>Gajus, Pokljuka, Sava, Krn, Soča, Boka, Muzci, Celje, Kozjak, Ptuj, Mežica, Košuta, Pavla, Osp, Sečovelje, Piave, Zasavje, Cezlak ...</i>
--

Tabela 10: Rezultati iz modula zaporedja črk.

4.9. Končnica -ska/-ske/-sko

Ta modul je uporaben za prepoznavanje večbesednih lastnih imen. Pri analizi besedila in rezultatov je bilo razvidno, da se pri številnih slovenskih večbesednih lastnih imenih prvi del konča s končnico -ska/-ske/-sko. To je bil tudi edini modul, ki je iskal večbesedna lastna imena, ostali moduli so iskali le po enobesednih imenih.

<i>Šaleška dolina, Osapska Reka, Radgonsko-Kapelske gorice, Kamniška Bistrica, Mučka Bistrica, Dolenjske Toplice, Matarsko podolje, Kočevska Mala, Potočka zijalka, Sorško polje, Baška grapa, Selška dolina, Polhograjska Gora, Slatenska plošča</i>

Tabela 11: Imena s končnico -ska/-ske/-sko.

4.10. Statistika

Kot smo že omenili, se za prepoznavanje imen uporabljajo tudi statistične metode. Čeprav smo se v prispevku bolj posvetili imenom in njihovi okolici, statističnih metod nismo povsem zanemarili.

Ta modul se je posvetil zaporedju različnih besednih vrst

Tabela 12: Kombinacija besednih vrst.

ART NN ADJA NN NN PUNCT NN APPR ART ADJA APPR ART NN ART PUNCT ART APPR NE NN VFIN PUNCT APPR

pred lastnimi imeni, pri čemer smo uporabili oblikoskladenjsko označen korpus. V tem modulu ni bila pomembna struktura imena, temveč besedne vrste in njihova pojavnost.

Tabela 12 prikazuje najpogostejše kombinacije dveh sosednjih besednih vrst. V tej tabeli je razvidno, da je v nemškem jeziku najpogostejša kombinacija člena in samostalnika. Lastna imena najpogosteje nastopajo s predlogi, ki stojijo pred njimi. Takšni primeri so bili: in *Železniki*, von *Škofje? Loka*, in *Slowenien*, is *Bohinj*, bis *Bovec* ...

Z nadaljnjo analizo smo se osredotočili na zaporedje besednih vrst do sedem mest pred imenom. Te kombinacije prikazuje spodnja tabela.

APPR ART NN ART ADJA NN +NE (APPR)an (ART)den (NN)Stellen (ART)der (ADJA)heutigen (NN)Städte (NE)Čromelj
NN APPR ART NE APPR NE +NE (NN)Brücke (APPR)über (ART)die (NN)Save (APPR)bei (NE)Zidani (NE)Most
ADJA NN APPR ART ADJA NN +NE (ADJA)entscheidende (NN)Schlacht (APPR)zwischen (ART)den (ADJA)konkurrierenden (NN)Kaisern (NE)Theodosius
PUNCT APPR ART NN APPR NE +NE (PUNCT). (APPR)Unter (ART)dem (NN)Blegoš (APPR)in (NE)Poljanska (NE)dolina
NN APPR ART NN APPR +NE (NN)Gegend (APPR)an (ART)der (NN)Save (APPR)zwischen (NE)Litija

Tabela 13: Kombinacija besednih vrst pred lastnimi imeni.

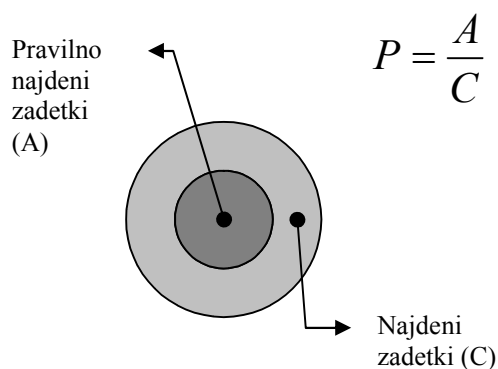
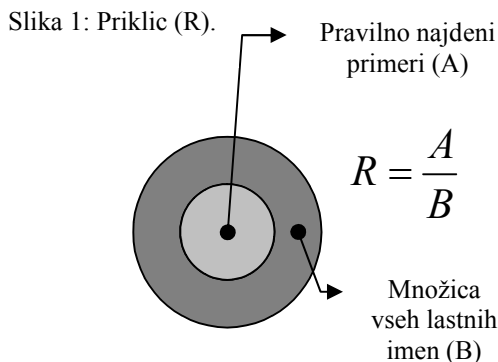
Razlika med statističnimi metodami in metodami z uporabo pravil je ta, da statistični sistem poda vrednost v odstotkih, ki pa ni zavezujoča. Pri opisu gora z nadmorsko višino je npr. zelo velika verjetnost, da pred opisom stoji lastno ime. Statistična tabela pa nam poda samo verjetnost (npr. 66 %), ne izključuje pa možnosti, da najdena beseda ni lastno ime.

Za popoln statistični sistem pa seveda ni dovolj, da gledamo samo, kaj stoji pred imeni, temveč moramo opazovati celotno okolico imen.

5. Evalvacija

Učinkovitost sistema smo izračunali z vrednostjo F (ang. F-Measure), ki primerja dve vrednosti, in sicer priklic (ang. recall, R) in natančnost (ang. precision, P).

Priklic (R) predstavlja razmerje med pravilno najdenimi imeni in množico vseh lastnih imen v korpusu. Natančnost (P) pa predstavlja razmerje med pravilno najdenimi zadetki in najdenimi zadetki v korpusu (Kemayou Yamga 2006).



Slika 2: Natančnost (P).

Za posamezne module je bilo mogoče izračunati samo natančnost (P), ker se posamezen modul nanaša na specifično lastnost lastnih imen in ne more zajeti vseh ostalih imen, ki te lastnosti nimajo.

	A	C	P
abeceda	153	210	0,73
rodilnik	14	19	0,74
predlog »von«	90	109	0,83
seznam zemljep. imen	112	112	1,00
seznam osebnih imen	12	138	0,09
iztočnice in sobesedila	22	24	0,91
zaporedje črk	45	59	0,76
končnica -ska/-ske/-sko	38	42	0,90

Tabela 14: Vrednosti P za posamezne module.

Na koncu raziskave smo vse module povezali in izračunali celotno vrednost F po enačbi:

$$F = \frac{2 \times (R \times P)}{R + P}$$

Za izračun vrednosti F smo morali celotno besedilo z lastnimi imeni ročno označiti.

pravilno najdeni primeri	402
najdeni zadetki	592
natančnost (P)	0,68

Tabela 15: Natančnost (P) za celotni korpus.

pravilno najdeni primeri	402
množica lastnih imen	660
priklic (R)	0,61

Tabela 16: Priklic (R) za celotno besedilo.

Vrednost F za nemško besedilo:

$$F_{de} = \frac{2 \times (0,61 \times 0,68)}{0,61 + 0,68} = 0,643$$

Za primerjavo smo izračunali še vrednost F za slovensko besedilo. Tudi za to besedilo smo uporabili iste module, z manjšimi modifikacijami.

pravilno najdeni primeri	849
najdeni zadetki	1283
natančnost (P)	0,66

Tabela 17: Natančnost (P) za celotno besedilo.

pravilno najdeni primeri	849
množica lastnih imen	1006
priklic (R)	0,84

Tabela 18: Priklic (R) za celotno besedilo.

Vrednost F za slovensko besedilo:

$$F_{sl} = \frac{2 \times (0,84 \times 0,66)}{0,84 + 0,66} = 0,739$$

Slovenski pravopis 1: Pravila (1994): 4., pregledana izdaja (s stvarnim kazalom), Ljubljana, DZS

6. Zaključek

Prispevek prikazuje, da obstajajo različne možnosti za luščenje lastnih imen iz tekočega besedila. Velika začetnica sicer zelo pomaga pri prepoznavanju lastnih imen, ampak samo, če so imena enobesedna. Pri večbesednih imenih označuje velika začetnica v slovenskem jeziku samo začetek lastnega imena. Te pomoči pa v nemškem jeziku ni, ker vsaka velika začetnica ne označuje nujno lastnega imena.

Učinkovitost smo izračunali s pomočjo vrednosti F, ki je za nemško besedilo znašala 0,64, za slovensko pa 0,74 točke. Pri tem je razvidno, kako lahko velika začetnica v slovenskem jeziku poveča učinkovitost sistema.

Nadaljnji korak bi bila sintaktična analiza (ang. parsing) za prepoznavanje večbesednih lastnih imen in označevanje teh kot eno samo enoto.

V veliko pomoč bi bil tudi večji učni korpus, na katerem bi se sistem lahko učil. Z večjim korpusom bi se tudi povečalo število različnih lastnih imen.

Ena večjih pomanjkljivosti sistema je pomanjkanje slovnčnih pravil, ki jih je v izdelanem sistemu zaenkrat deset, medtem ko razvitejši sistemi uporabljajo več tisoč takšnih pravil.

Ugotavljamo, da je luščenje lastnih imen z uporabljenimi postopki učinkovito, vendar sistem kot vsak drug potrebuje čas za rast in učenje.

7. Literatura

- Bekavac, Božo (2005): Strojno prepoznavanje naziva u suvremenim hrvatskim tekstovima, Sveučilište u Zagrebu Filozofski fakultet, Zagreb
- Grah, Käthe (2002): Slovenska stvarna imena v nemških besedilih – Slowenische Sachnamen in deutschen Texten, Ljubljana, Znanstveni inštitut Filozofske fakultete
- Kemayou Yamga, Syriane (2006): Informationsextraktion http://kontext.fraunhofer.de/haenelt/kurs/Referate/KemayouYamga_W05/Informationsextraktion.pdf (15. 5. 2006)
- Peterlin, Janez (2003a): Pozdravljena, Slovenija, prenovljena izdaja, Mladinska knjiga, Ljubljana
- Peterlin, Janez (2003b): Grösse aus Slowenien, überarbeitete Ausgabe, Mladinska knjiga, Ljubljana
- Roth, Jaennette (2002): Der Stand der Kunst in der Eigennamen-Erkennung, Mit einem Fokus auf Produktenamen-Erkennung, Lizentiatsarbeit der Philosophischen Fakultät der Universität Zürich, Zürich
<http://www.ifi.unizh.ch/cl/study/lizarbeiten/lizjeannetroth.pdf> (15. 5. 2006)
- Schippian, Thea (1992): Lexikologie der deutsche Gegenwartsfrage, Tübingen, Niemeyer Verlag