Zbornik 13. mednarodne multikonference **INFORMACIJSKA DRUŽBA - IS 2010**

Proceedings of the 13th International Multiconference **INFORMATION SOCIETY - IS 2010**

Zvezek C / Volume C

Jezikovne tehnologije

Language Technologies

Uredila / edited by:
Tomaž Erjavec
Jerneja Žganec Gros

14. do 15. oktober 2010, Ljubljana

14th - 15th October 2010, Ljubljana, Slovenia

Zbornik 13. mednarodne multikonference
INFORMACIJSKA DRUŽBA - IS 2010
Zvezek C

Proceedings of the 13th International Multiconference
INFORMATION SOCIETY - IS 2010
Volume C

Zbornik
Sedme konference JEZIKOVNE TEHNOLOGIJE

Proceedings of the
Seventh Language Technologies Conference

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

14. do 15. oktober 2010 / October 14th - 15th, 2010
Ljubljana, Slovenia

Uredniki:

Tomaž Erjavec
Odsek za tehnologije znanja
Institut »Jožef Stefan«, Ljubljana

Jerneja Žganec Gros
Alpineon d.o.o, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Tisk: Birografika BORI d.o.o.
Priprava zbornika: Mitja Lasič, Jana Krivec
Oblikovanje naslovnice: BONS d.o.o.
Tiskano iz predloga avtorjev
Naklada: 35

Ljubljana, oktober 2010

Konferenco IS 2010 sofinancirata
Ministrstvo za visoko šolstvo, znanost in tehnologijo
Javna agencija za raziskovalno dejavnost RS (ARRS)
Institut »Jožef Stefan«

Informacijska družba
ISSN 1581-9973

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

004.934(082)
81'25:004.6(082)

KONFERENCA Jezikovne tehnologije (7 ; 2010 ; Ljubljana)
   Zbornik Sedme konference Jezikovne tehnologije, 14. do 15. oktober 2010 : zbornik 13. mednarodne multikonference Informacijska družba - IS 2010, zvezek C = Proceedings of the Seventh Language Technologies Conference, October 14th-15th, 2010, Ljubljana, Slovenia : proceedings of the 13th International Multiconference Information Society - IS 2010, volume C / uredila, edited by Tomaž Erjavec, Jerneja Žganec Gros. - Ljubljana : Institut Jožef Stefan, 2010. - (Informacijska družba, ISSN 1581-9973)

ISBN 978-961-264-026-2
1. Jezikovne tehnologije 2. Language Technologies 3. Informacijska družba 4. Information society 5. Erjavec, Tomaž, 1960- 6. Mednarodna multikonferenca Informacijska družba (13 ; 2010 ; Ljubljana)
252779520

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2010

V svojem trinajstem letu je multikonferenca Informacijska družba (http://is.ijs.si) znova dokazala, da je ena vodilnih srednjeevropskih konferenc, ki združuje znanstvenike z različnih raziskovalnih področij, povezanih z informacijsko družbo. V letu 2010 smo v multikonferenco povezali deset odličnih neodvisnih konferenc. V Sloveniji in po svetu mgroli konferenc. Naša multikonferenca izstopa po širini in obsegu tem, ki jih obravnava, predvsem pa po akademski odprtosti in širini, ki spodbuja nove ideje.

Multikonferenca temelji na sinergiji interdisciplinarnih pristopov, ki obravnavajo različne vidike informacijske družbe ter poglabljajo razumevanje informacijskih, komunikacijskih in družbenih storitev v najširšem pomenu besede. Na multikonferenci predstavljamo, analiziramo in preverjamo nova odkritja in pripravljamo teren za njihovo praktično uporabo, saj je njen osnovni namen promocija raziskovalnih dosežkov in spodbujanje njihovega prenosa v prakso na različnih področjih informacijske družbe tako v Sloveniji kot tujini.

Na multikonferenci bo na vzporednih konferencah predstavljenih 300 referatov, vključevala pa bo tudi okrogle mize in razprave. Referati so objavljeni v zbornikih multikonference, izbrani prispevki pa bodo izšli tudi v posebnih številkah dveh znanstvenih revij, od katerih je ena Informatica, ki se ponaša s 34-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2010 sestavljajo naslednje samostojne konference:

- Odprta delavnica mednarodnega projekta Confidence
- Inteligentni sistemi
- Jezikovne tehnologije
- Kognitivne znanosti
- Robotika
- Rudarjenje podatkov in podatkovna skladišča (SiKDD 2010)
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Soočanje z demografskimi izzivi
- Vzgoja in izobraževanje v informacijski družbi
- 3. Minikonferenca iz teoretičnega računalništva 2010.

Zanimivo je, da finančna recesija ni zmanjšala zanimanja za informacijsko družbo, saj je prispevkov primerljivo z lansko konferenco, kljub temu, da se je državno sofinanciranje močno zmanjšalo. Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija. Zahvaljujemo se tudi Agenciji za raziskovalno dejavnost RS ter Ministrstvu za visoko šolstvo, znanost in tehnologijo za sodelovanje in podporo. V imenu organizatorjev konference pa se želimo posebej zahvaliti udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V letu 2010 sta se programski in organizacijski odbor odločila, da bosta podelila posebno priznanje Slovencu ali Slovenki za izjemen življenjski prispevek k razvoju in promociji informacijske družbe v našem okolju. Z večino glasov je letošnje priznanje pripadlo dr. Tomažu Kalinu. V letu 2010 tudi prvič podeljujemo nagrado za tekoče dosežke. Za aktivno delo pri računalniških tekmovanjih in drugih računalniških dogodkih sta odbora izmed predlogov izbrala Marka Grobelnika. Čestitamo obema nagrajencema!

Franc Solina, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2010

In its 13th year, the Information Society Multiconference (http://is.ijs.si) again demonstrated that is is one of the leading conferences in Central Europe gathering scientific community with a wide range of research interests in information society. In 2010, we organized ten independent excellent conferences forming the Multiconference. There are plenty of conferences in Slovenia and all over the world. The broad range of topics and the open academic environment fostering new ideas makes our event unique among similar conferences.

The Multiconference flourishes the synergy of different interdisciplinary approaches dealing with the challenges of information society. The major driving forces of the Multiconference are search and demand for new knowledge related to information, communication, and computer services. We present, analyze, and verify new discoveries in order to prepare the ground for their enrichment and development in practice. The main objective of the Multiconference is presentation and promotion of research results, to encourage their practical application in new ICT products and information services in Slovenia and also broader region.

The Multiconference is running in parallel sessions with 300 presentations of scientific papers. The papers are published in the conference proceedings, and in special issues of two journals. One of them is Informatica with its 34 years of tradition in excellent research publications.

The Information Society 2010 Multiconference consists of the following conferences:

- Confidence Project Open Workshop
- Intelligent Systems
- Languate technologies
- Cognitive Sciences
- Robotics
- Data Mining and Data Warehouses (SiKDD 2010)
- Collaboration, Software and Services in Information Society
- Demographic Challenges in Europe
- Education in Information Society
- The Third Mini Conference on Theoretical Computing 2010.

Interestinglly, the economic recession is not affecting Information society, judging from the number of single conferences; however, the national funding significantly decreased as a result of crisis. The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM. We would like to express our appreciation to the Slovenian Government for cooperation and support, in particular through the Ministry of Higher Education, Science and Technology and the Slovenian Research Agency..

In 2010, the Programme and Organizing Committees decided to award one Slovenian for his/her life-long outstanding contribution to development and promotion of information society in our country. With the majority of votes, this honor went to Dr. Tomaž Kalin. Congratulations!
In addition, a reward for current achievements was pronounced for the first. It goes to Marko Grobelnik for his support of the ACM computer competitions.

On behalf of the conference organizers we would like to thank all participants for their valuable contribution and their interest in this event, and particularly the reviewers for their thorough reviews.

Franc Solina, Programme Committee Chair
Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Izrael
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Finland
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Miklós Krész, Hungary
József Békési, Hungary

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek, co-chair
Lana Jelenkovič
Jana Krivec
Mitja Lasič

Programme Committee

Franc Solina, chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Marjan Pivka
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Tomaž Šef
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
David B. Vodušek
Baldomir Zajc
Blaž Zupan
Boris Žemva
Janez Žibert
Leon Žlajpah

KAZALO / TABLE OF CONTENTS

Zbornik 13. mednarodne multikonference
INFORMACIJSKA DRUŽBA - IS 2010
Zvezek C

Proceedings of the 13th International Multiconference
INFORMATION SOCIETY - IS 2010
Volume C

Zbornik
Sedme konference JEZIKOVNE TEHNOLOGIJE

Proceedings of the
Seventh Language Technologies Conference

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

14. do 15. oktober 2010 / October 14th - 15th, 2010
Ljubljana, Slovenia

PREDGOVOR K ZBORNIKU SEDME KONFERENCE »JEZIKOVNE TEHNOLOGIJE«

V pričujočem zborniku so objavljeni prispevki s sedme konference "Jezikovne tehnologije", ki je potekala 14. in 15. oktobra 2010 v Ljubljani, v okviru multikonference "Informacijska družba" IS'2010. Letošnja konferenca je bila namenjena članom Slovenskega društva za jezikovne tehnologije (SDJT) in drugim, ki jih to področje zanima, kot forum, kjer lahko predstavijo svoje delo v preteklih dveh letih, kolikor je minilo od zadnje konference o jezikovnih tehnologijah, organizirane v okviru IS. Zbornik vsebuje 18 prispevkov, ki obravnavajo široko paleto raziskav; posebej izstopa veliko število prispevkov s področja govornih tehnologij. Organizatorji bi se radi zahvalili vsem, ki so prispevali k uspehu konference: vabljenim predavateljem, avtorjem prispevkov, programskemu odboru za recenzentsko delo ter organizatorjem IS'2010.

Preface to the Proceedings of the Seventh Language Technologies Conference

These proceedings contain the contributions for the Seventh Language Technologies Conference, which took place on October 14th and 15th 2010 in Ljubljana, in the scope of the Information Society Multiconference, IS'2010. The conference was aimed at the members of the Slovenian Language Technology Society and others interested in the field, as a forum where they could present their work in the last two years, which have passed since the previous conference on Language Technologies organised in the scope of IS. The proceedings contain 18 contributions, which present a wide variety of research topics; especially numerous are contributions dealing with speech technologies. The organisers would like to thank the many people who contributed to the success of the conference: the invited speakers and the authors of contributions, the programme committee of the conference and the organising committee of IS 2010.

Tomaž Erjavec, Jerneja Žganec Gros
Ljubljana, October 2010.

RECENZENTI

- doc. dr. Simon Dobrišek, Fakulteta za elektrotehniko, Univerza v Ljubljani
- doc. dr. Tomaž Erjavec (predsednik), Odsek za tehnologije znanja, Institut "Jožef Stefan"
- dr. Darja Fišer, Filozofska fakulteta, Univerza v Ljubljani
- doc. dr. Vojko Gorjanc, Filozofska fakulteta, Univerza v Ljubljani
- prof. dr. Ivo Ipsić, Faculty of Engineering, Univerza v Reki (Hrvaška)
- prof. dr. Zdravko Kačič, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
- Simon Krek, Amebis, d.o.o.
- doc. dr. Cvetana Krstev, Filozofska fakulteta, Univerza v Beogradu (Srbija in Črna gora)
- dr. Domen Marinčič, Odsek za inteligentne sisteme, Institut "Jožef Stefan"
- prof. dr. France Mihelič, Fakulteta za elektrotehniko, Univerza v Ljubljani
- doc. dr. Dunja Mladenić, Odsek za tehnologije znanja, Institut "Jožef Stefan"
- prof. dr. Marko Stabej, Filozofska fakulteta, Univerza v Ljubljani
- dr. Tomaž Šef, Odsek za inteligentne sisteme, Institut "Jožef Stefan"
- prof. dr. Rastislav Šuštaršič, Filozofska fakulteta, Univerza v Ljubljani
- prof. dr. Marko Tadić, Oddelek za jezikoslovje, Univerza v Zagrebu (Hrvaška)
- doc. dr. Darinka Verdonik, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
- doc. dr. Špela Vintar, Filozofska fakulteta, Univerza v Ljubljani
- dr. Jerneja Žganec Gros (predsednica), Alpineon, d.o.o.
- doc. dr. Janez Žibert, Fakulteta za matematiko, naravoslovje in informacijske tehnologije, Univerza na Primorskem

Analysing similarities and differences between corpora

Serge Sharoff

Centre for Translation Studies
School of Modern Languages
University of Leeds
LS2 9JT, Leeds, UK
s.sharoff@leeds.ac.uk

Abstract

The notion of comparable corpora rests on our ability to assess the difference between corpora which are claimed to be comparable, but this activity is still art rather than proper science. Here I will discuss attempts at approximating the content of corpora collected from the Web using various methods, also in comparison to traditional corpora, such as the BNC. The procedure for estimating the corpus composition is based on selecting keywords, followed by hard clustering or by building topic models. This can apply to corpora within the same language, e.g., the BNC against various crawled Internet corpora, as well as to corpora in different languages, e.g., webpages collected using the same procedure for English and Russian. I will also discuss the differences in using hard clustering vs topic models.

1. Introduction

The British National Corpus (BNC) has been collected in the beginning of the 1990s using various written and spoken sources from the 1970s-1980s. It aims at representing modern English, but in many respects it is outdated and it does not cover many domains. The web is huge, it is easy to collect data from it either by using search engine queries (Baroni and Bernardini, 2004; Sharoff, 2006) or crawling respective websites (Joho and Sanderson, 2004). However, once we have a corpus, we still do not know its composition, e.g., the proportion of webpages for medical doctors and patients in the corpus of (Baroni and Bernardini, 2004), or to what extent the BNC is similar to a web corpus.

The problem of not knowing the content gets another dimension when we use comparable corpora, i.e., two corpora which are claimed to be similar in one aspect or the other. It is possible that comparable corpora are collected in their own ways and are drawn from different distributions, e.g., from different website or for different languages. For example, within the TTC project (Blancafort et al., 2010) our aim is to explore the possibility of mining terminological resources from specialised comparable corpora. For this task, we used parallel seeds to collect corpora on comparable topics by retrieving webpages returned in response to queries containing identical or nearly identical terms in several languages (below shown for English and Russian):

fossil fuel	ископаемое топливо
power station	электростанция
hydroelectricity	гидроэнергетика
photovoltaics	фотоэлектричество

However, the crucial question is: do we get comparable pages by sending comparable queries? One approach to comparing corpora across languages is by translating the features obtained from documents, usually such features are keywords (Adafre and de Rijke, 2006). However, translations listed in dictionaries are not always suitable in a given context. Also for many language pairs we lack reliable resources.

In this study I analyse ways of comparing the composition of different corpora using unsupervised methods for keyword selection and corpus classification.

2. Methodology

2.1. Corpora used for analysis

The BNC classifies its documents using a complex classification scheme (Lee, 2001), which includes such categories as

- domains (eight labels in total: natsci, appsci, socsci, belief, imaginative, leisure, business, world affairs);
- genre (seventy labels in total, W.advert, W.newsp.brdsheet.national.arts, etc)
- information about the audience, author, publication medium, etc.

In spite of the complexity of the classification system, it does not cover many substantial differences between the BNC texts, e.g., a text from socsci can be from the subdomain of linguistics or history, the domain of world affairs covers both local British and international politics. The composition of the BNC can be compared to ukWac and itWac (Baroni et al., 2009), large corpora crawled from the .uk and .it domains to represent respectively British English and Italian (with subsequent language filtering). Another task is to compare the composition of specialised corpora collected using comparable seeds (Table 1).

2.2. Keyword selection

The procedure for selecting the keywords per each document was based on the Log-likelihood (LL) score (Rayson et al., 2004). Like the commonly used tf*idf score the procedure is language-independent, but unlike tf*idf it takes into account the relative frequency of a term in a document against the reference corpus as well as the absolute number of its occurrences as the evidence of its statistical significance. The LL-score also allows setting a straightforward threshold using the value of 10.83 ($p = 0.001$) or 15.13 ($p = 0.0001$), for the justification see (Rayson et al., 2004). An example of keywords is shown below:[1]

[1]Extracted from the ukWac page `http://www.comp.leeds.ac.uk/biosystems/neuroscience.shtml`

Table 1: Corpora used in this study (sizes given in tokens and documents)

Corpus	BNC	ukWac	itWac	EN-Energy	RU-Energy	ZH-Energy
Tokens	111246939	2119891296	179512658	7505765	7766462	12431752
Documents	4054	2541926	175646	5762	5126	3287

LL-score	**N**	**Keyword**
126.07	7	Womble
101.47	9	neural
91.26	6	elegans
62.55	13	model
47.22	6	simulation
46.47	10	network
39.59	3	locomotion
36.66	3	biologically
26.71	3	constrain
21.95	3	nervous
21.78	3	cognitive
	...	

They are all useful for describing this individual webpage, but some of them (*Womble, elegans*) are too specific for the purpose of clustering webpages of a general-purpose corpus: *Womble* is a keyword for only 22 pages in ukWac, so it is less useful for finding its clusters. Also a corpus of the size of ukWac generates 89,394 unique keywords for clustering its 2,541,926 documents, causing dimensionality problems. Restricting the keyword lexicon down to the most common words[2] limits the ability to select pages with specific topics, so a simple approach used in this study was to reduce the amount of keywords down to the 10,000 words most often selected as keywords in the entire corpus. The use of the complete set of keywords tested on the BNC was marginally detrimental to the results. The clusters remained the same with the exception of lumping the clusters of investment with environment and medicine with biology and splitting the clusters of education and politics.

2.3. Clustering methods

For unsupervised detection of domains two scenarios were used, hard clustering and generation of topic models. Because of the need to cluster a fairly large number of webpages, the repeated bisection (RB) algorithm was used in Cluto (Zhao and Karypis, 2004), as it is quite efficient and tends to produce clearly interpretable results, cf also a study in (Steinbach et al., 2000). A cluster is treated as a subcorpus and described in terms of its keywords against the entire the corpus.

Generation of topic models is based on Latent Dirichlet Allocation (LDA), which estimates the distribution of probabilities of keywords belonging to different topics (as in traditional clustering, the number of topics is set at the beginning of the experiment, but not the allocation of topics to documents), while the distribution is derived from the distribution of hidden variables (like in Hidden Markov models). This setup helps in generalising inherent similarities between the keywords, see (Blei et al., 2003). For each topic we get the degree of its association with documents and keywords. The advantage of topic models in comparison to RB clustering is that each document gets a degree of its association with each topic, so that it can belong to several topics at once. However, this does not allow us to estimate the partitioning of the entire corpus into a set of clusters in order to compare their relative sizes.

It is known from prior experience that there are more topics in the BNC than the set indicated in its classification scheme, so more than eight clusters need to be used. In all experiments with general-purpose corpora I used 20 clusters. This number helped in producing interpretable models covering a diverse set of topics, while a larger number of clusters makes the comparison task inherently difficult. As discussed in (Chang et al., 2009), there is no correlation between human perception of the consistency of clusters and automatic measures (such as perplexity or I_2), so it is difficult to estimate the right amount of clusters automatically, while in the context of this study there was no scope for a proper human evaluation of the quality of each clustering solution for each corpus. CLUTO (Zhao and Karypis, 2004) uses I_2 as the standard objective function, which maximises the cosine similarity between each node and the centroid of the clusters it belongs to. I_2 on the BNC grew slowly when the number of clusters varied from 10 to 27 (reflecting the reduced size of each cluster), so no definite estimate towards the desired number of clusters can be made:

N	I_2	**N**	I_2	**N**	I_2
10	1.38	16	1.49	22	1.57
11	1.40	17	1.50	23	1.58
12	1.42	18	1.52	24	1.59
13	1.44	19	1.53	25	1.60
14	1.46	20	1.55	26	1.61
15	1.47	21	1.56	27	1.62

3. Results

Even though corpus analysis via clustering is more important for web-derived corpora, because we do not know their exact composition, it makes sense to evaluate our methods on a corpus we know better, such as the BNC (Section 3.1.). Then, in Section 3.2. I will compare ukWac and itWac, two general-purpose Internet corpora, to the BNC. Finally, I will present a study of specialised comparable corpora across different languages (Section 3.3.).

3.1. Clusters in the BNC

Table 2 lists the clusters and topic models with their keywords. The clusters produced by Cluto are numbered according to their internal consistency (the smallest number for the greatest consistency). Also, at the bottom of the list of clusters I indicate the most frequent genre categories from the BNC associated with this genre. In case the genre is not informative (W.misc), the most common domain

[2] As done, for example, in (Kilgarriff, 2001).

Table 2: Clusters (above) and topic models (below) in the BNC

0	3.3%	Inc, Corp, software, user, Unix, system, lifespan, IBM, module, application, version, package, file
1	1.5%	studio, video-tape, speaker, voice, read, over, report, say, yesterday, male, police, Oxford, Swindon
2	1.7%	pollution, environmental, conservation, nuclear, emission, waste, energy, forest, ozonosphere
3	7.6%	Yeah, oh, get, yeah, well, know, yes, go, No, na, think, there, cos, gon, what, no, just, like, say
4	3.7%	player, win, game, Cup, season, club, team, play, match, goal, championship, score, League, ball
5	1.7%	aircraft, engine, railway, station, car, fly, train, pilot, squadron, locomotive, steam, line, crew
6	2.0%	patient, disease, treatment, study, cell, infection, gastric, health, acid, clinical, concentration, care
7	2.8%	God, Jesus, church, Christian, Christ, faith, king, bishop, prayer, gospel, spirit, pope, Holy
8	3.4%	school, award, teacher, student, education, pupil, course, curriculum, research, study, ref, subject
9	6.6%	okay, right, yeah, get, what, so, yes, if, just, well, go, think, know, actually, mean, there, oh
10	9.9%	government, Minister, party, political, election, Soviet, labour, state, country, president, Prime
11	4.1%	court, case, Act, law, defendant, plaintiff, contract, section, person, appeal, any, solicitor, under
12	4.9%	music, film, guitar, band, play, song, album, Eliot, bass, movie, sound, musical, pop, actor
13	3.5%	cell, gene, DNA, protein, energy, equation, sequence, molecule, temperature, surface, particle
14	6.5%	company, market, rate, price, share, cost, profit, tax, business, firm, UK, investment, bank
15	4.6%	Darlington, local, Council, council, housing, pension, councillor, authority, scheme, service, area
16	5.4%	art, painting, century, artist, exhibition, Edward, building, William, museum, town, king, church
17	5.5%	social, language, theory, word, may, text, behaviour, process, individual, information, meaning
18	15.5%	her, him, me, my, say, look, eye, smile, back, go, feel, door, tell, hand, know, think, like, could
19	5.6%	fish, water, your, plant, knit, bird, food, colour, breed, tank, can, horse, leaf, use, dry, egg, specie

0: W.non.ac.tech.engin; 1: W.news.script; 2: W.misc/W.app.science; 3: S.conv; 4: W.pop.lore/W.newsp.brdsht.nat.sports; 5: W.misc/W.leisure (texts for railway/aircraft enthusiasts); 6: W.ac.medicine; 7: W.religion; 8: W.misc/W.soc.science; 9: S.consult; 10: W.non.ac.polit.law.edu,W.newsp.brdsht.nat.reports; 11: W.ac.polit.law.edu; 12: W.newsp.brdsht.nat.arts,W.pop.lore; 13: W.non.ac.nat.science,W.ac.nat.science; 14: W.commerce; 15: S.meeting; 16: W.non.ac.humanities.arts,W.biography; 17: W.ac.soc.science; 18: W.fict.prose; 19: W.pop.lore,W.instructional/W.leisure

Fiction	door, smile, love, girl, herself, walk, moment, voice, himself, mind, watch, mother, remember
Polit1	political, labour, economic, national, European, union, million, Britain, trade, authority, Prime
Health1	care, parent, patient, hospital, health, mother, person, experience, doctor, body, baby, drug
Spoken	yeah, Right, okay, actually, hundred, pound, twenty, alright, sort, nice, thank, fifty, thirty, anyway
Sports1	game, player, team, season, match, England, goal, race, minute, score, championship, ball, League
Biz1	income, rate, payment, contract, pension, benefit, amount, scheme, account, value, loan, tenant
Health2	patient, cell, disease, acid, protein, gene, datum, concentration, value, occur, energy, method
Comp	user, software, computer, module, datum, file, Corp, application, Unix, database, version, product
News	police, yesterday, voice, studio, John, crime, officer, speaker, prison, council, murder, video-tape
Spo2/Brit	club, Road, Darlington, Ireland, Royal, hour, award, Hall, town, Sunday, School, Middlesbrough
Travel	fish, animal, plant, land, bird, road, tree, site, river, specie, mile, wind, rock, aircraft, forest, tank
Biz2	price, firm, share, bank, profit, product, investment, industry, sale, management, customer, rate
History	church, century, John, king, Christian, England, Roman, Jesus, death, himself, Edward, land
Research	award, research, process, organisation, analysis, subject, datum, worker, institution, type, model
Linguistics	language, theory, sense, example, class, human, individual, kind, experience, society, structure
Leisure1	music, film, artist, painting, exhibition, American, sound, picture, song, band, John, collection
Polit2	Minister, Soviet, president, military, election, March, force, National, party, announce, leader
Legal	court, order, person, section, legal, matter, defendant, appeal, authority, shall, plaintiff, decision
Edu	education, student, teacher, course, staff, training, authority, Council, community, pupil, management
Leisure2	colour, garden, hotel, plant, food, design, flower, wine, knit, hair, holiday, light, wall, restaurant

category is given after a slash. Obviously not all cluster members share the same code, but the pattern is consistent. The following is the distribution of the BNC genre codes of documents belonging to Cluster 18 in Table 2:

N	BNC codes	N	BNC codes
424	W.fict.prose	5	W.letters.personal
49	W.misc	5	S.brdcast.discussn
43	S.oral.history	4	W.religion
29	W.fict.poetry	4	W.non.ac.medicine
26	W.biography	4	S.speech.unscripted
19	W.non.ac.soc	4	S.classroom
5	W.pop.lore	3	W.non.ac.humanities
5	W.essay.school	3	W.newsp.brdsht.social

The documents outside of W.fict.prose are still reasonably similar, as they include ADG (a book for teenage girls), ADM (accounts of travelling through Ireland), AP7 (experiences of old age), etc.[3]

The topics produced by the LDA package in R lack information about their relative size and consistency, so I gave them indicative labels on the basis of their keywords.

The two methods agreed on a number of domains, such as fiction, local British and international affairs. However, for some domains (underlined in Table 2) the clusters and topic models disagreed. One considerable difference stems from the ability of topic models to differentiate between the everyday language and professional discourse (the topics labelled as business and health care). In addition to this, hard clustering did not produce clusters related to history

[3]The ids are from the BNC index (Lee, 2001).

and research, which constitute relatively small (but consistent) topics in the BNC, e.g., documents HHY, HJ0, HPN, etc, containing research project applications and texts on project management. Also, LDA generated a cluster combining references to sport clubs with other local British texts, possibly generalising on the membership of their keywords to place names in the UK.

3.2. Comparing Internet corpora to the BNC

The same procedure was applied to ukWac and generated the clusters and topics reported in Table 3. The clusters offer a quick way into comparing the composition of ukWac to that of the BNC. For instance, there is a cluster without specific keywords (Cluster 0 in ukWac). This closely corresponds to the fiction cluster in the BNC (Cluster 18 in Table 2). However, unlike the BNC the ukWac cluster contains little fiction and mostly consists of diary-like blogs, discussion forums on everyday topics and chats.

Domains that reasonably match both the BNC and ukWac include political news, sports, health, business and religion. The differences concern a broader collection of topics in ukWac within each domain, as well as the presence of more modern texts. Clustering of ukWac detects two clusters with keywords related to computing, one of which (Cluster 7) is quite similar to Cluster 0 in the BNC. Another computing cluster of ukWac (15) belongs to the field of web-based communications (the longer list of its keywords also includes *HTML, Internet, click, Google*, etc).

RB clustering in Cluto seems to be considerably more efficient on a large corpus than LDA. Clustering of the entire ukWac took 1494 sec, while LDA was not able to deal with a selection of more than 500,000 documents from ukWac (on a computer with 4GB memory), and producing topic models even for this subset took 4896 sec (clustering of the same subset with Cluto took 304 sec).

Even though the clusters and the topic models in Table 3 are drawn from two slightly distributions, there is a broad similarity between their results. The clusters of music, movies, health, computing, communication, food and gardening, religion, politics, travel and business have compatible keywords.

Table 4 lists the clusters generated for itWac, an Italian web-corpus collected using the same compilation procedure as ukWac. It again showed broad compatibility with the domains of pages in itWac, including film (Cluster 0), food (1), blogs and chats (2 and 7), legal texts (3,9), music (4), education (5 and 17), computers (6), religion (8), sports (10), health (11), politics (12 and 13), business (14), culture (15), communication (16), travel (18) and local news (19). Some differences between the keywords in itWac and ukWac are related to the realities in individual countries (*Berlusconi* vs *Labour*). In ukWac there is a Do-It-Yourself cluster (4) and a cluster on NHS (10), while itWac contains two legal clusters (3 and 9), their longer keyword lists reveal that Cluster 9 is more about criminal cases, while Cluster 3 is about more general legislation.

3.3. Comparing comparable corpora

Energy corpora differ from the BNC or ukWac in their size (less than 10 million words). It was also expected that we can deal with fewer subdomains in them. The I_2 value was again not indicative, so the experiment was set for 15 clusters. As in the previous experiments both clustering and topic models were used. However, on a smaller corpus hard clustering produced less clearly interpretable results (also probably because of the difficulty in clear separation of relatively similar topics), while LDA produced fairly reasonable models listed in Table 5.

The analysis of topic models contradicted the original assessment made on the basis of term lists extracted from the corpora using the standard BootCat procedure (Baroni and Bernardini, 2004):[4]

7467 renewable energy	1508 возобновлять энергия
3127 greenhouse gas	1401 парниковый газ
3049 natural gas	2106 природный газ
2320 solar energy	2274 солнечный энергия
1994 carbon dioxide	1124 углекислый газ
1920 solar cell	2710 солнечный батарея
1529 fuel cell	862 топливный элемент

Given the overlap between the terms with similar frequencies and the fact that the corpora were collected with the set of keywords, it was expected that their content is sufficiently similar, and overall the corpora are good candidates for extracting and aligning term lists. However, LDA identified some topics present in both corpora, which are less ideal for term extraction, such as information for investors and news about utility companies, forums (we can expect less consistency in terminology used there) and legal texts. The Russian corpus was also contaminated with documents relating to computers (because of links to *power supply*), general news and student essays, which contain low-level introductions into a range of topics in this field. In general such lists of topics are useful in cleaning the corpus to achieve its greater consistency.

4. Conclusions

Three main outcomes of this study concern:

1. the possibility of rapid unsupervised assessment of the content of large corpora (clusters still need human interpretation, but this can be done quickly as long as they are consistent);
2. the possibility of comparing the content of corpora across languages without using any external resources, such as dictionaries;
3. the difference between hard clustering and LDA.

With respect to (1), we can use clustering to reveal more information even about a well-annotated corpus, such as the presence of a considerable cluster of texts for railway and aircraft enthusiasts in the BNC (otherwise obscurely coded as W.misc or W.leisure). It also shows broad classes of texts in a corpus of unknown composition such as ukWac or itWac.

With respect to (2), the clusters can be used to compare the content of individual corpora, but the interpretation needs to be done manually again. Intersection between the keywords of clusters in two corpora can be also done automatically.

[4]Lemmatisation of term elements in Russian affects the syntactic pattern of the composite term (Sharoff et al., 2008).

Table 3: Keywords clusters and topic models in ukWac

0	12.2%	I, her, he, she, my, his, me, him, i, it, do, say, PM, post, man, go, have, love, think, but
1	7.5%	road, mile, town, Canal, park, Road, walk, route, fn, Park, Street, Museum, village, north
2	2.5%	game, poker, player, ball, play, Sudoku, puzzle, sudoku, score, Games, win, goal, casino, match, tournament
3	4.6%	school, teacher, education, pupil, learn, skill, learning, child, student, training, teaching, learner
4	5.8%	water, cleared, colour, dive, light, surface, diver, valve, boat, wheel, wall, battery, bike, engine
5	10.0%	sector, organisation, local, business, development, management, sustainable, environmental
6	5.6%	student, University, module, research, study, course, academic, Studies, degree, science, graduate
7	6.3%	datum, system, software, model, use, computer, network, data, technology, user, NUN, solution
8	1.4%	insurance, mortgage, loan, property, Insurance, auto, quot, lender, rate, insurer, Agents
9	2.1%	God, Jesus, Christ, Christian, church, Lord, he, Church, Bible, his, sin, faith, prayer, verse, Him
10	3.1%	patient, care, NHS, health, nurse, hospital, clinical, nursing, mental, medical, service, Trust, Nursing
11	3.5%	plant, fish, bird, garden, flower, tree, food, specie, fruit, wine, seed, vegetable, cook, sauce, soil, sugar
12	3.1%	disease, drug, treatment, cancer, patient, cell, blood, infection, symptom, therapy, animal, gene, risk
13	3.2%	club, race, season, win, League, Cup, team, player, championship, match, lap, football, Championship
14	7.0%	pension, government, Committee, Labour, union, that, political, shall, Minister, vote, member
15	5.4%	file, search, user, text, server, use, Windows, web, page, site, library, directory, Web, browser
16	5.9%	customer, company, your, you, payment, business, sale, product, card, any, charge, market, service, price
17	3.5%	hotel, bedroom, room, holiday, beach, accommodation, apartment, restaurant, bathroom, cottage
18	4.8%	music, song, band, album, guitar, sound, musical, dance, vocal, gig, jazz, bass, track, play, concert
19	2.4%	film, movie, cinema, camera, DVD, comedy, actor, star, Jolen, scene, Hollywood, character

Music	music, song, game, player, band, album, club, season, sound, guitar, ball, score, match, goal, love
Movies	film, game, seem, never, something, quite, story, love, watch, race, turn, friend, movie, though
Health1	patient, care, treatment, drug, hospital, medical, doctor, clinical, woman, nurse, disease, mental
Comp	file, image, user, program, value, type, text, code, object, datum, function, display, click, version
History	John, William, Royal, James, history, David, George, King, Thomas, century, Robert, Peter, Hall
Goods	poker, wedding, game, price, artist, film, online, gift, wine, food, card, exhibition, colour, puzzle
Health2	cell, disease, protein, model, gene, effect, blood, datum, test, human, patient, food, cause, analysis
Garden	water, plant, animal, fish, tree, bird, specie, energy, garden, food, land, soil, farm, farmer, crop, waste, climate
Travel	room, hotel, walk, mile, road, town, house, bedroom, village, boat, holiday, garden, accommodation
Comm	Internet, software, user, customer, network, computer, technology, online, phone, server, mobile
Soc	social, language, political, history, human, society, theory, century, culture, word, text, idea, cultural, seem
Biz1	sector, market, organisation, industry, investment, million, management, financial, growth, cent, opportunity
Science	Research, library, resource, Institute, Centre, Science, publication, journal, Society, College, Department
Polit1	Iraq, Labour, political, attack, police, force, party, election, British, vote, military, weapon, Minister
Polit2	management, Committee, consultation, proposal, ensure, environmental, authority, safety, risk, review
Biz2	insurance, loan, payment, property, rate, mortgage, credit, claim, pension, income, employer, employee
Edu	skill, teacher, education, pupil, teaching, training, module, teach, learning, School, academic, language
Legal	shall, person, payment, court, request, party, apply, legal, notice, agreement, provision, liability, register
Local	Centre, Community, Road, meeting, West, council, volunteer, East, County, Trust, Borough, resident, District
Religion	Jesus, church, Christian, love, Christ, Lord, word, shall, faith, Bible, never, death, spirit, himself, heart

With respect to (3), clustering is a better approach to web-sized corpora, as LDA cannot reasonably handle ukWac and it takes considerably more time on the BNC or itWac. At the same time, on smaller corpora, LDA detects models which are easier to interpret in comparison to clustering solutions.

Acknowledgements

This research was funded by European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no 248005 (TTC).

5. References

S.F. Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in Wikipedia. In *Proc. 11^th EACL*, pages 62–69, Trento.

Marco Baroni and Silvia Bernardini. 2004. Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC2004*, Lisbon.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

H. Blancafort, B. Daille, T. Gornostay, U. Heid, C. Mechoulam, and S. Sharoff. 2010. TTC: Terminology extraction, translation tools and comparable corpora. In *Proc. EURALEX2010*, Leeuwarden, 5-6 July.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proc. Neural Information Processing Systems*.

Hideo Joho and Mark Sanderson. 2004. The SPIRIT collection: an overview of a large web collection. *SIGIR Forum*, 38(2):57–61.

Adam Kilgarriff. 2001. Comparing corpora. *International*

Table 4: Keywords in clusters of itWac

0	2.59%	film, cinema, regista, attore, personaggio, regia, scena, cinematografico, protagonista, pellicola, cast, storia
1	1.07%	vino, olio, cucchiaio, uovo, burro, cuocere, pasta, piatto, latte, zucchero, formaggio, pepe, farina, cucina
2	11.53%	mio, mi, che, suo, non, essere, dire, avere, io, uomo, tuo, fare, lui, quello, ti, ma, occhio, cosa, amore
3	13.99%	articolo, numero, comma, decreto, legge, cui, previsto, lavoro, relativo, presente, servizio, lavoratore
4	2.37%	musica, disco, canzone, album, musicale, brano, band, concerto, rock, suonare, chitarra, musicista, cd
5	3.19%	scuola, scolastico, docente, insegnante, alunno, classe, formativo, didattico, formazione, istruzione, educativo
6	1.71%	file, server, Windows, utente, software, Microsoft, versione, programma, web, browser, utilizzare, Linux
7	7.39%	mi, ciao, io, non, messaggio, ti, mio, avere, fare, se, inviare, ma, me, scrivere, dire, sapere, cosa, tuo, Posted
8	3.86%	chiesa, Dio, Gesù, Cristo, cristiano, santo, papa, fede, suo, cattolico, vescovo, religioso, uomo, padre
9	6.79%	articolo, emendamento, commissione, legge, esame, numero, relatore, corte, sentenza, comma, giudice
10	4.27%	squadra, gioco, giocatore, gara, partita, giocare, atleta, campionato, calcio, vincere, gol, atletico, tifoso, sport
11	3.01%	paziente, malattia, cellula, medico, farmaco, terapia, clinico, medicina, tumore, patologia, salute, sanitario
12	5.87%	guerra, Iraq, americano, pace, Bush, contro, palestinese, militare, terrorismo, popolo, israeliano, iracheno
13	7.56%	politico, governo, partito, europeo, paese, presidente, politica, Berlusconi, parlamento, maggioranza, voto
14	3.91%	banca, euro, mercato, consumatore, milione, prezzo, aumento, miliardo, sindacato, paese, lavoratore, Cgil
15	3.89%	teatro, arte, mostra, artista, opera, libro, spettacolo, romanzo, poesia, museo, autore, storia, scena, artistico
16	3.20%	Internet, rete, utente, software, prodotto, tecnologia, sito, servizio, cliente, web, elettronico, azienda, digitale
17	6.67%	università, ricerca, scientifico, studio, corso, universitario, laurea, professore, scienza, concorso, biblioteca
18	3.14%	Hotel, mare, hotel, acqua, albergo, isola, zona, situare, vacanza, spiaggia, metro, antico, città, km, ristorante
19	3.96%	Provincia, assessore, comune, sindaco, provinciale, parco, regione, comunale, territorio, area, cittadino

Journal of Corpus Linguistics, 6(1):1–37.

David Lee. 2001. Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Paul Rayson, Damon Berridge, and Brian Francis. 2004. Extending the Cochran rule for the comparison of word frequencies between corpora. In *Proc 7th International Conference on Statistical analysis of textual data (JADT 2004)*, pages 926–936, Louvain-la-Neuve.

Serge Sharoff, Mikhail Kopotev, Tomaž Erjavec, Anna Feldman, and Dagmar Divjak. 2008. Designing and evaluating a Russian tagset. In *Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008*, Marrakech.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini, editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

Ying Zhao and George Karypis. 2004. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331.

Table 5: Comparing English and Russian energy corpora

1	Table, Equipment, Market, Consumption, Capacity, Production, Industry, Generation, Distribution, Transformers, Sector
2	earth, atmosphere, dioxide, surface, cool, cause, warming, fluid, radiation, methane, reservoir, human, rock, warm
3	reactor, Nuclear, uranium, radioactive, barrel, mine, Uranium, fission, cent, Petroleum, reserve, billion, mining, safety
4	ocean, wave, OTEC, Ocean, tide, Intel, Tidal, Wave, marine, Hawaii, offshore, Conversion, device, surface, conversion
5	Commission, Public, shall, bill, Utility, credit, Federal, contract, FERC, eligible, District, regulation, federal, County
6	distribute, consumer, distribution, network, peak, meter, period, investment, value, datum, average, sector, reliability, factor
7	speed, rotor, blade, field, magnetic, shaft, circuit, wire, engine, transformer, phase, connect, rotate, frequency, torque
8	cogeneration, hydrogen, ethanol, engine, wood, combustion, Biomass, boiler, burn, crop, convert, landfill, residue, gasoline
9	Green, News, Hydro, Business, India, Stock, Development, Sustainable, Geothermal, Alternative, Environmental
10	river, hydroelectric, hydro, reservoir, fish, River, head, blade, hydroelectricity, Hydro, Hydropower, height, stream
11	post, read, bill, want, look, article, problem, money, green, really, question, link, warming, news, storey, idea, What, talk
12	announce, billion, Tags, investment, megawatt, Kansas, expect, sign, April, News, green, California, feed-in, release, Farm
13	sustainable, policy, economic, sector, reduction, Development, management, national, international, community
14	module, light, silicon, inverter, watt, sunlight, hour, roof, saving, device, appliance, save, film, array, tower, house, electron
15	Program, National, Department, California, Center, Association, Resources, Public, Information, Efficiency, Service, page
1	новость (news), ноутбук (laptop), процессор (processor), комментарий (comment), компьютер (computer), телефон (phone), ученый (scientist), рубрика (heading), память (memory), мобильный (mobile), intel, energy
2	финансовый (financial), государство (state), инвестиция (investment), мера (measures), доля (proportion), политика (politics/strategy), правительство (government), национальный (national), потребность (demand)
3	ооо (ltd), электронный (electronic), бесперебойный (uninterruptible), продажа (sale), дизельный (diesel), купить (buy), ремонт (repair), чертеж (drawing), ибп (UPS), фирма (firm), каталог (catalogue), товар (goods)
4	вот (this/yeah), кто (who), да (yes), сделать (do), сейчас (now), говорить (say), там (there), ни (neither), ли (would), надо (should), просто (simply), сам (-self), знать (know)
5	коммунальный (utilities), жилой (resident), федеральный (federal), водоснабжение (water supply), отопление (heating), оплата (payment), жкх (utilities), помещение (room), энергосбережение (energy saving)
6	рф (Russia), рао еэс (Unified Energy System), подробно (details), инвестиционный (investment), директор (director), правительство (government), президент (president), глава (chairman), новость (news), январь (January)
7	конференция (conference), выставка (exhibition), устойчивый (sustainable), научный (scientific), наука (science), охрана (protection), энергосбережение (energy saving), энергоэффективность (energy efficiency)
8	геотермальный (geothermal), биомасса (biofuel), водород (hydrogen), отходы (waste), возобновляемый (renewable), ветроэнергетика (wind generation), топливный (fuel), сжигание (burning), вэу (wind farm), солнце (sun)
9	море (sea), глобальный (global), океан (ocean), растение (plant), лес (forest), загрязнение (pollution), потепление (warming), природа (nature), парниковый (greenhouse), животное (animal), организм (organism), планета (planet)
10	плотина (dam), добыча (production), сооружение (installation), водохранилище (reservoir), месторождение (deposit), запас (resource), очистка (purification), сырье (raw materials), сточный (sewage)
11	аккумулятор (battery), воздушный (air), насос (pump), емкость (capacity), нагрузка (load), рисунок (drawing), ротор (rotor), конструкция (design), вал (shaft), корпус (body), вращение (rotation), мм (mm), охлаждение (cooling)
12	тэц (CHP), энергосистема (grid), газотурбинный (gas turbine), когенерация (cogeneration), киловольт (kV), нагрузка (load), котельная (boiler station), генерация (generation), теплоснабжение (heating supply), выработка (production)
13	паровой (steam), котел (boiler), силовой (power), трансформатор (transformer), гост (GOST), частота (frequency), пар (steam), линия (line), power, преобразователь (converter), передача (transmission), реактивный (reactive), переменный (alternating), провод (wires)
14	русский (Russian), улица (street), война (war), язык (language), московский (Moscow), республика (republic), школа (school), век (century), ребенок (child), франция (France), текст (text), германия (Germany), игра (game), километр, книга, транспорт, история, культура,
15	реферат (essay), реактор (reactor), движение (movement), ядро (nucleus), наука (science), поле (field), атом (atom), реакция (reaction), нейтрон (neutron), физика (physics), планета (planet), магнитный (magnetic)

Konkordančnik za govorni korpus GOS

Darinka Verdonik,* Ana Zwitter Vitez,† Miro Romih‡, Simon Krek‡

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova 17, 2000 Maribor
darinka.verdonik@uni-mb.si
† Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, 4220 Škofja Loka
ana.zwitter@guest.arnes.si
‡ Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik
miro.romih@amebis.si, simon.krek@guest.arnes.si

Povzetek

Prispevek predstavlja funkcije konkordančnika GOS za iskanje po referenčnem govornem korpusu slovenščine. Konkordančnik bo od oktobra 2010 na voljo na spletnem naslovu www.korpus-gos.net zainteresiranim uporabnikom. Bistvena prednost konkordančnika pred drugimi sorodnimi jezikovnimi orodji je, da ponuja povezavo med zvokom in zapisom: vsako konkordanco lahko tudi poslušamo v kontekstu. Konkordančnik izkorišča podatke, vsebovane v korpusnem gradivu, tudi za zelo poglobljeno, specifično iskanje zahtevnih uporabnikov, hkrati pa se skuša približati manj zahtevnim potencialnim uporabnikom z enostavnim in hitro usvojljivim načinom uporabe.

Concordancer for the speech corpus GOS

The paper presents the functions of the GOS concordancer designed for browsing the Slovenian reference speech corpus. The concordancer will be available on the web address www.korpus-gos.net from October 2010. The main advantage of the concordancer compared to other similar language tools is its link between the sound and the transcription: every concordance can be listened to in its context. The concordancer uses all the metadata from the corpus to enable complex search function for highly-demanding users, but at the same time tries to attract less demanding potential users with a simple, user friendly interface.

1. Uvod

Potreba po izdelavi referenčnega govornega korpusa slovenščine je bila v zadnjem desetletju v slovenskem jezikoslovnem prostoru večkrat eksplicitno izražena. Prvi je na to opozoril Stabej, ko je pri predstavitvi besedilnovrstne sestave korpusa FIDA poudaril, da bi bil "seveda v slovenskem prostoru še bolj dragocen korpus, ki bi vseboval tudi govorjena besedila" (Stabej, 1998: 100); ideja je bila podrobneje predstavljena l. 2000 (Stabej, Vitez, 2000: 79). Tudi Weiss (2001: 422) je poudaril nujnost vključevanja govorjenih besedil v elektronsko zbirko. Gorjanc je pozival k začetku gradnje: "Čim prej bi bilo treba oblikovati skupino, ki bi začela s pripravami govornega dela korpusa." (Gorjanc, 2005: 53) Tudi v okviru dialektoloških študij je novo tisočletje vzbudilo pričakovanja po govornem korpusu: "/…/ širitev korpusa na spontani nejavni govor vzbuja upanje na drugačne čase" (Kenda Jež, 2004: 271), v okviru govornih tehnologij pa je bila potreba po vključitvi spontanega govora v raziskave izražena med drugim v Verdonik (2006: 40).

Teoretična izhodišča gradnje govornega korpusa, preizkušena na manjšem učnem govornem korpusu, so bila izdelana l. 2007 (Zemljarič Miklavčič, 2007; 2008). Leta 2008 so bila v okviru projekta *Sporazumevanje v slovenskem jeziku (SSJ, 2008−2013)*[1] zagotovljena sredstva za izgradnjo govornega korpusa slovenščine v obsegu 1 milijona besed ali 110 ur govora[2], s čimer so bili natanko desetletje po prvem pozivu h gradnji izpolnjeni vsi pogoji za začetek. Tekom dela je korpus dobil ime GOS (korpus GOvorjene Slovenščine) in bo konec leta 2010 v celoti dokončan in na voljo uporabnikom. Prve predstavitve njegove zasnove najdemo v Zemljarič et al. (2009) in Zwitter Vitez et al. (2009).

Enako kot sam korpus je za njegovo širšo dostopnost in uporabo pomemben tudi spletni vmesnik – konkordančnik – za iskanje po korpusu. Zapise iz govornega korpusa bi seveda lahko vključili tudi v konkordančnik za pisni korpus (za slovenščino npr. konkordančniki korpusov Nova beseda ali FidaPLUS), vendar bi s tem izgubili mnoge posebnosti govornega materiala, zlasti povezavo med zvokom in zapisom, govorni korpus GOS pa ima še eno pomembno posebnost, ki bi se v konkordačniku pisnega korpusa prav tako izgubila: dvojni zapis govora, v pogovorni in standardizirani različici.

Sofinanciranje projekta izdelave konkordančnika, prirejenega posebej za govorni korpus GOS, je jeseni 2009 odobrilo Ministrstvo za visoko šolstvo, znanost in tehnologijo ob sofinanciranju iz Evropskega regionalnega sklada, in tako bo od jeseni 2010 na spletnem naslovu www.korpus-gos.net na voljo konkordančnik GOS, s katerim bo mogoče spočetka iskati po delu korpusnega gradiva, po celotnem gradivu zaključenega korpusa GOS pa od januarja 2011.

V tem prispevku se bomo po uvodnem kratkem pregledu nekaterih sorodnih tujih konkordančnikov in

[1] http://www.slovenscina.eu/

[2] Lastnik korpusa GOS je Ministrstvo za šolstvo in šport Republike Slovenije na podlagi pogodbe »Pogodba o sofinanciranju izvedbe projekta Sporazumevanje v slovenskem jeziku v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013«, št. pogodbe 3311-08-986003, sklenjene med Republiko Slovenijo, Ministrstvom za šolstvo in šport, ter podjetjem Amebis, d.o.o., Kamnik.

skupin potencialnih uporabnikov osredotočili predvsem na predstavitev iskalnih funkcij in informacij o gradivu, ki jih bo ponujal konkordančnik GOS.

2. Pregled nekaterih tujih sorodnih konkordančnikov

Britanski nacionalni korpus (http://www.natcorp.ox.ac.uk/) je prosto dostopen na spletu, vendar govorna podsekcija zajema samo transkripcije brez izvornih zvočnih posnetkov. Obsega 10 milijonov besed. Iskanje po korpusu je med drugim omogočeno z vmesnikom XAIRA, s pomočjo katerega dobimo informacije o frekvencah besed, slovničnih struktur in kolokatorjev ter primerih konkordanc. Iskanje omogoča po celotnem korpusu ali omejenem naboru gradiva.

Tudi pri ruskem nacionalnem korpusu (http://www.ruscorpora.ru/en/index.html) je govorna podsekcija korpusa dostopna samo z zapisom, brez zvoka. Iskanje je mogoče po obliki, kanalu, slovničnih ali semantičnih značilnostih oz. po metapodatkih o govorcih in diskurzu.

Spletni konkordančnik omogoča tudi brskanje po francoskem Corpus de la parole (http://www.corpusdelaparole.culture.fr/). Iščemo lahko po jeziku (zastopani so namreč vsi govorjeni jeziki v Franciji), po besedah ali frazah ali po metapodatkih o diskurzu in govorcih.

Pregled še ostalih tujih korpusov kaže, da tako kot navedeni večinoma[3] ne omogočajo dostopa do zvočnih posnetkov, pri iskanju pa izkoriščajo možnosti, ki jih ponuja korpusno gradivo (poleg osnovnega iskanja po besedah/frazah še iskanje po jezikovnih oznakah na različnih ravneh ter metapodatkih o govorcih in diskurzih).

3. Uporabniki konkordančnika GOS

Pri zasnovi slovenskega spletnega konkordančnika za iskanje po govornem korpusu GOS smo upoštevali predvidene potrebe naslednjih potencialnih skupin uporabnikov konkordančnika:

- raziskovalci – zahtevni uporabniki, ki jim je treba omogočiti čim več iskalnih funkcij in podatkov o gradivu
- izobraževanje – v izobraževanju (pri učenju slovenskega jezika) so pomembni predvsem enostavnost iskanja, kratka, jasna in priročna navodila ter grafična privlačnost konkordančnika
- poklici v stiku z govorom (razni pisci, tolmači in prevajalci, poklicni govorci...) – po eni strani enostavnost in privlačnost uporabe, po drugi strani pa vseeno ustrezno pester nabor iskalnih funkcij, zlasti glede na metapodatke o govorcih in diskurzih

Glede na ta predvidevanja ter glede na podatke, ki jih vključuje korpusno gradivo, in z navezavo na predvideni prenovljeni konkordančnik za pisni korpus slovenskega jezika, ki nastaja prav tako v okviru projekta Sporazumevanje v slovenskem jeziku, bo imel konkordančnik GOS iskalne funkcije, kot so opisane v sekciji 4.

4. Iskalne funkcije konkordančnika GOS

Konkordančnik podpira dva osnovna tipa prikaza rezultatov: v obliki konkordančnega niza ali v obliki seznama. Več o prikazu rezultatov je v naslednjem poglavju, v nadaljevanju pa so opisane iskalne funkcije konkordančnika, ki jih delimo na:

- iskanje po konkordančnem nizu (enostavno iskanje, napredno iskanje)
- iskanje po seznamu

Enostavno iskanje omogoča naslednje načine iskanja:

Enostavno iskanje	
po pogovornem zapisu	***po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"
filter po govorcih in diskurzih	

Tabela 1: Enostavno iskanje.

Z enostavnim iskanjem lahko iščemo poljubno po pogovornem ali po standardiziranem zapisu govora v korpusu. Razlika med enim in drugim zapisom je naslednja:

- **Pogovorni zapis:** Govor je zapisan v veljavnem slovenskem črkopisu (ni fonetični zapis) in upoštevane so veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pa je pri tem kolikor mogoče verno predstavljena glasovna podoba govora. Nekaj primerov: *tud, neki, mislm, prov, navm, najraj b vidu...*
- **Standardizirani zapis:** Pri pretvorbi v standardizirani zapis so odpravljene glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Ciljna oblika je knjižna različica istega leksema. Na drugih jezikovnih ravneh besede niso spremenjene. Če določenega leksema ni v knjižni normi, je ohranjen v obliki, ki se pojavlja v govoru. Nekaj primerov: *tudi, nekaj, mislim, prav, ne bom, najraje bi videl, štengica, sva šle...*

Primer iste izjave v pogovornem in standardiziranem zapisu:

P: a	*veš*	*boš*	*mov*	*veš*	*kok*	*boš*
S: a	*veš*	*boš*	*imel*	*veš*	*koliko*	*boš*

P: traku zgubu	*al*	*kva*	*maš*	*not*
S: traku izgubil	*ali*	*kaj*	*imaš*	*notri*

Pri pogovornem zapisu je mogoče samo iskanje po kanalu besede (iščemo samo in točno tisto obliko, ki smo ji vpisali), medtem ko je pri iskanju po standardiziranem zapisu privzeto iskanje po kanalu leme, kar pomeni, da se iskana beseda avtomatsko pretvori v osnovno obliko in se prikažejo rezultati za vse oblike te besede. Šele če vpišemo iskani niz med narekovaji, iščemo po kanalu besede, torej samo in točno tisto obliko, ki smo ji vpisali.

[3] Poslušanje zadetkov omogoča npr. nemški vmesnik (http://dsav-oeff.ids-mannheim.de/DSAv/SUCHMASK.H_TM), mnogi drugi pa ne, npr. češki - http://ucnk.ff.cuni. cz/english/hledat_v_cnk.php, italijanski - http://badip.uni-graz.at/, poljski – http://korpus.ia.uni.lodz.pl/conversational/...

Potem ko se prikažejo rezultati iskanja, lahko le-te **filtriramo po tipih diskurzov/govorcev:** ta funkcija omogoča iskanje znotraj vključenih metapodatkov o diskurzih in govorcih.

Metapodatki o govorcih vključujejo naslednje izbire:

- spol
- starost
- izobrazba
- regionalna pripadnost:
    - enotna (posameznik je vse življenje preživel v eni regiji)
    - razpršena (posameznik je del življenja [vsaj eno leto: šolanje, služba, selitev...] preživel v kateri drugi regiji/-ah)
- prvi jezik (slovenski ali tuji)

Regionalna pripadnost je v korpusu GOS označena glede na večja regionalna mestna središča, h katerim gravitira posamezno področje in ki sovpadajo z registrskimi območji v Sloveniji (MB, MS, SG, CE, KK, NM, LJ, KR, PO, GO, KP), oz. glede na pripadnost regijam zunaj Slovenije (zamejski Slovenci: Italija, Avstrija, Madžarska, življenje v tujini). V skladu s temi označbami so tudi iskalne možnosti v konkordančniku.

Metapodatki o diskurzih vključujejo izbire po:

- regiji snemanja
- letu snemanja
- klasifikaciji diskurza, kot prikazuje tabela 3

Tip diskurza	Kanal	Govorni dogodek
javni informativno-izobraževalni	televizija	novinarski prispevek
		moderirani pogovor
	radio	moderirani program
		moderirani pogovor
	osebni stik	osnovnošolska učna ura
		srednješolska učna ura
		tečaj
		javno predavanje
		fakultetno predavanje
javni razvedrilni	televizija	moderirani pogovor
		moderirana oddaja
		resničnostni šov
		športni prenos
	radio	moderirani program
		moderirani pogovor
nejavni nezasebni	osebni stik	formalni delovni sestanek
		neformalni delovni sestanek
		konzultacija na fakulteti
		storitev
		razgovor
	telefon	storitev
nejavni zasebni	osebni stik	pogovor v družini
		pogovor med prijatelji/znanci
	telefon	pogovor v družini
		pogovor med prijatelji/znanci

Tabela 3: Pregled iskanj glede na klasifikacijo diskurzov.

Za zahtevnejše uporabnike je na voljo napredno iskanje:

Napredno iskanje	
iskanje po bližini (1 in 2)	
iskanje po oblikoslovnih oznakah (samo 2)	
1 po pogovornem zapisu	***2 po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"
filter po govorcih in diskurzih	

Tabela 2: Napredno iskanje.

Pri naprednem iskanju lahko poleg iskalnih funkcij, ki jih omogoča že enostavno iskanje (torej iskanje po pogovornem ali po standardiziranem zapisu in filtriranje rezultatov glede na podatke o govorcih in/ali podatke o diskurzih), uporabljamo še dve funkciji:

- **Iskanje po bližini:** Omogoča iskanje glede na bližino/nebližino drugih besed. Iščemo lahko v okolici do 5 besed.
- **Iskanje po oblikoslovnih oznakah:** Samo pri iskanju po standardiziranem zapisu lahko iskano besedo tudi dodatno omejimo z besedno vrsto in ostalimi oblikoslovnimi oznakami (iščemo po kanalu oblikoslovnih oznak). Ker bo korpus GOS v prvi fazi samo avtomatsko označen z označevalnikom, naučenim na pisnih besedilih, še ni znano, kolikšna bo natančnost kanala z oblikoslovnimi oznakami.

Drugi način iskanja po korpusu GOS je iskalna funkcija seznam:

Seznam	
iskanje z nadomestnimi znaki (1 in 2)	
iskanje po oblikoslovnih oznakah (samo 2)	
1 po pogovornem zapisu	***2 po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"

Tabela 4: Pregled iskalne funkcije seznam.

Funkcija seznam omogoča iskanje z nadomestnimi znaki, in sicer naslednjimi:

- znak * nadomešča poljubno število znakov

- znak ? nadomešča en znak

Enako kot pri iskanju po konkordančnem nizu je mogoče izbirati med iskanjem po pogovornem in po standardiziranem zapisu, pri slednjem je na voljo tudi filtriranje po oblikoslovnih oznakah. Filtriranje po tipih diskurzov in govorcev v prvi fazi ne bo omogočeno, je pa predvidena tovrstna nadgradnja naknadno.

Rezultati iskanja pri seznamu niso prikazani v obliki konkordanc, tako kot pri iskanju po konkordančnem nizu, ampak v obliki seznama besed s podatki o frekventnosti posamezne besede in njeni standardni obliki.

Iskalna načina seznam in konkordančni niz sta med seboj povezana: s klikom na besedo v seznamu se v zavihku konkordančni niz izdela ustrezen konkordančni niz, ki vsebuje primere rabe v kontekstu za besedo s seznama.

5. Prikaz zadetkov in podatki o njih

Rezultati iskanja se prikažejo:

- v obliki seznama konkordanc pri iskanju po konkordančnem nizu
- v obliki seznama besed s podatki o frekvenci in standardizirani obliki pri iskanju po seznamu

Rezultati iskanja pri konkordančnem nizu so:

- prikazani v pogovornem zapisu
- označeni glede na tip diskurza (JI – javni informativno-izobraževalni, JR – javni razvedrilni, NN – nejavni nezasebni, NZ – nejavni zasebni)

Desno od vsake konkordance je zvočnik. S klikom nanj lahko poslušamo eno ali več izjav, v kateri(h) je bil izgovorjen iskani niz.

Če želimo več podatkov o posamezni konkordanci, kliknemo na konkordanco. Prikažejo se naslednji podatki:

- razširjeni kontekst v pogovornem zapisu (tj. izjava ali izjave, v kateri(h) je bil izgovorjen iskani niz + 1 predhodna + 1 naslednja izjava)
- podatki o diskurzu, iz katerega je konkordanca (tip diskurza, kanal, opis govornega dogodka, regija, kjer je potekal diskurz, datum in čas poteka diskurza, vir posnetka, opis diskurza)
- podatki o govorcu (spol, starost, izobrazba, regionalna pripadnost, prvi jezik)

Če želimo še več informacij, imamo na voljo:

- da poslušamo razširjeni kontekst
- da pogledamo standardizirani zapis razširjenega konteksta
- da pogledamo korpusne oznake (lema, oblikoslovne oznake)
- da shranimo podatke v odložišče

Rezultate lahko tudi:

- sortiramo
- izvozimo

Pri urejanju rezultatov iskanja lahko le-te razvrstimo:

- glede na jedrno besedo/besede
- glede na levo sobesedilo
- glede na desno sobesedilo

6. Zaključek

Prispevek predstavlja funkcije konkordančnika GOS za iskanje po referenčnem govornem korpusu slovenščine. Konkordančnik bo od oktobra 2010 na voljo na spletnem naslovu www.korpus-gos.net zainteresiranim uporabnikom. Bistvena prednost konkordančnika pred drugimi sorodnimi jezikovnimi orodji je, da ponuja povezavo med zvokom in zapisom: vsako konkordanco lahko tudi poslušamo v kontekstu. Konkordančnik izkorišča podatke, vsebovane v korpusnem gradivu, tudi za zelo poglobljeno, specifično iskanje zahtevnih uporabnikov, hkrati pa se skuša približati manj zahtevnim potencialnim uporabnikom z enostavnim in hitro usvojljivim načinom uporabe. Ob tem bo seveda opremljen s priročnimi navodili, priročniki, videom ... o uporabi in bo na voljo tudi v angleški različici.

V prihodnosti bo poleg funkcionalnosti konkordančnika za uporabo korpusa GOS ključna seveda tudi sama obsežnost in opremljenost korpusnega gradiva. Korpus bo ob zaključku v okviru sedanjega financiranja obsegal 1 mio. besed. V tujini referenčni govorni korpusi že rastejo na 10 mio. besed ali več (angleški, poljski, nizozemski, portugalski...), in tudi za slovenščino je ob sedanjem obsegu gradiva izraz »referenčni« le pogojno upravičen, saj je v 1 mio. besed avtentičnih pogovorov nemogoče zajeti vso pestrost govorjene slovenščine. Za rabo v jezikoslovju in jezikovnih tehnologijah pa bi si poleg večjega obsega želeli tudi dodano opremljenost gradiva, npr. skladenjsko označevanje, fonetični zapis ipd.

7. Literatura

Gorjanc, V., 2005. *Uvod v korpusno jezikoslovje*. Domžale, Izolit.

Kenda Jež, K., 2004. Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju. Kržišnik, E. (ur.), *Obdobja 22*. Ljubljana, Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko: 263−276.

Stabej, M., Vitez, P., 2000. KGB (korpus govorjenih besedil) v slovenščini. *Informacijska družba IS'2000: Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan.

Stabej, M., 1998. Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6.

Verdonik, D., 2006. *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora*. Doktorska disertacija, Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Weiss, P., 2001. Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU: utemeljitev. *Jezikoslovni zapiski* 7, 1−2. Ljubljana, Založba ZRC: 419−428.

Zemljarič Miklavčič, J., 2007. *Načela oblikovanja govornega korpusa za slovenščino*. Doktorska disertacija, Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Zemljarič Miklavčič, J., 2008. *Govorni korpusi*. Univerza v Ljubljani, Filozofska fakulteta.

Zemljarič Miklavčič, J., Stabej, M., Krek, S., Zwitter Vitez, A., 2009. Kaj in zakaj v referenčni govorni korpus slovenščine. Stabej, M. (ur.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani: 437–442

Zwitter Vitez, A., Zemljarič Miklavčič, J., Stabej, M., Krek, S., 2009. Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Stabej, M. (ur.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani: 437–442

Samodejno razpoznavanje govorcev v prostoru supervektorjev

Boštjan Vesnicer[†], Jerneja Žganec Gros[†], France Mihelič[‡]

[†]Alpineon d.o.o.
Ulica Iga Grudna 15, SI-1000 Ljubljana
{bostjan,jerneja}@alpineon.si

[‡]Univerza v Ljubljani
Fakulteta za elektrotehniko
Tržaška 25, SI-1000 Ljubljana
france.mihelic@fe.uni-lj.si

Povzetek

V prispevku predstavljamo postopke samodejnega razpoznavanja govorcev, ki temeljijo na predpostavki, da lahko posameznega govorca obravnavamo kot točko v visoko razsežnem prostoru, ki mu pravimo prostor supervektorjev. Na tej predpostavki temelji večina trenutno najbolj učinkovitih sistemov za razpoznavanje govorcev. V prispevku podrobneje predstavimo tudi konkreten primer takega sistema, katerega učinkovitost ovrednotimo na uveljavljeni zbirki za vrednotenje sistemov razpoznavanja govorcev.

1. Uvod

Problem razpoznavanja govorcev spada na področje biometrije, zato ima veliko skupnega z večino drugih biometričnih postopkov, med katere uvrščamo tudi razpoznavanje obrazov, prstnih odtisov in šarenice. Ker pa je govorni signal — v nasprotju s "slikovnimi" signali — predstavljen s časovno vrsto poljubne dolžine, ga moramo obravnavati nekoliko drugače kot sliko obraza, prstnega odtisa ali šarenice.

Eden izmed najbolj uveljavljenih načinov za modeliranje časovnih vrst so *prikriti Markovovi modeli* (angl. hidden Markov models, HMM). Ti predstavljajo osnovne gradnike v praktično vseh omembe vrednih sistemih za samodejno razpoznavanje govora, zato se zdi logično, da bi morali biti modeli HMM najprimernejši tudi pri sorodnem problemu razpoznavanja govorcev. Ker pa je vrstni red akustičnih dogodkov — za razliko od razpoznavanja govora — za (tekstovno neodvisno) razpoznavanje govorcev v večji meri nepomemben, je fleksibilnost, ki nam jo ponujajo modeli HMM, nepotrebna. Zaradi tega se je na področju razpoznavanja govorcev uveljavil enostavnejši model *mešanice Gaussovih porazdelitev* (angl. Gaussian mixture model, GMM)[1].

Model GMM je parametrična verjetnostna porazdelitev, sestavljena iz poljubnega števila K Gaussovih porazdelitev. V večrazsežnem primeru lahko model GMM zapišemo kot:

$$f(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (1)$$

kjer smo z λ označili vse parametre (apriorne verjetnosti π_k, povprečne vektorje $\boldsymbol{\mu}_k$ in kovariančne matrike $\boldsymbol{\Sigma}_k$), ki nastopajo v modelu.

Osnovna predpostavka razpoznavanja govorcev s pomočjo modelov GMM je, da lahko vsakega govorca predstavimo kot točko v prostoru parametrov modela GMM. Ko želimo preveriti identiteto nekega govorca, moramo znati izračunati le razdaljo med dvema točkama v tem prostoru.

V praksi se je izkazalo, da je bolje, če govorca predstavimo le s podmnožico vseh parametrov modela GMM — povprečnimi vektorji. Če le-te zapišemo enega za drugim v nov vektor, dobimo visokorazsežen vektor, za katerega se je prijelo ime *supervektor*[2].

2. Splošni model govorca

Če želimo uporabiti model GMM za razpoznavanje govorcev, moramo poiskati način, kako iz govornega signala poljubne dolžine oceniti parametre modela GMM. (Tukaj predpostavljamo, da je govorni signal predstavljen v obliki niza vektorjev značilk, npr. značilk melodičnega kepstra).

Čeprav bi lahko parametre modela GMM ocenili neodvisno za vsakega govorca posebej, je smiselneje, če to storimo posredno preko modela GMM, ki smo ga naučili z uporabo govorne zbirke, v kateri imamo na voljo veliko število posnetkov čim večjega števila govorcev. Takemu modelu pravimo splošni model govorca (angl. universal background model, UBM) (Reynolds et al., 2000). Parametre modela UBM ocenimo po kriteriju *največjega verjetja* (angl. maximum likelihood, ML) z iterativnim postopkom *maksimizacije upanja* (angl. expectation maximization, EM) (Dempster et al., 1977), katerega lastnost je, da zagotavlja monotonost konvergence k lokalnemu maksimumu in je zaradi tega občutljiv na začetne vrednosti.

Za razliko od modela UBM, za učenje katerega imamo na voljo poljubno veliko podatkov, imamo za oceno modela GMM posameznega govorca na voljo le omejeno količino podatkov, tipično od deset sekund do nekaj minut dolg govorni posnetek. V takem primeru, ko je količina podatkov premajhna za zanesljivo oceno parametrov, je bolje, če kriterij ML nadomestimo s kriterijem *največje vrednosti posteriorne porazdelitve* (angl. maximum a posteriori, MAP). Uporaba kriterija MAP povzroči, da bodo ocenjene vrednosti parametrov odvisne od količine podatkov. Vrednosti bodo namreč določene kot utežena vsota ocene parametrov modela UBM in ocene, ki bi jo dobili po kriteriju ML. Več kot bomo imeli podatkov, bolj bo ocena MAP določena z oceno ML in manj z vrednostjo parametrov modela UBM

[1]Model GMM lahko vidimo kot poseben primer modela HMM, ki ima zgolj eno stanje.

[2]Čeprav lahko govorni signal preslikamo v supervektor tudi na druge načine, se bomo v tem prispevku omejili na preslikavo, ki temelji na modelu GMM.

ter *vice versa*, manj kot bo podatkov, manj bomo zaupali oceni ML, zato bo ocena MAP v večji meri določena z vrednostjo parametrov modela UBM[3].

Model UBM si lahko predstavljamo kot fotografski aparat, s katerim "posnamemo" sliko govorca — supervektor. Ta slika bo zaradi končne dolžine govornega signala nepopolna — krajši kot bo posnetek, bolj bo zamegljena, daljši kot bo posnetek, bolj bo ostra. Prav tako bo slika odvisna od izgovorjenega besedila. Ta analogija s fotografskim aparatom nazorno pokaže na pomembno razliko med razpoznavanjem govorcev in razpoznavanjem obrazov (in drugimi postopki, ki temeljijo na (statični) slikovni informaciji), kjer je za uspešno razpoznavanje ponavadi dovolj, če zajamemo sliko v enem samem časovnem trenutku. Razliko med dinamično in statično informacijo lahko še nazorneje ponazorimo z naslednjim miselnim eksperimentom. Denimo, da imamo fotografski aparat z zelo omejenim zornim kotom, ki se v vsakem trenutku usmeri na naključno mesto in posname majhno sličico. Če to snemanje ponavljamo nekaj časa, dobimo niz sličic, iz katerih želimo sestaviti sliko obraza. Pri tem se moramo zavedati, da zaradi končnega časa snemanja najverjetneje nismo zajeli dovolj sličic, da bi lahko sestavili sliko celotnega obraza in da za posamezno sličico ne vemo a priori, iz katerega dela obraza (ali celo ozadja) izhaja oz. povedano drugače, sličice niso opremljene s podatkom, na katero mesto je bil usmerjen aparat, ko je zajel sličico. Opisani miselni eksperiment je zelo podoben realni situaciji, s katero se soočamo pri besedilno neodvisnemu razpoznavanju govorcev.

3. Model sejne spremenljivosti

Žal govorski supervektor, ki ga ocenimo s pomočjo kriterija MAP, ne vsebuje le (za nas koristne) informacije o identiteti govorca, ampak vsebuje tudi informacijo o akustičnem kanalu, t.j. lastnostih mikrofona in prenosnih poti, ki smo jih uporabili pri zajemu govornega signala. Spremenljivost, ki ima negativen vpliv na uspešnost razpoznavanja, označimo z izrazom *sejna spremenljivost*. Vzroke za sejno spremenljivost ponavadi pripišemo različnim lastnostim akustičnega kanala, čeprav so za sejno spremenljivost pogosto odgovorni tudi manj oprijemljivi dejavniki kot so psihofizično stanje govorca, staranje, izgovorjeno besedilo ter celo dolžina govornega posnetka.

Proti sejni spremenljivosti se lahko borimo na različne načine. Postopke normalizacije sejne spremenljivosti ločimo v tri skupine: (i) postopki normalizacije na nivoju signala oz. značilk, (ii) postopki normalizacije na nivoju modela in (iii) postopki normalizacije na nivoju rezultatov razpoznavanja. Potencialno najbolj uspešni so tisti postopki, ki spadajo v drugo skupino, zato je največ raziskovalnega napora usmerjenega ravno v razvoj in izboljšavo postopkov te vrste.

Do sedaj je bilo predlaganih kar nekaj metod normalizacije sejne spremenljivosti na nivoju modela. Med njimi sta se najbolj uveljavili dve metodi, ki temeljita sicer na enaki predpostavki, vendar se problema lotevata na različne načine. Prva metoda je *projekcija motečih lastnosti* (angl. nuisance attribute projection, NAP) (Solomonoff et al., 2005), druga pa metoda *analiza vezanih faktorjev* (joint factor analysis, JFA) (Kenny et al., 2007), na katero se bomo osredotočili v tem prispevku.

Osnovna predpostavka, na kateri temelji metoda JFA, je, da lahko supervektor $\boldsymbol{m}$ razstavimo na vsoto dveh supervektorjev, govorskega supervektorja $\boldsymbol{s}$ in kanalskega supervektorja $\boldsymbol{c}$:

$$\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{c}, \tag{2}$$

pri čemer privzamemo, da sta $\boldsymbol{s}$ in $\boldsymbol{c}$ statistično neodvisna in normalno porazdeljena. Izkaže se, da je ločitev na govorsko in kanalsko komponento (2) izvedljiva le, če je kanal omejen na nižjerazsežen podprostor. Ta zahteva je smiselna, saj bi v nasprotnem primeru bilo mogoče, da bi kanal preslikal enega govorca v drugega. To bi pomenilo, da je razpoznavanje govorcev nerešljiv problem. Omejenost kanala na nizkorazsežen podprostor matematično izrazimo z zapisom:

$$\boldsymbol{c} = \boldsymbol{U}\boldsymbol{x}, \tag{3}$$

kjer je $\boldsymbol{x}$ nizkorazsežen standardno-normalno porazdeljen vektor, matrika $\boldsymbol{U}$ pa matrika lastnih kanalov (angl. eigenchannel). S podobnim izrazom omejimo tudi strukturo kovariančne matrike govorskega supervektorja[4]:

$$\boldsymbol{s} = \boldsymbol{m}_0 + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\boldsymbol{z}, \tag{4}$$

kjer nastopajo: središčni supervektor $\boldsymbol{m}_0$, nizkorazsežen standardno-normalno porazdeljen vektor $\boldsymbol{y}$, matrika lastnih glasov (engl. eigenvoice) $\boldsymbol{V}$, standardno-normalno porazdeljen vektor $\boldsymbol{z}$ in diagonalna matrika $\boldsymbol{D}$.

Če želimo uporabiti JFA kot model sejne spremenljivosti, moramo imeti na voljo obsežno govorno zbirko z velikim številom različnih govorcev, posnetih v čim večjem številu različnih sej. Na podlagi zadostne (Baum-Welch) statistike, ki jo s pomočjo modela UBM izluščimo iz govornega posnetka, moramo znati (i) poiskati posteriorne porazdelitve *prikritih spremenljivk*[5] x, y in z ter (ii) oceniti vrednosti *hiperparametrov*[6] $\boldsymbol{m}_0$, $\boldsymbol{U}$, $\boldsymbol{V}$ in $\boldsymbol{D}$. Postopki, ki jih pri tem potrebujemo, so detaljno predstavljeni v (Kenny, 2005). Opozorimo le, da je izvedba teh postopkov v splošnem precej zahtevna, prav tako pa so postopki zahtevni tudi s stališča prostorske in časovne kompleksnosti. Na srečo lahko v praksi stvari poenostavimo tako, da se zatečemo k aproksimativnim rešitvam.

[3] Tukaj smo opisali izkustveno interpretacijo ocenjevanja parametrov po kriteriju MAP. Zdi se, da se le-ta ne razlikuje bistveno od kriterija ML. Vendar bolj poglobljena obravnava pokaže, da med njima obstaja pomembna konceptualna razlika, saj slednji obravnava parametre verjetnostnega modela kot sicer neznane, a konstantne vrednosti, medtem ko prvi parametre obravnava kot naključne spremenljivke. Ta sila pomembna razlika v obravnavi parametrov je v največji meri odgovorna, da se statistiki delijo na dva tabora — frekvencioniste in Bayesovce.

[4] Enačba (4) nastopa v najbolj splošni obliki modela JFA. Če v tej enačbi izpustimo člen $\boldsymbol{V}\boldsymbol{y}$, dobimo klasični MAP, če pa izpustimo člen $\boldsymbol{D}\boldsymbol{z}$, dobimo EMAP (iz angl. eigenvoice MAP).

[5] Prikrite spremenljivke so tiste naključne spremenljivke verjetnostnega modela, ki jih lahko ocenimo le posredno — preko *opaženih* spremenljivk, katerih vrednosti nastopajo kot podatki.

[6] V modelu JFA poleg naštetih hiperparametrov nastopa tudi matrika $\boldsymbol{\Sigma}$, ki je prevzeta iz modela UBM.

4. Merjenje podobnosti

Denimo, da imamo dva govorna posnetka. Osnovno vprašanje, na katerega moramo pri razpoznavanju govorcev znati odgovoriti, je, ali je oba posnetka izgovoril isti govorec ali pa gre za dva različna govorca. Če želimo odgovoriti na to vprašanje, moramo na podlagi govornih posnetkov znati izmeriti podobnost med govorci. Podobnost lahko merimo na različne načine. Kadar obravnavamo problem razpoznavanja v okviru teorije verjetnosti, je najnaravneje, če podobnost merimo v obliki *razmerja verjetij* (angl. likelihood ratio).

4.1. Razmerje verjetij

Merjenje podobnosti dveh posnetkov (oz. govorcev) poteka tako, da enega izmed posnetkov (po navadi daljšega, označimo ga z $\mathcal{D}_1$) uporabimo za oceno parametrov modela GMM oz. za določitev govorskega supervektorja, drug posnetek (označimo ga z $\mathcal{D}_2$) pa prilegamo na ta model tako, da izračunamo vrednost funkcije verjetja pri ocenjeni vrednosti parametrov. Vrednost verjetja je potrebno še normirati, kar storimo tako, da izračunamo še verjetje pri vrednostih parametrov od govorca neodvisnega modela UBM. Razmerje verjetij izračunamo tako, da obe izračunani vrednosti med seboj delimo.

Govorski supervektor izračunamo tako, da poiščemo največjo vrednost posteriorne porazdelitve prikritih spremenljivk, kar zapišemo kot:

$$\boldsymbol{s} = \arg\max_{\boldsymbol{s}} P(\mathcal{D}_1|\boldsymbol{s},\boldsymbol{c})P(\boldsymbol{x})P(\boldsymbol{y})P(\boldsymbol{z}). \qquad (5)$$

Ker model JFA — za razliko od klasičnega postopka MAP — eksplicitno izpostavi apriorne porazdelitve prikritih spremenljivk ($\boldsymbol{x}$, $\boldsymbol{y}$ in $\boldsymbol{z}$), lahko kanal posnetka, ki ga prilegamo na model, izločimo s seštevanjem preko vseh možnih vrednosti parametrov, tako da upoštevamo porazdelitev spremenljivke $\boldsymbol{x}$:

$$P(\mathcal{D}_2|\boldsymbol{s}) = \int P(\mathcal{D}_2|\boldsymbol{s},\boldsymbol{c})P(\boldsymbol{x})d\boldsymbol{x}. \qquad (6)$$

4.2. Metoda podpornih vektorjev

Poleg razmerja verjetij se je na področju razpoznavanja govorcev kot način merjenja podobnosti uveljavil tudi postopek, ki temelji na diskriminatorni metodi podpornih vektorjev (angl. support vector machine, SVM). Pri tem načinu merjenja podobnosti najprej oba govorna posnetka preslikamo v prostor supervektorjev. Prvi govorski supervektor nato uporabimo za oceno modela SVM, ki ga predstavimo s hiperravnino, ki ločuje govorski supervektor od vnaprej izbrane množice neodvisnih supervektorjev. Podobnost med obema supervektorjema izračunamo tako, da izračunamo oddaljenost drugega supervektorja do te hiperravnine.

Čeprav metoda SVM s pomočjo *jedrnega trika* zmore poiskati tudi nelinearno ločilno mejo med dvema razredoma, se izkaže, da zaradi velike razsežnosti prostora supervektorjev nelinearnost v primeru razpoznavanja govorcev ni potrebna.

5. Eksperimenti

5.1. Akustične značilke

Surove govorne signale smo najprej obdelali tako, da smo iz njih izluščili značilke melodičnega kepstra. To smo storili tako, da smo vsakih 10 ms iz 25 ms dolgega okna izračunali 20 koeficientov MFCC, ki smo jim dodali še dinamične značilke prvega in drugega reda. Skupna dolžina vektorja značilk je bila tako 60. Ponavadi se značilke normirajo z eno izmed metod normalizacije, kot je npr. ukrivljanje značilk (Pelecanos and Sridharan, 2001), mi pa smo ta korak izpustili, saj smo želeli preveriti, ali je mogoče normiranje značilk nadomestiti z metodo normalizacije na nivoju modela v obliki modela JFA. Iz dobljenega niza vektorjev značilk smo izločili tiste vektorje, ki so ustrezali negovornim delom signala. Pri tem smo se zanašali na časovno poravnane besedne prepise, ki so bili pridobljeni s samodejnim razpoznavalnikom govora.

5.2. Model UBM

Za učenje parametrov modela UBM smo uporabili NIST-ove govorne zbirke iz let 2004, 2005 in 2006. Po odstranjevanju negovornih odsekov nam je za ocenjevanje parametrov modela UBM ostalo približno 600 ur govornega gradiva. Učenja smo se lotili tako, da smo najprej ocenili le eno Gaussovo porazdelitev in nato s postopno delitvijo povečevali število porazdelitev vse do številke 2048. Čeprav je običajna praksa, da obravnavamo ženske in moške govorce ločeno in da torej zgradimo dva neodvisna modela UBM, smo se sami odločili, da takšne ločitve ne izvedemo in za oba spola zgradimo le en model UBM.

5.3. Model JFA

Za učenje hiperparametrov modela JFA smo uporabili iste podatke, s katerimi smo učili že model UBM. Podatke smo razdelili na dve množici; večjo množico smo uporabili za učenje matrik $\boldsymbol{U}$ in $\boldsymbol{V}$ ter vektorja $\boldsymbol{m}_0$, manjšo pa za oceno matrik $\boldsymbol{D}$ in $\boldsymbol{\Sigma}$. Pri tem smo se zgledovali po postopku, ki so ga predlagali Kenny et al. (2008).

5.4. Merjenje podobnosti

Podobnost smo merili na dva načina, ki smo ju že predstavili v razdelkih 4.1. in 4.2. Ker je metoda SVM občutljiva na različen razpon vrednosti posameznih koeficientov vektorjev (atributov), je atribute potrebno predhodno ustrezno normirati. Običajno se vrednosti skalira na interval med 0 in 1, sami pa smo uporabili neparametrično metodo normalizacije ranga, za katero se je izkazalo, da dodatno pripomore k izboljšanju rezultatov (Stolcke et al., 2008).

V nasprotju z uveljavljeno prakso rezultatov razpoznavanja nismo normirali, saj se je izkazalo, da z normalizacijo le poslabšamo učinkovitost razpoznavanja.

Znano je, da lahko z združevanjem rezultatov razpoznavanja — če le-ti vsebujejo dovolj komplementarne informacije — izboljšamo učinkovitost razpoznavanja, zato smo poskusili rezultate, ki smo jih dobili s kriterijema razmerja verjetij in metodo podpornih vektorjev združiti tako, da smo posamezne rezultate sešteli. Izkazalo se je, da se na tak način uspešnost razpoznavanja še precej izboljša (glej tretjo vrstico v tabeli 1).

5.5. Rezultati

Sistem smo preizkusili na uveljavljeni zbirki NIST SRE 2008[7], ki je namenjena vrednotenju sistemov za verifikacijo govorcev. Osredotočili smo se le na tisti del glavne naloge, ki se nanaša na telefonske posnetke. Skupaj smo obdelali 37001 poskusov, od katerih je bilo 3826 takšnih, v katerih je govorec v obeh posnetkih isti, ostalih 33175 pa takšnih, v katerih gre za dva različna govorca.

Dobljeni rezultati so v skladu s standardnim načinom predstavljanja rezultatov biometričnih postopkov podani v tabeli 1 in na sliki 1. Vidimo lahko, da so rezultati primerljivi s tistimi, ki so jih dosegli najuspešnejši sistemi v okviru NIST-ovega vrednotenja sistemov za razpoznavanje govorcev iz leta 2008, kljub temu, da smo v nasprotju z uveljavljeno prakso uporabili spolno-neodvisen sistem UBM in da nismo izvedli normalizacije značilk in rezultatov razpoznavanja.

6. Zaključek

V prispevku smo podali pregled postopkov razpoznavanja govorcev, ki temeljijo na ideji, da lahko vsakega govorca predstavimo v obliki supervektorja. Opisali smo tudi model sejne spremenljivosti (JFA), s pomočjo katerega lahko sejni supervektor razstavimo na vsoto govorskega in kanalskega supervektorja. Model JFA je poleg praktične uporabnosti zanimiv tudi zaradi tega, ker predstavlja enega izmed prvih uspešnih poskusov uporabe Bayesovskih metod na področju razpoznavanja govorcev, kakor tudi na širšem področju govornih tehnologij. Bayesovske metode so bile namreč zaradi računske zahtevnosti do nedavnega omejene na manj kompleksne probleme, z večanjem računske moči računalnikov in z napredkom na področju aproksimativnih Bayesovskih metod pa prihajajo vse bolj v ospredje (glej npr. (Kenny, 2010)) tudi pri kompleksnejših (angl. large-scale) problemih, kakršen je tudi problem samodejnega razpoznavanja govorcev.

Tabela 1: Rezultati v obliki vrednosti EER in DCF, ki smo jih dobili z uporabo različnih mer podobnosti: razmerja verjetij (LR), metode podpornih vektorjev (SVM) in njuno kombinacijo (LR + SVM).

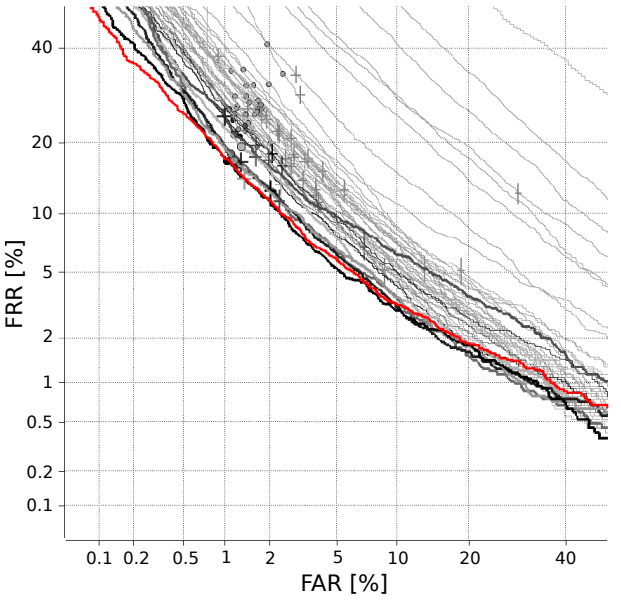
Mera podobnosti	EER [%]	DCF
LR	8,52	0,36
SVM	6,54	0,30
LR + SVM	5,50	0,27

7. Literatura

A. P. Dempster, N. M. Laird, in D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

P. Kenny, G. Boulianne, P. Ouellet, in P. Dumouchel. 2007. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447.

[7]Vse podatke v zvezi s to zbirko in pripadajočimi eksperimentalnimi protokoli je mogoče najti v dokumentu, ki je dostopen na naslovu http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan/release4.pdf.



Slika 1: Primerjava doseženih rezultatov z uradnimi rezultati NIST SRE 2008 (označeni s sivo barvo). Z rdečo barvo je vrisana DET krivulja sistema, pri katerem smo združili rezultate dveh mer podobnosti (LR + SVM).

P. Kenny, P. Ouellet, N. Dehak, V. Gupta, in P. Dumouchel. 2008. A study of inter-speaker variability in speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 16(5):980–988.

P. Kenny. 2005. Joint factor analysis of speaker and session variability: Theory and algorithms. Tehnično poročilo 06/08-13, Centre de recherche informatique de Montréal (CRIM). [Online] http://www.crim.ca/perso/patrick.kenny/.

P. Kenny. 2010. Bayesian speaker verification with heavy-tailed priors. V: *Proc. Odyssey*, Brno, Czech Republic. [Online] http://www.crim.ca/perso/patrick.kenny/.

J. Pelecanos in S. Sridharan. 2001. Feature warping for robust speaker verification. V: *Proc. Odyssey*, str. 213–218, Crete, Greece.

D. A. Reynolds, T. F. Quatieri, in R. B Dunn. 2000. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1–3):19–41.

A. Solomonoff, W. Campbell, in I. Boardman. 2005. Advances in channel compensation for SVM speaker recognition. V: *Proc. ICASSP*, str. 629–632, Philadelphia.

A. Stolcke, S. Kajarekar, in L. Ferrer. 2008. Nonparametric feature normalization for SVM-based speaker verification. V: *Proc. IEEE ICASSP*, str. 1577–1580.

Zahvala

Razvojno raziskovalno delo je delno financirala Evropska unija (št. pogodbe 164/2009), in sicer iz sredstev Evropskega socialnega sklada v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013, razvojne prioritete Spodbujanje podjetništva in prilagodljivosti, prednostne usmeritve Strokovnjaki in raziskovalci za konkurenčnost podjetij, in sicer iz proračunske postavke 6882 Usposabljanje strokovnjakov in raziskovalcev 07-13-EU v višini 85 % vrednosti sofinanciranja in iz proračunske postavke 6966 Usposabljanje strokovnjakov in raziskovalcev 07-13-slovenska udeležba v višini 15 % vrednosti sofinanciranja.

Slovenska baza izgovarjav z Lombardovim efektom - SiLSD

Damjan Vlaj, Aleksandra Zögling Markuš, Marko Kos, Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor, Slovenija
{damjan.vlaj, sandra.zogling, marko.kos, kacic}@uni-mb.si

Povzetek

Članek predstavlja korake pri pridobivanju govornega materiala in postopke označevanja Slovenske baze izgovarjav z Lombardovim efektom (SiLSD – Slovenian Lombard Speech Database), katere snemanje se je začelo v letu 2008. SiLSD[1] je bila posneta na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Namen tega članka je opisati strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, opis poteka snemanja, predstavitev strukture baze izgovarjav in predstavitev orodja, ki je bilo uporabljeno za obdelavo baze izgovarjav. Baza izgovarjav zajema posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk. Vsak govorec je posnel besedila osmih različnih korpusov, in to v dveh snemalnih sejah v razmiku vsaj enega tedna. Struktura korpusa je podobna strukturi, ki je bila uporabljena v bazi izgovarjav SpeechDat II. Posneto je bilo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo. Govorni material je bil ročno obdelan in označen z orodjem LombardSpeechLabel, razvitim na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Govorni material in pripadajoč označevalni material je shranjen na 10 DVD medijih (en govorec na enem DVD mediju).

Abstract

This paper presents the steps for acquisition of speech material and steps for annotation of Slovenian Lombard Speech Database (SiLSD), the recording of which started in the year 2008. The database was recorded at the Faculty of Electrical Engineering and Computer Science, University of Maribor. The goal of this paper is to describe the hardware platform used for the acquisition of speech material, description of recording scenarios, presentation of database structure and tools used for the annotation of SiLSD. The database consists of recordings of 10 Slovenian native speakers. Five males and five females were recorded. Each speaker pronounced a set of eight corpuses in two recording sessions with at least one week pause between recordings. The structure of the corpus is similar to the SpeechDat II database. Approximately 30 minutes of speech material per speaker and per session was recorded. The manual annotation of speech material was performed with the LombardSpeechLabel tool developed at the Faculty of Electrical Engineering and Computer Science, University of Maribor. The speech and annotation material was saved on 10 DVDs (one speaker on one DVD).

1. Uvod

Namen članka je predstaviti strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, potek snemanja, strukturo baze izgovarjav in orodje, ki je bilo uporabljeno za obdelavo Slovenske baze izgovarjav z Lombardovim efektom (SiLSD – Slovenian Lombard Speech Database). Baza izgovarjav je bila posneta z namenom, da bi v posnetkih govora zajeli Lombardov efekt.

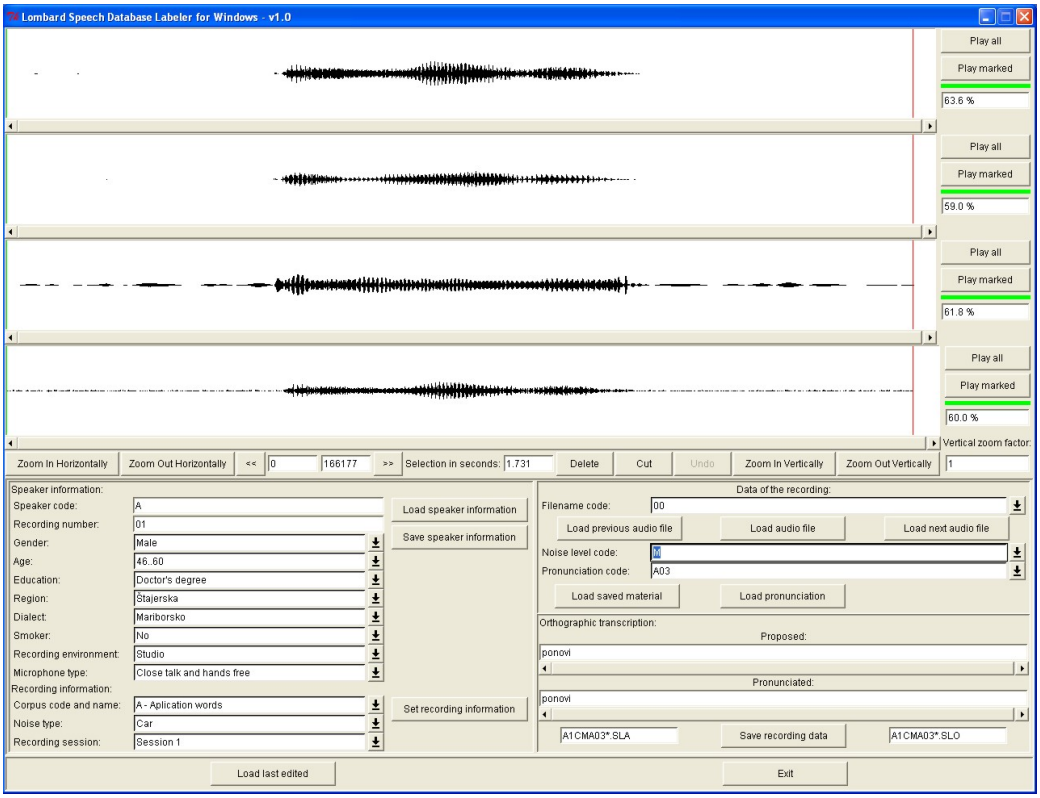
Lombardov efekt je bil prvič omenjen leta 1911, ko je Etienne Lombard (Lombard, 1911) odkril fiziološki efekt pri tvorjenju govora ob prisotnosti šuma. Lombardov efekt je pojav, pri katerem govorec poveča glasnost govora ob povečanju glasnosti šuma ozadja. Skozi zgodovino sta se pojavili dve interpretaciji Lombardovega efekta. Prva dokazuje, da je Lombardov efekt fiziološki avdio-fonetični refleks (Lombard, 1911). Druga razlaga pa temelji na domnevi, da so spremembe v govoru, ki jih povzroči Lombardov efekt poledica slabše razumljivosti običajnega govora, ki ga govorec posluša v šumnem okolju (Lane in Tranel, 1971). Nekateri avtorji so tudi trdili, da lahko oba mehanizma prispevata k spremembam govornih karakteristik, kadar je govorec v šumnem okolju (Junqua, 1993) in tako povečata razumljivost govora. Ker povečanje glasnosti govora privede do sprememb karakteristik govornega signala, ima to pri avtomatskem razpoznavanju govora največkrat za posledico slabšo uspešnost avtomatskega razpoznavanja govora.

V procesu avtomatskega razpoznavanja govora ima postopek izločanja značilk velik pomen in zelo vpliva na uspešnost sistemov avtomatskega razpoznavanja govora. Z uporabo standardnih postopkov izločanja značilk, pri tem imamo v mislih uporabo mel-kepstralnih značilk govornega signala in uporabo prikritih modelov Markova (Young in dr., 2000), se uspešnost avtomatskega razpoznavanja govora v studijskih razmerah približa 100 odstotkom (Hirsch in Pearce, 2000). Kakor hitro pa preidemo iz studijskega v naravno šumno okolje, uspešnost razpoznavanja upade glede na nivo razmerja med signalom in šumom (Hirsch in Pearce, 2000). Uspešnost avtomatskega razpoznavanja govora pa upade tudi zaradi sprememb karakteristik govornega signala, ki so se zgodile zaradi vpliva Lombardovega efekta na govorni signal. Vzrok za to so postopki izločanja značilk in učenja prikritih modelov Markova, ki so izvedeni na bazah izgovarjav, ki niso posnete v okoljih, kjer je bil v govornem signalu prisoten Lombardov efekt.

Raziskave na področju Lombardovega efekta kažejo, da se Lombardov govor razlikuje od normalnega govora na več načinov. Glavne spremembe značilnosti Lombardovega govora so lahko vidne v zvišanju osnovne harmonske frekvence, v povečanju časovnega trajanja vokalov in v premiku formantov F1 in F2. Avtorja Hanley in Steer (Hanley in Steer, 1949) sta tudi ugotovila, da se hitrost govora zniža, ko je govor tvorjen v šumnem okolju.

Baze izgovarjav posnete v realnem okolju zagotavljajo dragocen material za sisteme avtomatskega razpoznavanja govora, vendar je v primeru glasnejšega šumnega okolja

[1] Lastnik baze izgovarjav je podjetje SVOX, ki namerava ponuditi bazo izgovarjav preko organizacije ELDA/ELRA.



Slika 1: Programsko orodje LombardSpeechLabel za ročno obdelavo in označevaje govornega materiala.

(hrup v avtomobilu, govor v ozadju, …) zaradi pomešanosti govora s šumom ozadja analiza prisotnosti Lombardovega efekta zelo otežena (Bořil in dr., 2006). Da bi bilo mogoče potrditi prisotnost Lombardovega efekta v govornem signalu, je potrebno analizirati govorni signal, ki vsebuje čim manj šuma v ozadju govora. Zato so Bořil in dr. posneli bazo izgovarjav, v kateri so želeli poudariti vpliv Lombardovega efekta v govornem signalu (Bořil in dr., 2006).

Z namenom nadaljevati raziskave na tem področju smo zasnovali snemalno okolje in izvedli snemanje baze izgovarjav SiLSD. Namen baze je omogočiti raziskave vpliva Lombardovega efekta v primeru različnih šumov okolja, različnih šumnih nivojev, ki jih v času snemanja sliši govorec in analizirati konsistentnost Lombardovega efekta glede na čas snemanja. Uporabili smo dva različna tipa šumov in izvedli snemanje pri dveh različnih nivojih šuma ozadja ter v dveh snemalnih sejah. S takšno bazo smo želeli omogočiti analize, ki bi potrdile, da povečanje nivoja šuma ozadja poveča vpliv Lombardovega efekta. Želeli smo tudi analizirati konsistentnost Lombardovega efekta znotraj dveh snemalnih sej, saj je med snemanjem posameznih sej s posameznim govorcem bilo vsaj en teden razmika.

V nadaljevanju tega članka, to je v drugem poglavju, bomo predstavili način pridobivanja govornega materiala, in opisali potek snemanja. V tretjem poglavju bomo predstavili način obdelave govornega materiala. V četrtem poglavju bomo predstavili strukturo baze izgovarjav SiLSD. V petem poglavju bomo podali zaključke.

2. Pridobivanje govornega materiala

Baza izgovarjav SiLSD je bila posneta v studijskem okolju. Vsak govorec je izgovoril nabor osmih korpusov besedil v dveh snemalnih sejah z vsaj enim tednom razmika med snemalno sejo za posameznega govorca. Posneli smo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo.

Pri snemanju so bili hkrati uporabljeni prostoročni mikrofon AKG C 3000 B, obustni mikrofon Shure Beta 53 in dvokanalni laringograf EG2. Hkrati smo izvedli snemanje štirih kanalov:

- prostoročni mikrofon,
- obustni mikrofon,
- laringograf in
- snemanje šuma ozadja, pomešanega z govorom govorca, ki je bil predvajan na slušalke govorca med samim snemanjem.

Snemalna platforma je bila sestavljena iz zunanje zvočne kartice Audigy 4 PRO za štirikanalno zajemanje zvoka in mešalne mize Phonic MU244X. Zajemanje je bilo izvedeno pri frekvenci vzorčenja 96 kHz in 24-bitni linearni kvantizaciji.

Pri snemanju sta bila uporabljena dva tipa šumov in sicer hrup v avtomobilu in govor v ozadju. Šuma sta bila vzeta iz baze izgovarjav Aurora 2 (Hirsch in Pearce, 2000). Šuma sta bila normalizirana in predvajana na govorčeve slušalke AKG K271.

Nivo predvajanega šuma je bil pred začetkom vsakega snemanja nastavljen na način, kot je bilo predlagano v (Bořil in dr., 2006). Zahtevan nivo šuma je bil nastavljen glede na efektivno napetost zvočne kartice pri odprtih sponkah. Za dosego Lombardovega efekta smo izbrali nivo šuma 80 dB SPL[2] in 95 dB SPL pri navidezni razdalji med 1 in 3 metri.

[2] SPL je okrajšava za Sound Pressure Level (raven zvočnega tlaka).

Znotraj ene snemalne seje so bila izvedena tri snemanja:

- referenčno snemanje brez prisotnega šuma,
- snemanje pri 80 dB SPL in
- snemanje pri 95 dB SPL.

Med snemanjem posameznih besedilnih korpusov (besede, števila, povezane številke, stavki, …) je bil narejen kratek premor, da se je lahko govorec prilagodil na okolje brez šuma. Daljši premor je bil narejen po snemanju vseh osmih besedilnih korpusov.

Med samim snemanjem je bilo vzpostavljeno sodelovanje med govorcem in snemalcem. Snemalec je slišal govor govorca z dodanim šumom. Pri tem je snemalec ocenil, ali je govor, ki ga sliši, razumljiv ali ne. Snemalec je preko LCD monitorja vzpodbujal govorca k bolj glasnemu govorjenju in ponovitvi izrečenega, če mu govor, ki ga je poslušal, ni bil dovolj razumljiv.

Baza izgovarjav zajema tako čiste posnetke brez dodanega šuma, kot tudi posnetke s šumom, ki so bili predvajani na slušalke govorca med snemanjem.

3. Obdelava govornega materiala

Govorni material je bil ročno pregledan, obdelan in označen s pomočjo programskega orodja Lombard-SpeechLabel (slika 1), ki je bilo razvito na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Programsko orodje je napisano v programskem jeziku Tcl/Tk/Tix, ki omogoča vizualno programiranje. Čeprav je bilo razvito na platformi operacijskega sistema Microsoft Windows, ga lahko brez večjih težav z manjšimi spremembami prenesemo na druge platforme operacijskih sistemov.

Okolje programskega orodja LombardSpeechLabel je razdeljeno v tri polja. Zgornje polje vsebuje prikaz štirih časovnih potekov signalov (prostoročni mikrofon, obustni mikrofon, laringograf in signal, predvajan na slušalke govorca), ki so bili zajeti med samim snemanjem baze izgovarjav. Vse posnetke je možno predvajati oz. poslušati s pritiskom na gumbe, ki se nahajajo na desni strani zgornjega polja. Spodnji del orodja je razdeljen na dva dela. Na levi strani se nahaja informacija o govorcu in snemanju. Na desni strani pa so podani dodatni podatki o snemanju in sama ortografska transkripcija posnetega signala.

Programsko orodje LombardSpeechLabel izvede tudi shrambo avdio materiala v končni format, tako da je signal vzorčen s frekvenco 96 kHz in kvantiziran s 16-bitno linearno kvantizacijo.

4. Struktura baze izgovarjav

Baza izgovarjav SiLSD vsebuje posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk. Kot smo že omenili, je vsak govorec izgovoril nabor osmih korpusov besedil v dveh snemalnih sejah z vsaj enim tednom razmika med snemalno sejo za posameznega govorca. Struktura korpusa je podobna strukturi, ki je bila uporabljena v bazi izgovarjav SpeechDat II (Kaiser in Kačič, 1997). Več informacij o sami bazi izgovarjav bomo podali v naslednjih podpoglavjih.

4.1. Format avdio in opisne datoteke

Govorni signal je zapisan v avdio datotekah kot zaporedje otipkov kvantiziranih s 16-bitno linearno kvantizacijo pri frekvenci vzorčenja 96 kHz. Otipki so shranjeni v surovem Intel formatu, brez kakršnekoli glave datoteke. Vsaka izgovarjava je shranjena v svoji avdio datoteki. Velikost datotek se razlikuje glede na besedilni korpus. Vsaka avdio datoteka ima pripadajočo opisno datoteko v opisnem formatu SAM s kodiranjem simbolov UTF-8.

A	Koda govorca (A-Z)
S	Koda seje (1-9) – uporabljeni samo 1 in 2
T	Koda tipa šuma: • R: brez šuma • C: hrup v avtomobilu • B: govor v ozadju
R	Koda snemanja: • N: snemanje referenčnega signala brez prisotnosti šuma • L: snemanje signala brez prisotnosti šuma • M: snemanje signala s prisotnostjo šuma pri nivoju šuma 80 dB SPL • H: snemanje signala s prisotnostjo šuma pri nivoju šuma 95 dB SPL
NNN	Koda korpusa besedil (A00 – Z99): A – aplikacijske besede B – povezane številke D – datumi I – izolirane številke N – naravna števila S – fonetično bogati stavki T – časi W – fonetično bogate besede
C	Koda snemalnega kanala: • 1: prostoročni mikrofon • 2: obustni mikrofon • 3: signal laringografa • 4: signal, ki je bil predvajan na slušalke govorca med snemanjem
LL	Dve črki jezikovne kode ISO 639
F	Koda tipa datoteke: O – datoteka z ortografsko označevalno vsebino A – datoteka z avdio vsebino

Tabela 1: Opis datotečne nomenklature.

4.2. Datotečna nomenklatura

Imena datotek sledijo datotečnemu zapisu ISO 9660 (8 + 3 znaki) glede na osnovni standard zgoščenk. Zaradi velike količine avdio materiala so bili vsi podatki shranjeni na DVD mediju.

Za datotečno nomenklaturo je bila uporabljena naslednja predloga:

```
A S T R NNN C. LL F
```

Datotečna nomenklatura je podrobno opisana v Tabeli 1.

4.3. Struktura map

Struktura map je sestavljena iz petih nivojev in je podana na naslednji način:

```
\<database>
    \<speaker>
        \<session>
            \<condition>
                \<corpus>
```

<database>	Določena kot: <name><language code> i.e. LOMBSPSL kjer je: <name> LOMBSP in predstavlja Lombard Speech <LL> ISO 2-črki kode SL za Slovenian
<speaker>	Določen kot: SPK_<a> kjer <a> predstavlja naraščajočo črko abecede od A do Z. Ta črka je enaka prvi črki v imenu datoteke (glej poglavje 4.2).
<session>	Določena kot: SES_<s> kjer <s> predstavlja naraščajočo številko od 1 do 9. Ta številka je enaka drugi številki v imenu datoteke (glej poglavje 4.2).
<condition>	Določeni so trije tipi okoliščin: • REF: snemanje referenčnega signala brez prisotnosti šuma, • CAR: snemanje signala s prisotnostjo šuma hrupa v avtomobilu in • BABBLE: snemanje signala s prisotnostjo šuma govora v ozadju.
<corpus>	Določen kot: CORPUS_<c> kjer <c> predstavlja črko korpusa besedil: A – aplikacijske besede B – povezane številke D – datumi I – izolirane številke N – naravna števila S – fonetično bogati stavki T – časi W – fonetično bogate besede

Tabela 2: Struktura map za bazo izgovarjav SiLSD.

Struktura map je postavljena tako, da so posnetki vsakega govorca shranjeni na svojem DVD mediju. Vsak govorec je posnel svoj del baze izgovarjav v dveh sejah. V vsaki seji se nahajajo referenčni posnetki in posnetki, posneti pri dveh šumih okoljih, ki jih je slišal govorec. Vsako okolje vsebuje osem govornih korpusov.

Struktura map za bazo izgovarjav SiLSD je podrobneje podana v Tabeli 2.

4.4. Definicija kod korpusov besedil

V bazi izgovarjav smo definirali za vsak korpus besedil eno črko, ki določa korpus besedil, in dve številki, ki določata zaporedno številko posnetka v samem korpusu. Takšno označevanje je vneseno tudi v imena datotek, da bi uporabnik enostavno razbral iz imena datoteke, v kateri korpus besedil sodi njena vsebina. Definicija kod korpusov besedil je podana v Tabeli 3.

Predvsem pri izbiri fonetično bogatih besed in stavkov smo stremeli k čim bolj enakomerni pokritosti fonemov.

5. Zaključek

V članku smo opisali strojno platformo, ki je bila uporabljena za pridobivanje govornega materiala, podali smo potek snemanja, predstavili strukturo baze izgovarjav in orodje, ki smo ga uporabili za obdelavo baze izgovarjav SiLSD. Baza izgovarjav zajema posnetke desetih slovenskih govorcev, od tega petih moških in petih žensk.

Črka korpusa	Zaporedna številka	Vsebina korpusa besedil
A	00-29	aplikacijske besede (30 besed)
B	00-04	povezane številke (sekvenca 10 številk izgovorjena 5 krat)
D	00-04	datumi (5 datumov)
I	00-11	izolirane številke (12 številk)
N	00-04	naravna števila (5 števil)
S	00-29	fonetično bogati stavki (30 stavkov)
T	00-06	časi (7 časov)
W	00-49	fonetično bogate besede (50 besed)

Tabela 3: Definicija kod korpusov besedil.

Vsak govorec je posnel besedila osmih različnih korpusov in to v dveh snemalnih sejah v razmiku vsaj enega tedna med snemanji. Posneto je bilo približno 30 minut govornega materiala na enega govorca in na eno snemalno sejo.

Že med samim snemanjem so bile narejene nekatere analize govornega materiala, s katerimi smo potrdili prisotnost Lombardovega efekta v posnetem govornem signalu. Te analize so bile izvedene z opazovanjem sprememb osnovne harmonske frekvence, časovnega trajanja vokalov in s premiki formantov F1 in F2. V prihodnosti želimo izvesti poglobljeno analizo posnetega govornega materiala baze predvsem v smeri ugotavljanja vpliva različnih nivojev šuma ozadja na jakost Lombardovega efekta, odvisnost od vrste šuma ozadja in oceniti konsistentnost prisotnosti Lombardovega efekta v govornem signalu različnih govorcev.

6. Literatura

Lombard, E. (1911). Le signe de l'elevation de la voix, *Annals maladiers oreille, Larynx, Nez, Pharynx*, 37: 101-119.

Lane, H. in Tranel, B. (1971). The Lombard sign and the role of hearing in speech, *Journal of Speech and Hearing Research*, 14(4): 677-709.

Junqua, J. C. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers, *Journal of the Acoustical Society of America*, W(l): 510-524.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. in Woodland, P. (2000). *The HTK Book - Version 3.0*, Microsoft Corporation, ZDA.

Hirsch, H. G. in Pearce, D. (2000). The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *ISCA ITRW ASR'00 Proceedings*. Pariz, Francija.

Hanley, T. in Steer, M. (1949). Effect of level of distracting noise upon speaking rate, duration and intensity, *Journal of Speech and Hearing Disorders*, 14(4): 363-368.

Bořil, H., Bořil, T. in Pollák, P. (2006). Methodology of Lombard speech database acquisition: Experiences with CLSD, *Proceedings of the fifth Conference on Language Resources and Evaluation – LREC'06*, 1644-1647.

Kaiser, J. in Kačič, Z. (1997). *SpeechDat Slovenian Database for the Fixed Telephone Network*, Univerza v Mariboru, Maribor, Slovenija.

Zmanjševanje odvečnosti končnih pretvornikov za učinkovito gradnjo razpoznavalnikov slovenskega govora z velikim besednjakom

Simon Dobrišek, France Mihelič

LUKS, Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 2, 1000 Ljubljana
{simon.dobrisek,france.mihelic}@fe.uni-lj.si

Povzetek

V članku je predstavljena zasnova razpoznavalnika slovenskega govora z velikim besednjakom, katerega gradnja temelji na izvirnem postopku učinkovitega prirastnega zmanjševanja odvečnosti končnih pretvornikov. Učinkovitost postopka je predstavljena z rezultati pri gradnji končnega pretvornika, ki kodira slovar izgovarjav s preko šeststo osemdeset tisoč različnih besed v različnih pregibnih oblikah. Predstavljeni postopek se je izkazal za znatno bolj učinkovitega od postopkov, ki so na razpolago v uveljavljeni zbirki orodij OpenFST.

Redundancy reduction of finite-state transducers for the efficient construction of large-vocabulary Slovenian speech recognizers

This paper presents the development of a large-vocabulary speech recognition system for the Slovenian language. Its construction is based on an efficient algorithm for the incremental redundancy reduction of finite-state transducers. The efficiency of the algorithm is demonstrated by using it to construct a finite-state pronunciation dictionary model that encodes over 680 thousand different word forms. The experimental results show that the incremental algorithm creates considerably smaller finite-state models than the conventional algorithms implemented in the OpenFST toolkit.

1. Uvod

Samodejni razpoznavalnik govora lahko obravnavamo kot informacijski dekodirnik, ki ga predstavimo kot sklop deterministič nih, nedeterministič nih in verjetnostnih končnih avtomatov. Enako velja za samodejni sintetizator oziroma kodirnik govora. Najbolj pregledna in enovita predstavitev tovrstnih sistemov je podana v obliki večnivojskega sklopa končnih pretvornikov (angl. finite-state transducers), kjer se posamezni nivoji nanašajo na slovnico, slovar, akustično-fonetična pravila izgovarjav besed, akustične modele glasov itd (Mohri et al., 2002). S takšno predstavitvijo se problem samodejnega razpoznavanja govora poenostavi na problem iskanja najbolj verjetnega zaporedja prehodov med stanji končnih pretvornikov pri danem vhodnem zaporedju akustičnih govornih vzorcev (Jelinek, 1997). Podobno velja za samodejne sintetizatorje govora, kjer je problem le obrnjen in se iz danega zaporedja besed določa najbolj verjetna zaporedja prehodov med stanji končnih pretvornikov ter s tem najbolj verjetne izhodne akustične govorne vzorce (Black et al., 2007).

Pri razpoznavalnikih govora z velikim besednjakom postane struktura sklopa končnih pretvornikov zelo obsežna, saj lahko vsebuje na desetine milijonov stanj in prehodov med stanji. Udejanjenje tako obsežnega sklopa končnih pretvornikov na današnjih običajnih računalnikih nujno zahteva njegovo optimizacijo glede potrebne količine pomnilnika in predvsem glede zahtevnosti postopka iskanja najbolj verjetnih zaporedij prehodov med stanji avtomatov. Zamisel uteženih končnih pretvornikov in računalniška programska orodja, ki jih je razvil Mohri s sodelavci (Mohri et al., 2008), ponujajo enega od možnih pristopov k takšni optimizaciji. Orodja so na razpolago v prosto-dostopni zbirki OpenFST (Allauzen et al., 2007), ki ponuja učinkovite izvedbe algoritmov za ustvarjanje, sklapljanje in optimizacijo posplošenih uteženih končnih pretvornikov.

Ena od naših dolgoročnih raziskovalnih usmeritev je razvoj čim bolj prilagodljivega in zanesljivega lastnega razpoznavalnika slovenskega govora z velikim besednjakom. Pri njegovem razvoju smo se doslej deloma naslanjali na obstoječa prosto dostopna računalniška programska orodja za gradnjo razpoznavalnikov govora, kot so orodja CUED HTK in CMU Sphinx. V zadnjem času pa razvoj preusmerjamo v smeri zamisli sklopa končnih pretvornikov, ki jih ta orodja le deloma podpirajo. Za ta namen so bolj primerna orodja OpenFST, ki pa zaenkrat ne vključujejo orodij za gradnjo akustičnih modelov glasov.

Naše raziskovalne izkušnje z navedenimi orodji so pokazale, da se ne moremo v celoti naslanjati le na zamisli, ki so se uveljavile pri razvoju razpoznavalnikov govora z velikim besednjakom za angleške govorjene jezike. Zapleti pri razvoju razpoznavalnika za slovenski govorjeni jezik se pojavljajo predvsem zaradi precej večjega števila pregibnih oblik besed in manj strogega besednega reda, kar v primerjavi z razpoznavalnik angleškega govora precej povečuje obsežnost strukture sklopa končnih pretvornikov. Po naši grobi oceni se zaradi velikega števila pregibnih oblik besed pri približno istem številu leksemov obsežnost razpoznavalnika slovenskega govora v primerjavi s primerljivim razpoznavalnikom angleškega govora vsaj podeseteri. Zato smo posebno pozornost posvetili prav optimizaciji končnih pretvornikov v smislu zmanjševanja njihove odvečnosti, pri čemer smo poskušali upoštevati posebnosti slovenskega jezika.

V zadnjem času smo v programskem jeziku Java povsem na novo zasnovali razpoznavalnik govora z velikim besednjakom, ki podpira vsa tri omenjena prosto-dostopna programska orodja, predvsem v smislu datotečnih formatov modelov končnih avtomatov ipd. Na ta način bo, denimo, mogoče zgraditi akustične modele glasov z orodji CUED HTK ali CMU Sphinx in jih uvoziti v naš razpoznavalnik.

2. Zasnova razpoznavalnika govora z velikim besednjakom

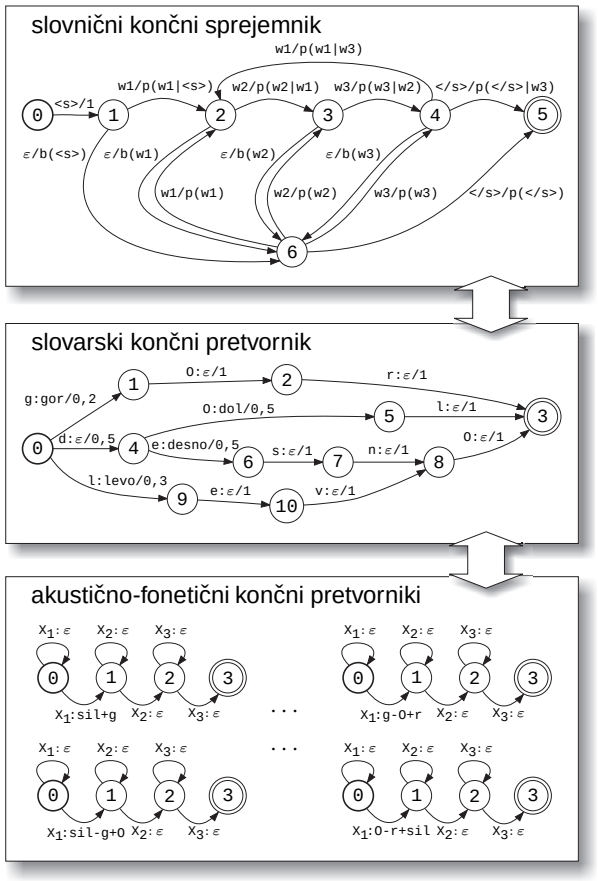
Pri zasnovi razpoznavalnika smo v grobem sledili zgledu, ki so ga predlagali Mohri, Pereira in Riley (Mohri et al., 2008). Razpoznavalnik smo zasnovali kot klasični trinivojski sklop uteženih končnih avtomatov. Sklop vsebuje najvišji slovnični uteženi končni sprejemnik (angl. weighted finite-state acceptor), vmesni slovarski uteženi končni pretvornik in na najnižjem nivoju množico akustično-fonetičnih uteženih končnih pretvornikov. Sklop je simbolično ponazorjen na sliki 1. Akustično-fonetični končni pretvorniki modelirajo utežno (ponavadi verjetnostno) preslikavo med nizi akustičnih govornih vzorcev in kontekstno odvisnimi alofoni. Vhod teh pretvornikov ponavadi modeliramo z zveznimi naključnimi spremenljivkami po vektorskem prostoru akustičnih značilk, izhod pa z diskretnimi naključnimi spremenljivkami po fonetični abecedi. Slovarski uteženi končni pretvornik modelira utežno preslikavo med nizi alofonov in besednimi oblikami. Slovnični uteženi končni sprejemnik modelira utežno preslikavo nizov besednih oblik, kar z drugimi besedami pomeni, da pripisuje nizom besednih oblik uteži.

Končne avtomate smo zasnovali kot kar se da ločene komponente, kar omogoča prilagodljivost razpoznavalnika morebitnim posebnostim govorjenega jezika, predvsem na slovničnem in slovarskem nivoju. Postopek razpoznavanja (dekodiranja) govora temelji na zgledu klasičnega iskalnega algoritma, ki predpostavlja prehajanjem t.i. utežnih žetonov med stanji končnih avtomatov (Young et al., 1989). Iskanje najbolj verjetnega niza prehodov med stanji avtomatov pri danem vhodnem nizu akustičnih govornih vzorcev poteka tako, da si avtomati med sabo izmenjujejo žetone s pripisanimi utežnimi vrednostmi (Dobrišek et al., 2009). Žetoni so opremljeni z dodatnimi povratnimi referencami, ki omogočajo vrnitev žetona v stanje, po katerem se je zgodil prehod v začetno stanje avtomata drugega nivoja.

V opisani zasnovi ima po naši oceni z vidika algoritemske zapletenosti najpomembnejšo vlogo slovarski končni pretvornik, iz katerega se žetoni selijo navzdol v akustično-fonetične končne pretvornike ali navzgor v slovnični končni sprejemnik. Pri razvoju celotnega razpoznavalnika se trenutno posvečamo predvsem optimizaciji slovarskega končnega pretvornika. Razvili smo izvirni optimizacijski postopek, ki precej zmanjša njegovo odvečnost in s tem precej pohitrili postopek dekodiranja govora. Manjše število stanj in prehodov med stanji namreč pomeni tudi manjše število žetonov, ki jih med dekodiranjem govora obdelujemo.

3. Zmanjševanje odvečnosti slovarskega končnega pretvornika

Izgovarjave besed so praviloma končne, zato slovarje z izgovarjavami besede modelirano z necikličnimi končnimi pretvorniki. Vhod v takšne pretvornike so nizi alofonov, izhod pa nizi grafemov in ničelnih simbolov, ki služijo kot identiteta za operacijo sklapljanja nizov grafemov. Posameznemu prehodu med dvema stanjema končnega pretvornika pripisujemo posamezen vhodni alofon. Pravimo, da pripisani vhodni alofon povzroči prehod med dvema stanjema



Slika 1: Razpoznavalnik govora kot tri-nivojski sklop uteženih končnih avtomatov.

pretvornika. Posameznemu prehodu med dvema stanjema pripisujemo še nize izhodnih grafemov ali izhodni prazen simbol ter utež, ki določa *ceno* prehoda.

Nizu prehodov od začetnega do končnega stanja pretvornika pravimo pot, ki jo ta opravi pri vhodnem nizu alofonov. Na tej poti pretvornik enkrat odda niz grafemov, ki predstavlja besedno obliko in je pripisana vhodnemu nizu alofonov. Pri preostalih prehodih na tej poti pretvornik odda prazne simbole. Pri tem ni pomembno, pri katerem prehodu končni pretvornik odda izhodni niz grafemov in pri katerih prazni simbol. Končna pretvornika, ki se razlikujeta le po tem, pri katerem prehodu na poti, ki jo opravita pri istem vhodnem nizu, oddata izhodni niz grafemov, štejemo za ekvivalentna. Ekvivalentne končne pretvornike ustvarjamo tako, da, kjer je to mogoče, premaknemo izhodni niz grafemov na naslednje ali predhodne prehode med stanji.

Slovarski končni pretvorniki so lahko deterministični ali nedeterministični. Pri determinističnih končnih pretvornikih nobeno stanje nima dveh ali več prehodov z istim vhodnim alofonom, ker bi ta sicer povzročil prehod v več različnih možnih stanj avtomata. Zaradi manjše odvečnosti imajo zato deterministični končni pretvorniki prednost pred nedeterminističnimi. Mohri s sodelavci (Mohri et al., 2002) je pokazal, da se lahko nedeterminiranost končnih pretvornikov vedno odpravi ali vsaj zmanjša z uporabo klasičnih postopkov za determinizacijo končnih sprejemnikov. Pri tem postopku se končni pretvornik začasno prekodira v

končni sprejemnik tako, da se prehodom pripisane izhode obravnava skupaj s pripisani vhodi kot združene vhode, ki povzročajo prehode med stanji končnega sprejemnika.

Odvečnost determinističnega končnega sprejemnika lahko nadalje zmanjšamo s klasičnimi postopki minimizacije, ki združijo njegova ekvivalentna stanja (Revuz, 1992). Za dve stanji končnega sprejemnika pravimo, da sta ekvivalentni, če isti niz vhodnih simbolov povzroči prehode od teh dveh stanj do končnega stanja sprejemnika. Determiniziran in minimiziran končni sprejemnik na koncu postopka prekodiramo nazaj v končni pretvornik tako, da razdružimo vhodne in izhodne simbole.

Opisani postopek zmanjševanja odvečnosti se izvaja v svežnju nad celotnim izhodiščnim slovarskim končnim pretvornikom. Izvedemo ga namreč tako, da najprej za celotni slovar zgradimo izhodiščni nedeterministični končni pretvornik, ki za vsak vhodni niz alofonov in pripisano izhodno besedno obliko vsebuje svojo pot od začetnega do končnega stanja. Pri odpravljanju odvečnosti sta nato možni dve strategiji. Pri prvi najprej premaknemo izhodne besedne oblike proti začetnemu stanju končnega pretvornika, nakar izvedemo njegovo minimizacijo. Nato premaknemo izhodne besedne oblike proti končnemu stanju tako daleč, kot je mogoče, da še ohranimo ekvivalentnost pretvornika, nakar izvedemo njegovo determinizacijo. Druga strategija je obratna. Najprej premaknemo izhodne besedne oblike proti končnemu stanju, izvedemo determinizacijo, premaknemo izhodne besedne oblike proti začetnemu stanju končnega pretvornika in izvedemo njegovo minimizacijo. Vse korake obeh različic postopka lahko izvedemo s programskimi orodji OpenFST. V nadaljevanju z OFST-I označujemo prvo in z OFST-F drugo različico postopka odstranjevanja odvečnosti, ki smo jo pri poskusih izvedli s temi orodji.

Na opisan način precej zmanjšamo odvečnost končnih pretvornikov, vendar te ne moremo povsem odpraviti. Možnost premikanja izhodnih nizov grafemov namreč ustvarja obsežno množico ekvivalentnih pretvornikov in pri vsakem od njih lahko po izvedbi determinizacije in minimizacije dosežemo drugačen končni pretvornik. Poskusi pokažejo, da sta slovarska končna pretvornika, ki ju pridobimo z eno ali drugo različico postopka, sicer ekvivalentna, vendar se med sabo znatno razlikujeta po številu stanj in prehodov med stanji (Tabela 1).

4. Prirastni postopek zmanjševanja odvečnosti

Druga možnost pri gradnji slovarskega pretvornika je njegova postopna gradnja s sprotnim zmanjševanjem odvečnosti za vsako dodano besedno obliko posebej. V tem primeru za vsako dodano besedno obliko tvorimo svoj končni pretvornik z eno samo potjo, ki ga nato z operacijo unije pridružimo obstoječemu naraščajočemu slovarskemu končnemu pretvorniku. Po vsaki operaciji unije nato izvedemo odstranjevanja njegove odvečnosti. Poskusi so pokazali, da na ta način pridobimo slovarske končne pretvornike, ki imajo znatno manjšo odvečnost od pretvornikov, pridobljenih s klasičnim postopkom. Žal pa je takšna izvedba postopka prirastnega odstranjevanja odvečnosti bistveno bolj računsko zahtevna in je pri velikih slovarjih že skoraj neuporabna.

Pri raziskovalnem delu na tem področju nam je uspelo razviti izvirni prirastni postopek, ki izvede zmanjševanja odvečnosti slovarskega končnega pretvornika v približno linearnem času. Prva različica je bila razvita za slovarske končne pretvornike, ki so vrste Moore (Dobrišek et al., 2009). Kasneje smo postopek prilagodili pretvornikom, ki so vrste Mealy (Dobrišek et al., 2010) in jih v osnovi podpirajo tudi orodja iz zbirke OpenFST.

5. Pridobivanje obsežnejšega slovarja izgovarjav

Prirastni postopek tvorjenja slovarskih končnih pretvornikov se je izkazal za zelo učinkovitega, vendar doslej še nismo raziskali, kje so njegove meje. Naša dolgoročna usmeritev pri razvoju slovenskega razpoznavalnika govora je povečati obseg njegovega besednjaka preko pol milijona različnih besednih oblik. Tako obsežnega slovarja z izgovarjavami v našem laboratoriju trenutno še ne premoremo in po naših informacijah ga od drugih raziskovalnih skupin ali podjetij v naši državi tudi ni mogoče pridobiti v proste raziskovalne namene. V okviru našega preteklega projekta razvoja spletnega bralnika za slepe in slabovidne pa smo za raziskovalne namene v obdobju od leta 2002 do 2007 pridobili besedila enajstih slovenskih periodičnih publikacij, to je glavnih slovenskih dnevnikov in nekaj revij. Skupni obseg teh besedil je prek dvesto milijonov besed. V raziskavi smo uporabili le besedila dnevnika Delo v skupnem obsegu dobrih šestdeset milijonov besed.

Besedila smo najprej obdelali s pomočjo besedilno-besedilnega pretvornika, ki je del našega sintetizatorja slovenskega govora S6TTS (Ž. Gros et al., 1997). Ta pretvornik spreminja posebne simbole, kot se ločila, števila, kratice ipd v razširjene besedne oblike. S tem smo prišli do približka tega, kako bi ta besedila bral ali narekoval človek. Pretvorjena besedila smo nato obdelali z orodji CMU SLM, ki se uporabljajo za pridobivanje n-gramskih jezikovnih modelov. Pridobljeni korpus različnih besednih oblik smo ročno pregledali in deloma odstranili napake, ki smo jih odkrili s preverjanjem frekvenc besednih oblik ipd. S tem smo pridobili besednjak s preko šesto osemdeset tisoč različnih besednih oblik.

Izgovarjave vseh besed v besednjaku smo pridobili z grafemsko-fonemskim prevornikom omenjenega sintetizatorja govora. S tem smo pridobili razmeroma obsežen slovar izgovarjav besed, ki je zaenkrat le približek ciljnemu slovarju. Kakovost slovarja bomo s pomočjo študentov ovrednotili s preizkušanjem razpoznavalnika govora, ki je v celoti udejanjen z orodji CMU Sphinx, in se bo uporabljal za izbrana omejena področja uporabe.

6. Poskusi z velikim slovarjem izgovarjav

V članku poročamo o rezultatih postopkov zmanjševanja odvečnosti slovarskih končnih pretvornikov pri dveh velikih slovarjih izgovarjav. Poleg navedenega slovenskega slovarja izgovarjav, označenega z DELO-SL, smo v poskusih uporabili tudi prosto dostopen slovar CMU-US, ki pokriva severno ameriški angleški govorjeni jezik in vključuje približno sto triintrideset tisoč različnih besednih oblik.

minimizacija	CMU-US (133k)	DELO-SL (680k)
brez	717k/850k	5499k/6179k
OFST-I	80k/214k	380k/1060k
OFST-F	64k/197k	432k/1113k
ITHM-O	33k/166k	142k/822k
ITHM-S	33k/166k	121k/801k

Tabela 1: Število (približno v tisočih) stanj in prehodov med stanji slovarskih končnih pretvornikov, pridobljenih z različnimi postopki zmanjševanja njihove odvečnosti.

Obseg tega slovarja je nekajkrat manjši kot je obseg uporabljenega slovenskega slovarja, vendar moramo pri tem upoštevati razlike med obema jezikoma glede števila pregibnih oblik besed. Po naši oceni sta slovarja po številu leksemov v grobem primerljiva.

Primerjali smo velikosti slovarskih končnih pretvornikov, ki smo jih dosegli s štirimi različnimi postopki oziroma strategijami zmanjševanja odvečnosti pri obeh slovarjih. Tabela 1 podaja dosežene velikosti slovarskih končnih pretvornikov v smislu števila njihovih stanj in prehodov med stanji. Oznaki OFST-I in OFST-F označujeta dve različici postopka, ki smo ju opisali v tretjem poglavju. Oznaki ITHM-O in ITHM-S pa označujeta dve različici prirastnega postopka ostranjevanja odvečnosti, ki se razlikujeta po tem, po kakšnem vrstnem redu dodajamo besedne oblike v naraščajoči slovarski končni pretvornik. Pri prvi različici so dodajane besedne oblike predhodno leksikografsko urejen, v drugem primeru pa naključno permutirane.

Predstavljeni rezultati potrjujejo našo ugotovitev, da je prirastni postopek zmanjševanja odvečnosti končnih pretvornikov bistveno bolj uspešen, kot klasični postopki, ki jih ponujajo orodja OpenFST. Jasno se tudi pokaže razlika med obema jezikoma. Zaradi pregibnih oblik besed, ki se med sabo pogosto razlikujejo le po spremembah v predponah in končnicah, je zmanjšanje odvečnosti pri slovenskem slovarju izgovarjav bistveno večja kot pri angleškem slovarju. Zanimiva je znatna razlika med različicama ITHM-O in ITHM-S pri slovenskem slovarju. Tudi to razliko pripisujemo značilnostim pregibnih oblik slovenskih besed. Rezultati kažejo, da je prirastni postopek malo manj učinkovit, če so vhodne besedne oblike urejen po črkah.

Poleg večje učinkovitosti se je prirastni postopke izkazal tudi za hitrejšega od orodij OpenFST, in to kljub temu, da smo ga udejanjili v programske jeziku Java in da so orodja OpenFST udejanjena v programskem jeziku C++. Potreben čas za izvedbo prirastnega postopka pri slovenskem slovarju je bil dvakrat, pri angleškem slovarju celo petkrat krajši kot pri uporabi orodij OpenFST.

Ker imamo opravka s splošnimi optimizacijskimi postopki, lahko pričakujemo podobne rezultate tudi pri odstranjevanju odvečnosti slovničnega končnega sprejemnika in akustično-fonetičnega končnega pretvornika. Pri slednjih je potrebno postopke zgolj prilagoditi dejstvu, da so vhodi pretvornika modelirani kot zvezne naključne spremenljivke. Običajne postopke vezave parametrov tovrstnih akustično-fonetičnih modelov namreč lahko prepoznamo kot eno od možnih različic postopkov minimizacije končnih pretvornikov.

7. Zaključek

Celoten razpoznavalnik govora z velikim besednjakom, ki bo udejanjen v programskem jeziku Java, je še v razvoju in trenutno povezujemo zgrajeni slovarski končni pretvornik z akustično-fonetičnimi končnimi pretvorniki, pri udejanjenju katerih se zgledujemo po zasnovi javanskih komponent orodij CMU Sphinx 4. S tem bomo omogočili neposredno uporabo akustično-fonetičnih modelov, ki jih lahko zgradimo s pomočjo orodij SphinxTrain.

Poskusno smo že izvedli pretvorbo pridobljenega slovarskega končnega pretvornika v fonetični graf, ki ga podpirajo orodja CUED HTK. Z njimi smo udejanjili slovarsko *siljeni* razpoznavalnik glasov, ki je brez večjih zapletov tekel v stvarnem času na običajnem računalniku, in to kljub temu, da ta orodja niso tako zelo časovno optimirana. Po dosedanjih izkušnjah zato pričakujemo, da bo z uporabo predstavljenih učinkovitih postopkov zmanjševanja odvečnosti končnih pretvornikov na današnjih običajnih osebnih računalnikih tudi v programskem jeziku Java možno udejanjili razpoznavalnik govora z besednjakom, ki obsega vsaj več sto tisoč besed.

8. Literatura

C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, in M. Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. V: *CIAA*, str. 11–23.

A. W. Black, H. Zen, in K. Tokuda. 2007. Statistical parametric speech synthesis. V: *Proceedings of ICASSP 2007*, str. 1229–1232, Honolulu, Hawaii, US.

S. Dobrišek, B. Vesnicer, in F. Mihelič. 2009. A sequential minimization algorithm for finite-state pronunciation lexicon models. V: *Speech and intelligence: proceedings of Interspeech 2009*, str. 720–723, Brighton, UK.

S. Dobrišek, J. Žibert, in F. Mihelič. 2010. Towards the optimal minimization of a pronunciation dictionary model. V: Petr Sojka, Ales Horák, Ivan Kopecek, in Karel Pala, ur., *TSD-2010*, Lecture Notes in Computer Science, str. 267–274, Brno, Czech Republic. Springer.

F. Jelinek. 1997. *Statistical methods for speech recognition*. MIT Press, Cambridge, MA, USA.

M. Mohri, F. C. N. Pereira, in M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88.

M. Mohri, F. C. N. Pereira, in M. Riley, 2008. *Speech recognition with weighted finite-state transducers*, pogl. Part E: Speech recognition. Springer-Verlag, Heidelberg, Germany.

D. Revuz. 1992. Minimisation of acyclic deterministic automata in linear time. *Theoretical Computer Science*, 92(1):181–189.

J. Ž. Gros, N. Pavešič, in F. Mihelič. 1997. Text-to-speech synthesis: a complete system for the slovenian language. *J. Comput. Inf. Technol.*, 5(1):11–19.

S. J. Young, N. H. Russel, in J. H. S. Thornton. 1989. Token passing: A simple conceptual model for connected speech recognition systems. Tehnično poročilo F/INFENG/TR.38, Cambridge University Engineering Department, Cambridge, UK.

Razpoznavalnik tekočega govora UMB Broadcast News 2010: nadgradnja akustičnih in jezikovnih modelov

Andrej Žgank, Mirjam Sepesy Maučec,

Inštitut za elektroniko in telekomunikacije, Univerza v Mariboru
Smetanova ul. 17, 2000 Maribor
andrej.zgank@uni-mb.si, mirjam.sepesy@uni-mb.si

Povzetek

V članku bomo predstavili nadgradnjo sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast News, katerega domena so dnevno-informativne oddaje. Verzijo sistem predstavljeno leta 2008 smo nadgradili z izboljšanimi akustičnimi in jezikovnimi modeli. Na področju akustičnega modeliranja smo se osredotočili na modul izločanja značilk, kjer smo primerjali med seboj različne konfiguracije tako s stališča pravilnosti razpoznavanja kot tudi s stališča hitrosti delovanja. Jezikovne modele smo nadgradili s podporo za trigrami, ki boljše modelirajo tekoči govor. V učni nabor za jezikovno modeliranje smo vključili dodatni besedilni korpus FidaPLUS. Z nadgrajenim sistemom smo dosegli pravilnost razpoznavanja besed 71,3%.

UMB Broadcast News 2010 continuous speech recognition system: new acoustic and language models

This paper presents an improved version of a Slovenian continuous speech recognition system for the Broadcast News domain. A new set of acoustic and language models was included in the system presented in year 2008. In the area of acoustic modeling, various configurations of feature extraction module were compared. The focus was given on improving accuracy and speed of speech recognition system. A trigram language model was introduced in current version of recogniser. The FidaPLUS text corpus was added to language model training. The improved speech recognition system achieved 71,3% word accuracy.

1. Uvod

Z razvojem širokopasovnih internetnih povezav ter naraščanjem števila različnih sodobnih telekomunikacijskih storitev se je izredno povečala količina multimedijskega gradiva, ki je dostopno uporabnikom. Kadar želimo v takšni količini gradiva poiskati želeno informacijo, smo prisiljeni uporabiti metode avtomatskega iskanja po vsebini. Ena izmed ključnih funkcionalnosti, ki to omogočajo, je avtomatsko razpoznavanje govora.

V članku[1] bomo predstavili nadgradnjo sistema avtomatskega razpoznavanja tekočega govora UMB Broadcast News, ki je trenutno najkompleksnejši sistem za razpoznavanje slovenskega govora. Domena sistema so dnevno-informativne televizijske oddaje. Prva različica sistema UMB BN je bila predstavljena leta 2006 (Žgank et al., 2006) in je delovala samo za bran govor v studijskem okolju brez zvočnega ozadja. Sistem s polno podpora za različne tipe govora in akustična ozadja smo predstavili leta 2008 (Žgank et al., 2008a). Verzijo iz leta 2008 smo nadgradili z izboljšanimi akustičnimi modeli, kjer smo spreminjali predvsem lastnosti povezane s parametri izločanja značilk. Cilj je bil izboljšati modeliranje v akustično-fonetičnem prostoru in preveriti možnosti uporabe različnih konfiguracij izločanja značilk. Izboljšanje modeliranja je potreben korak za razširitev možnosti dodajanja različnih modulov za predprocesiranje (npr. avtomatska segmentacija in klasifikacija, grozdenje govorcev). Hkrati smo preučili še eno izmed možnosti, kako pohitriti sistem razpoznavanja govora. Slednje se je pokazalo kot potreben korak pri pripravi sistema razpoznavanja govora za delovanje v dveh iteracijah, kar predstavlja naslednjo fazo razvoja.

Na področju jezikovnih modelov smo vključili uporabo trigramov, ki zaradi modeliranja daljših besednih zaporedij bistveno izboljšajo pravilnost delovanja. Hkrati smo v besedilne korpuse za učenje jezikovnih modelov vključili tudi korpus FidaPLUS (Arhar & Gorjanc, 2007), kar predstavlja njegovo prvo uporabo v sistemih avtomatskega razpoznavanja govora. Predvsem vključitev tako velikega korpusa kot je FidaPLUS omogoča prehod na N-gramske modele višjih redov.

V nadaljevanju članka bomo najprej predstavili jezikovne vire, ki smo jih uporabili pri izgradnji sistema avtomatskega razpoznavanja govora. V tretjem poglavju bo sledila predstavitev nadgradnje akustičnih in jezikovnih modelov eksperimentalnega sistema. Rezultate in analizo vrednotenja razpoznavanja govora bomo predstavili v četrtem poglavju. Zaključek in smernice za nadaljnje delo bomo podali v petem poglavju.

2. Jezikovni viri

Jezikovni viri predstavljajo ključno komponento v postopku izgradnje avtomatskega razpoznavalnika govora, saj so potrebni tako za gradnjo akustičnih kot tudi jezikovnih modelov. Pri tem imajo pomembno vlogo tudi značilnosti jezika, saj potrebujemo za doseganje primerljive kakovosti razpoznavanja govora zelo različen obseg jezikovnih virov. Analize so tako pokazale, da slovenski jezik s tega vidika sodi med kompleksnejše jezike. Za jezikovne vire v domeni Broadcast News je značilno, da je praviloma v njihovo izdelavo potrebno vložiti dosti ročnega dela (označevanje, zapisovanje izgovorjenega), kar pomembno vpliva na razpoložljivi obseg gradiva.

2.1. Govorna baza BNSI Broadcast News

Osnovno učenje akustičnih modelov sistema UMB BN smo izvedli z govornim korpusom slovenske baze BNSI Broadcast News (Žgank et al., 2004), ki obsega 36 ur zapisanega govornega materiala iz obdobja 1999-2003. V

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. P2-0069.

korpus so vključene različne dnevno-informativne oddaje RTV Slovenija (TV Dnevnik, Odmevi). Govorni posnetki so bili v celoti ročno segmentirani, označeni in zapisani. Slovenska baza BNSI je dostopna pri evropski organizaciji ELRA/ELDA (ELRA, 2008).

Postopek priprave na učenje akustičnih modelov zahteva dodatno ročno delo za pripravo in uskladitev vseh vključenih jezikovnih virov, predvsem če želimo doseči visoko kvaliteto razpoznavanja govora. V primerjavi s sistemom UMB BN, predstavljenim leta 2008 (Žgank et al., 2008a) tokrat nismo dodatno povečevali govornega korpusa. Vrednotenje vpeljanih metod smo ponovno izvedli s celotnim testnim naborom v skupni dolžini približno 3 ure govora, ki vsebuje posnetke v vseh različnih f-razredih (Schwartz et al., 1997).

2.2. Tekstovne baze

Za učenje jezikovnih modelov smo obstoječim besedilnim korpusom dodali še korpus FidaPLUS.

FidaPLUS korpus (Arhar & Gorjanc, 2007) predstavlja razširitev korpusa FIDA. FIDA je referenčni korpus slovenskega pisanega jezika in obsega 100 mio besed iz različnih tekstovnih virov iz obdobja 1990-1999. FidaPLUS obsega 621 mio besed in dodaja besedila iz tekstovnih virov iz obdobja 1999-2006. Večji del korpusa predstavljajo časopisni in revijalni članki ter knjige. Nekaj besedil izvira tudi iz Spleta. Korpus je lematiziran in označen z morfosintaktičnimi značkami, ki pa v pričujočem članku niso bile uporabljene.

Obstoječa korpusa BNSI-Speech in BNSI-Text sta korpusa govorjenega jezika, obstoječi korpus Večer in novo dodani korpus FidaPLUS pa sta korpusa pisanega jezika. Med govorjenim (predvsem spontano govorjenim) in pisanim jezikom je velika razlika (Stouten et al., 2006). Stavki v korpusih pisanega jezika so daljši od izjav v korpusih govorjenega jezika. Številnih pojavov, ki so značilni za govor (npr. uporaba mašil, ponavljanje, napačni starti ipd.), v pisanem jeziku ne zasledimo (Žgank et al., 2008). Uravnotežen vpliv različnih jezikovnih virov smo dosegli z linearno interpolacijo na BNSI-Devel korpusu. BNSI-Devel korpus je po strukturi enak BNSI-Speech korpusu in obsega 4 oddaje. Poskrbeli smo, da med korpusi BNSI-Speech, BNSI-Devel in BNSI-Eval (ki je namenjen testiranju) ni vsebinskega prekrivanja.

3. Nadgradnja sistema UMB BN

Okvirne karakteristike sistema UMB BN iz leta 2008 so predstavljene v tabeli 1, več informacij je na voljo v (Žgank & Kačič, 2005/1; Žgank et al., 2005/2; Žgank et al., 2008a; Young et al., 1994; Odell, 1995; Grašič et al., 2008).

UMB BN 2008	
Izločanje značilk	MFCC, PLP
Karakteristike značilk	Okno dolžine 32 ms, korak 10 ms, MFCC 12 koef., energija, 1. in 2. odvod, 26 filtrov
Akustični model	Medbesedni trigrafemi
Kompleksnost AM	16 gaussovih porazdelitev
Jezikovni modeli	Interpolirani bigrami
Velikost slovarja	64.000 besed

Tabela 1: Karakteristike sistema avtomatskega razpoznavanja tekočega govora UMB BN 2008.

Predstavljen sistem avtomatskega razpoznavanja govora je ob uporabi ročne segmentacije dosegel pravilnost razpoznavanja besed 66,0%, kar je predstavljajo izhodišče za nadgradnje trenutnega sistema.

3.1. Nadgradnja akustičnih modelov

Predhodni nabor akustičnih modelov predstavlja solidno izhodišče za nadaljnje eksperimente v smeri izboljšanja modeliranja v akustično-fonetičnem prostoru. Trenutno smer razvoja sistema z eno iteracijo smo zastavili v smeri izboljšav v modulu izločanja značilk.

Na osnovi dosedanjih rezultatov smo še vedno ostali pri uporabi obeh tipov izločanja značilk (mel kepstralni koeficienti (MFCC) in koeficienti perceptivnega linearnega napovedovanja (PLP (Hermansky, 1990))), saj različne konfiguracije dajejo različne rezultate za posamezni tip značilk.

Na področju izločanja značilk smo tako najprej izvedli primerjavo med različnima dolžinama okna za izločanje značilk. Dosedanji sistem je uporabljal dolžino 32 ms. Zaradi možnosti kombiniranja z ostalimi moduli smo postavili vzporedni sistem, pri katerem je dolžina okna 25 ms.

Na osnovi preliminarnih rezultatov dobljenih na avtomatskem razpoznavalniku govora za govorno vodene telefonske storitve (IVR) smo povečali število mel filtrov v izločanju MFCC značilk iz 26 na 42 (HTK, 2010). Na takšen način povečana razločljivost je pri IVR sistemu pokazala tendenco izboljšanja rezultatov.

Kot zadnji korak v nadgradnji akustičnih modelov smo pripravili vzporedni sistem, ki uporablja namesto 12 mel kepstralnih koeficientov samo 8 koeficientov. Tako se skupna velikost vektorja značilk zmanjša iz 39 na 27 (zmanjšanje za ~30%). S takšnim pristopom sicer nekoliko poslabšamo rezultate razpoznavanja govora, vendar lahko zaradi zmanjšane kompleksnosti akustičnih modelov bistveno pohitrimo razpoznavanje. Tako okrnjeni akustični modeli bodo uporabljeni v prvi iteraciji naslednje verzije sistem UMB BN, ki bo uporabljala dve iteraciji. Tako pri sistemu s spremenjenim številom mel filtrov kot tudi pri sistemu s spremenjenim številom koeficientov smo pri tvorbi trigrafemskih akustičnih modelov (Žgank & Kačič, 2005/1) uporabili spremenjen prag združevanja modelov, kar je bilo potrebno zaradi sprememb v značilkah.

3.2. Nadgradnja jezikovnih modelov

Za gradnjo jezikovnih modelov smo uporabili orodje SRI Language Modeling Toolkit (Stolcke, 2002). Na osnovi besedilnih korpusov smo zgradili bigramski in trigramski model. Jezikovni model je sestavljen iz štirih komponent: prvo komponento smo zgradili na korpusu BNSI-Speech, drugo na korpusu BNSI-Text, tretjo na korpusu Večer in četrto na korpusu FidaPLUS. V prvih treh komponentah smo ohranili vse bigrame in trigrame iz učnih korpusov, v četrti komponenti pa smo izločili vse n-grame s frekvenco 1.

Slovar je obsegal 64.000 besed. Vseboval je vse besede korpusov BNSI-Speech in BNSI-Text. Do velikosti 64.000 smo ga dopolnili z najpogostejšimi besedami iz korpusa Večer.

Uporabili smo Good-Turingovo glajenje in sestopanje po Katz-u. Preizkusili smo tudi modificirano Kneser-Ney glajenje, a bistveni razlik v rezultatih ni bilo. Interpolacijske koeficiente komponent smo določili tako, da smo minimizirali perpleksnost jezikovnega modela na BNSI-Devel. Interpolacijski koeficienti po komponentah so predstavljeni v tabeli 2. Perpleksnost bigramskega interpoliranega jezikovnega modela na BNSI-Eval je znašala 359, trigramskega modela pa 246. Delež besed izven slovarja (OOV) je bil 4.22%. Besed, ki so bile izven slovarja, nismo posebej modelirali na nivoju akustičnega modela. Razširitev bigramskega modela s komponento FidaPLUS je perpleksnost izboljšala za 12%. Prehod iz bigramskega na trigramski jezikovni model in uporaba korpusa FidaPLUS sta prinesla 40% izboljšanje perpleksnosti. Število bigramov se je podvojilo. Dodanih je bilo 33.6 mio trigramov.

Komponenta	2g	3g
BNSI-Speech	0,20	0,18
BNSI-Text	0,28	0,24
Večer	0,15	0,12
FidaPLUS	0,37	0,46

Tabela 2: Koeficienti λ komponent v interpoliranem bigramskem (2g) in trigramskem (3g) modelu.

4. Rezultati eksperimentov

Vrednotenje različnih konfiguracij sistema razpoznavanja govora UMB BN smo izvedli na evalvacijskem naboru baze BNSI, pri tem pa smo tokrat uporabljali izključno ročno segmentacijo, ki ne vnese dodatne napake v rezultate razpoznavanja govora (Žgank et al., 2008a; NIST, 2010). Rezultate podajamo v odstotku razpoznanih besed ter v odstotku pravilno razpoznanih besed. V slednjem primeru upoštevamo poleg razpoznanih besed še vrinjene in izbrisane besed, ki poslabšajo skupni rezultat.

V prvem koraku (tabela 3) smo ovrednotili vpliv trigramskih jezikovnih modelov na pravilnost razpoznavanja govora. Teste smo izvedli za oba tipa značilk (MFCC in PLP), saj smo želeli preveriti, ali je v kombinaciji trigramov in različnih tipov značilk prišlo do kakšne spremembe.

Sistem	Razpoznanih(%)	Pravilnih(%)
2g, MFCC	69,0	65,7
2g, PLP	69,6	66,0
3g, MFCC	70,7	67,5
3g, PLP	71,4	68,0

Tabela 3: Rezultati razpoznavanja govora z interpoliranimi trigramskimi jezikovnimi modeli.

Tudi rezultati razpoznavanja utemeljujejo prehod iz bigramskih na trigramske jezikovne modele. Prehod na trigramske modele je v primeru MFCC značilk prinesel 2.74% izboljšanje, v primeru trigramskih modelov pa 2.94%. Primerjava obeh tipov značilk je pokazala, da se tudi z uporabo trigramskih jezikovnih modelov ohrani prednost PLP značilk.

Sistem	Razpoz.(%)	Prav.(%)
JM_FP1, 2g, MFCC	70,0	67,4
JM_FP1, 3g, MFCC	73,6	71,0
JM_FP2, 2g, MFCC	60,9	57,7
JM_FP2, 3g, MFCC	73,4	71,1

Tabela 4: Rezultati razpoznavanja govora z dodanim besedilnim korpusom FidaPLUS.

Pri sistemih, katerih rezultati so predstavljeni v tabeli 4, je bil pri izgradnji jezikovnega modela dodatno uporabljen besedilni korpus FidaPLUS. V verziji JM_FP1, je bil korpus FidaPLUS dodan h preostalim trem osnovnim besedilnim korpusom. V verziji sistema JM_FP2 pa smo iz učenja jezikovnega modela izločili korpus Večer, saj je ta v veliki meri tako že vsebovan v korpusu FidaPLUS. V vseh različnih kombinacijah jezikovnih modelov smo uporabili identičen fonetični slovar s 64.000 besedami (glej poglavje 3.2). Rezultati, dobljeni z bigramskimi jezikovnimi modeli, so na prvi pogled presenetljivi. Izločitev komponente Večer je botrovala precejšnjemu poslabšanju rezultatov. Razlog vidimo predvsem v povezavi komponente Večer in načinom izbire besed, ki so v slovarju. Ker slovar vsebuje pretežno najpogostejše besede korpusa Večer, le-te najuspešneje modelira komponenta Večer interpoliranega modela. Čeprav je korpus Večer vsebovan v korpusu FidaPLUS, je zaradi obsega korpusa njegov vpliv bistveno bolj oslabljen kot v primeru uporabe ločene komponente, učene izključno na tem besedilnem gradivu. V primeru trigramskega modela je razlika med JM_FP1 in JM_FP2 zanemarljivo majhna. V tem modelu ima ključno napovedno moč komponenta FidaPLUS, ki ima tudi največjo interpolacijsko utež. Z vključitvijo trigramskih jezikovnih modelov in korpusa FidaPLUS smo dosegli 71,1% delež pravilno razpoznanih besed, kar predstavlja absolutno izboljšanje za 5,1%. Sistem podobne kompleksnosti (slovar 60k, 8% OOV), razvit za češki jezik je dosegel 70,8% delež pravilno razpoznanih besed (Podversky & Machek, 2005).

Sistem	Razpoz.(%)	Prav.(%)
2g, MFCC_32ms	70,0	67,4
2g, MFCC_25ms	70,4	67,8
2g, PLP_32ms	70,9	68,0
2g, PLP_25ms	70,5	67,7
3g, MFCC_32ms	73,6	71,0
3g, MFCC_25ms	73,5	71,0
3g, PLP_32ms	73,9	70,9
3g, PLP_25ms	73,6	70,7

Tabela 5: Rezultati razpoznavanja govora z jezikovnimi modeli JM_FP1 in različnim tipom značilk ter dolžino okna.

Primerjava sistemov UMB BN z dolžino okna izločanja značilk 32 ms in 25 ms je podana v tabeli 5. Za različne kombinacije tipov značilk (MFCC, PLP), dolžin okna (32 ms, 25 ms) in jezikovnih modelov (bigrami,

trigrami) smo sicer dobili manjše razlike v rezultatih, ki pa niso statistično signifikantne.

Sistem	Razpoz.(%)	Prav.(%)
2g, MFCCm	70,9	68,1
2g, MFCCm, FB42	70,3	67,6
2g, MFCCm, 8+1	66,0	64,1
3g, MFCCm	74,0	71,3
3g, MFCCm, FB42	73,7	71,0
3g, MFCCm, 8+1	69,7	67,7

Tabela 6: Rezultati razpoznavanja govora z jezikovnimi modeli JM_FP1 in različnimi konfiguracijami modula za izločanje značilk.

Vrednotenje različnih konfiguracij izločanja značilk je predstavljeno v tabeli 6. Povečanje števila mel filtrov (oznaka FB42 v tabeli 6) je v sistemu UMB BN poslabšalo rezultat za 0,5% (bigrami) oz. 0,3% (trigrami). Na osnovi analize in primerjave rezultatov s preliminarnimi rezultati na sistemu IVR je možno predpostaviti, da je za poslabšanje rezultatov kriv različen tip govora. Sistem IVR namreč podpira samo izolirane in vezane besed v telefonskem okolju, sistem UMB BN pa tekoč govor v različnih akustičnih okoljih.

V primeru uporabe samo 8 mel kepstralnih koeficientov (oznaka 8+1 v tabeli 6) je po pričakovanjih prišlo do poslabšanja rezultatov. Pri bigramskih jezikovnih modelih se je pravilnost poslabšala za 4,0%, pri trigramskih pa za 3,6%. Uporaba akustičnih modelov s tako okrnjeno kompleksnostjo je pohitrila delovanje razpoznavalnika govora za cca. 16% v primeru bigramov, ter za cca. 19% v primeru trigramov. Ob upoštevanju pohitritve delovanja razpoznavalnika govora je takšno poslabšanje pravilnosti sprejemljivo v primeru vključitve v sistem z dvema iteracijama.

5. Zaključek

V članku smo predstavili trenutno stanje razvoja sistema za razpoznavanje tekočega govora UMB Broadcast News. Z uporabo različnih nadgradenj akustičnih in jezikovnih modelov smo uspešno izboljšali rezultate razpoznavanja govora, ter dosegli najvišjo pravilnost razpoznanih besed 71,3%. V naslednjem koraku razvoja bomo sistemu UMB BN dodali dodatno iteracijo razpoznavanja govora.

Zahvala

Zahvaljujemo se avtorjem besedilnega korpusa FidaPLUS, ki so nam omogočili njegovo uporabo za jezikovno modeliranje avtomatskega razpoznavalnika govora.

6. Literatura

Arhar, Š., Gorjanc, V., (2007). Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. Jezik in slovstvo 52/2., 95--110.

ELRA *BNSI Catalog Reference : S0275*: www.elra.info.

Grašič, M., Kos, M., Žgank, A., Kačič, Z., (2008). Two Step Segmentation Method Using Bayesian Information Criterion and Adapted Gaussian Mixtures Models. *Proc. Interspeech 2008* (v tisku), Brisbane, Avstralija.

Hermansky, H. (1990). Perceptual Linear Predictive Analysis of Speech. *J. Acoustic Soc. Americ*, v87, n4.

HTK domača stran, http://htk.eng.cam.ac.uk.

NIST *Sclite* domača stran (2010). http://www.nist.gov/speech/tools/

Odell, J.J., (1995). *The Use of Context in Large Vocabulary Speech Recognition.* Doktorska disertacija, Univerza v Cambridgeu, Velika Britanija.

Podversky, P., Machek, P., (2005). Speech Recognition of Czech – Inclusion of Rare Words Helps. *Proc. ACL Student Research Workshop,* Ann Arbor, ZDA.

Schwartz, R., Jin, H., Kubala, F., Matsoukas, S., (1997). Modeling those F-Conditions - or not. *Proc. DARPA Speech Recognition Workshop*, Chantilly, ZDA.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. International Conference on Speech and Language Processing, II: 901--904.

Stouten, F., Duchateau, J., Martens, J.-P., Wambacq, P., (2006). Coping With Disfluencies In Spontaneous Speech Recognition:Acoustic Detection And Linguistic Context Manipulation, Speech Communication vol. 48, issue 11, 1590--1606.Woodland, P. et al. (2001). CU-HTK March 2001 Hub5 System. *Proc. 2001 LVCSR Workshop.*

Young, S., Odell, J., Woodland, P., (1994). Tree-based State Tying for High Accuracy Acoustic Modelling. *Proc. ARPA Human Language Technology Conference* Plainsboro.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Verdonik, D., Kitak, J., Vlaj, D., Hozjan, V., Kačič, Z., Horvat, B., (2004). Acquisition and annotation of Slovenian broadcast news database. *Fourth international conference on language resources and evaluation, LREC 2004*, Lizbona, Portugalska.

Žgank, A., Kačič, Z., (2005/1). Primerjava treh tipov akustičnih osnovnih enot razpoznavalnika slovenskega govora. *Elektrotehniški vestnik*, 2005, Ljubljana, Slovenija.

Žgank, A., Horvat, B., Kačič, Z., (2005/2). Data-driven generation of phonetic broad classes, based on phoneme confusion matrix similarity. *Speech Communication*, vol. 47, issue 3, 379--393, november 2005.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., Kačič, Z., (2006). Osnovna zgradba razpoznavalnika slovenskega tekočega govora UMB Broadcast News. *Jezikovne tehnologije 2006*, Ljubljana, Slovenija.

Žgank, A., Rotovnik, T., Sepesy Maučec, M., (2008). Modeling Filled Pauses for Spontaneous Speech Recognition Applications. *Proc. Applications of Electrical Engineering*, Trondheim, Norveška.

Žgank, A., Kos, M., Kotnik, B., Sepesy Maučec, M., Rotovnik, T., Kačič, Z. (2008a). Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast news. *Jezikovne tehnologije 2008*, Ljubljana, Slovenija.

Analiza značilk linearne transformacije MLLR pri samodejnem razpoznavanju spontanih čustvenih stanj govorca

Tadej Justin, Rok Gajšek, Simon Dobrišek

Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška 25, 1000 Ljubljana
{tadej.justin, rok.gajsek, simon.dobrisek}@fe.uni-lj.si

Abstract

V tem prispevku primerjamo dva tipa izpeljave značilk za namen razvrščanja vzbujenega ter emocionalnega normalnega stanja govorca. Prvi temelji na linearni transformaciji po kriteriju največjega verjetja (MLLR), drugi pa predstavlja skupek akustičnih značilk (frekvenca spremembe predznaka, energija RMS, osnovna frekvenca, razmerje zvenečenega proti nezvenečenemu delu govora in koeficienti MFCC), ki je v literaturi velikokrat zaznamovan z terminom "state-of-the-art". Zbirke spontanega emocionalnega govora velikokrat vsebujejo številčno neuravnotežene vzorce med razredi objektov razpoznavanja. Tako lastnost lahko pripišemo tudi uporabljeni zbirki slovenskega spontanega emocionalnega govora AvID. Da bi se v večji meri oddaljili od problema prenaučenosti (ang. overfitting) razvrščevalnikov uporabimo postopek SMOTE, ki omogoča tvorjenje novih umetnih vzorcev manj zastopanega razreda razvrščanja. S programskim okoljem WEKA, ki omogoča tako predobdelavo in kasnejše razvrščanje, udejanjimo razvrščevalnik na osnovi metode podpornih vektorjev. Rezultate razvrščanja za prvi in drugi tip izpeljave značik primerjamo z uveljavljenima merama uspešnosti razvrščanja; priklic (ang. recall - R) in natančnost (ang. precision - P). Rezultati pokažejo, da udejanjen razvrščevalnik emocionalnih stanj s "state-of-the-art" sklopom značilk bolje ločuje med testnimi vzorci obeh razredov (AR $70,42\%$) kot udejanjen razvrščevalnik z značilkami MLLR (AR $59,12\%$).

The analysis of features based on MLLR for classifying spontaneous speaker's emotional states

In this paper we compare two types of feature sets for the task of classifying aroused and normal emotional state. The first one is based on the maximum likelihood estimation of linear transformations (MLLR), and the second one is suprasegmental modeling of acoustic features (zero-crossing-rate, RMS energy, pitch, HNR and MFCC coefficients), known as the "state-of-the-art" in emotion recognition from speech. Databases of spontaneous emotional speech usually contains unbalanced emotional classes. This is also the case with AvID - the Slovenian emotional database of spontaneous speech, which was used in our experiments. We try to minimize the problem of overfitting using the SMOTE filter, that over-samples minority class by synthetically generating samples. The WEKA software was used for pre-processing and classification, where we used the support vector machine algorithm. As criterion for comparing the results recall and precision were used. The results show that the classifier of emotional state trained with the "state-of-the-art" features better distinguishes the samples of normal and aroused emotional state (AR $70,42\%$), as the classifier trained with the MLLR features (AR $59,12\%$).

1. Uvod

Želja raziskovalce na področju komunikacije med človekom in strojem je nedvomno, da bi lahko ljudje komunicirali s stroji na ljudem najbolj prijazen način. Uporaba govora v ta namen je zato ena izmed ustreznejših možnosti komunikacije. Za razvoj uspešnega in bolj naravnega sporazumevanja ni dovolj, da se stroj ne zaveda le vsebine sporočila, ampak tudi samega načina sporočanja. Raziskovanje postopkov avtomatskega prepoznavanja emocij je eden ključnih problemov pri doseganju naravne komunikacije med človekom in strojem.

Pridobitev podatkovne zbirke, ki vsebuje spontan emocionalni govor, je v razvoju avtomatskega razpoznavalnika emocionalnih stanj govorca ena zahtevnejših nalog. Učenje stroja je predvsem odvisno od dobro označene podatkovne zbirke emocionalnega govora. Le v primerih, ko so govorni posnetki konsistentno označeni lahko pridobimo značilke, ki nosijo pravo informacijo o emocionalnem stanju govorca. V tem prispevku se osredotočamo na podatkovno zbirko emocionalnega govora, ki je bila pridobljena kot del interdicsiplinarnega projekta z imenom *"AvID: Audiovisual speaker identification and emotion detection for secure communications"* (Gajšek et al., 2009b).

Surov govorni signal posameznika, ki mu je pripisana primerna oznaka emocionalnega stanja, sam po sebi ni primeren za razpoznavanje emocionalnih stanj govorca. Raziskovalci se poslužujejo različnih postopkov in orodij, da bi izluščili iz surovega govornega signala kar se da veliko ključnih informacij za učenje razvrščevalnikov. To delo preizkuša dva postopka izpeljave značilk z namenom razvrščanja normalnega in vzbujenega emocionalnega stanja. Prvi se opira na linearno transformacijo MLLR (ang. Maximum Likelihood Linear Regression), drugi pa na uveljavljen skupek značilk, ki zajema akustične značilke govornega signala (frekvenca spremembe predznaka, energija RMS , osnovna frekvenca, azmerje zvenečenega proti nezvenečenemu delu govora in koeficienti MFCC) ter dodatne parametre pridobljene s pomočjo statistične analize (povprečna vrednost, standardna deviacija, ekstreme, koeficiente linearne regresije,itd.) (Schuller et al., 2009). Pri izpeljavi referenčnega sklopa značilk se avtorji opiramo na prosto dostopno orodje OpenSmile[1] (Eyben et al., 2009), ki omogoča preprost izračun ter zapis skupka značilk v vektor, ki predstavlja govornega signala za razpoznavanje emocio-

[1] http://sourceforge.net/projects/opensmile/

nalnih stanj govorca.

Preizkus predlaganih tipov vektorjev značilk omogoča le udejanjen razvrščevalnik emocionalnih stanj. Za to nalogo je primerno programsko okolje WEKA[2] (Hall et al., 2009), ki ponuja hitro učenje in preizkušanje več vrst postopkov razvrščanja.

2. Zbirka emocionalnega govora

Raziskovalci se velikokrat odločijo za pridobivanje nove zbirke podatkov, namenjenih za znanstvene raziskave ali določeno aplikativno uporabo. Več-modalna podatkovna zbirka AvID je bila izdelana z namenom uporabe govornih in slikovnih tehnologij pri video komunikacijskih sistemih. Njena uporaba je namenjena nalogam identifikacije ali verifikacije ter detekcije emocionalnih stanj udeležencev v pogovoru. Čeprav je zbirka AvID zasnovana kot več-modalna zbirka podatkov, saj vsebuje govorne in slikovne zapise, je možno obravnavati vsak del gradiva ločeno. V tem prispevku se osredotočamo le na emocionalno označen govorni del zbirke.

Na področju raziskovanja emocionalnih stanj iz govornega signala obstajata dve strategiji zajema posnetkov. Namen prve je snemanje govornih signalov z naravnimi (spontanimi) emocijami. Pri drugi strategiji se avtorji zbirk opirajo na vsebino posnetkov igranega govora. Ena izmed težjih nalog pri zbirkah spontanega emocionalnega govora v primerjavi z zbirkami igranih emocijami predstavlja označevanje govornih posnetkov. V primeru zbirke AvID so za pravilno označevanje emocionalnih stanj poskrbeli trije študenti psihologije. Čeprav so označevali več vrst emocionalnih stanj se avtorji tega prispevka osredotočamo le na vzbujeno (jeza, gnus, žalost, strah, veselje in presenečenje) in normalno emocionalno stanje. Končne oznake posameznih govornih odsekov govorcev so se pripisale avtomatsko, glede na večinsko število oznak vseh označevalcev.

Zbirka AvID ponuja dve vrsti snemalne seje. V tem prispevku smo se odločili analizirati le prvo vrsto seje. Uporabljen govorni material obsega sedemnajst govorcev, od tega 7 moškega spola in 10 ženskega spola. Razpolagali smo z označenimi posnetki v skupnem časovnem obsegu več kot šest ur.

Tabela 1: Število posnetkov in časovna dolžina posnetkov

Oznaka emo. stanja	št. posnetkov	čas [s]
Vzbujeno stanje	820	2867,56
Normalno stanje	7780	33571,54
Skupaj	8600	3639,10

Pogled na število posnetkov v tabeli 1 kaže na neenakomerno zastopanje posnetkov v obeh razredih. Tistih, ki so označeni kot ne-emocionalni govor, je skoraj 10 krat več, kar napeljuje uporabo tehnik, ki omogočajo po postopku izpeljave značilk umetno tvorjenje novih vzorcev značilk manj zastopanega razreda. Za prenos ključne informacije o emocionalnem stanju govorca s pomočjo transformacije MLLR potrebujemo za vsako oceno značilke dovolj govornega materiala. Ob tej priložnosti smo v postopku predobdelave združimo posnetke istega govorca z enakimi oznakami na skupno dolžino, ki je večja od petnajstih sekund. Združeni posnetki govorca v zbirki AvID predstavljajo v obeh postopkih izpeljave značilk vzorec iz katerega želimo prenesti ključno informacijo o emocionalnih stanjih posameznika.

[2]http://www.cs.waikato.ac.nz/ml/weka/

3. Izpeljava značilk

S stališča razpoznavanja vzorcev so bistvene lastnosti objektov tiste, ki poudarijo posebnosti posameznih razredov vzorcev. Take lastnosti objektov imenujemo značilke (Pavešić, 2000). Objektu razpoznavanja želimo prirediti množico značilk, katere naj bi opisale lastnosti posameznih emocionalnih stanj. Do zdaj se raziskovalci še niso poenotili kaj predstavlja na splošno dobro množico značilk za razpoznavanje emocionalnih stanj iz govornega signala.

V nadaljevanju predstavljamo izpeljavo značilk, ki temelji na linearni transformaciji MLLR, drugo izpeljavo pa temelji na uporabi sestavlja skupek akustičnih statističnih parametrov z nekaterimi govornega signala.

3.1. Izpeljava akustičnih značilk MLLR

Linearne transformacije prikritih Markovovih modelov so splošno uporabljene na področjih adaptacije modela na akustično okolje. Pri tem predpostavljamo, da se informacija o okolju ali govorca nahaja v parametrih linearne transformacije, ki se uporablja za prilagajanje parametrov akustičnega model ali značilk. Informacijo o okolju ali govorcu se nahaja v linearni transformaciji, ki se kasneje nahaja v kombinaciji s parametri akustičnega modela ali na značilkah. Obstaja več vrst tovrstnih linearnih transformacij, glede na različne možnosti vezave parametov celotnega govornega modela. V tem prispevku se osredotočamo na vezano linearno transformacijo, ki jo je mogoče preslikati iz transformacije parametrov modela v transformacijo prostora akustičnih govornih značilk (Gajšek et al., 2009a). Pri CMLLR-transformaciji (ang. constrained maksimum likelihood linear regression) se vektorji srednjih vrednosti μ in kovariančne matrike o linearno transformirajo po spodnjih enačbah.

$$\hat{\mu} = \acute{\mathbf{A}}\mu - \acute{\mathbf{b}}$$
$$\hat{\mathbf{o}} = \acute{\mathbf{A}}\mathbf{o}\acute{\mathbf{A}}^T$$

Matrika $\acute{\mathbf{A}}$ in vektor $\acute{\mathbf{b}}$ predstavljata parametre linearne transformacije in njune koeficiente določamo po kriteriju največjega verjetja akustičnega modela za dane nize vektorjev govornih akustičnih značilk $\mathbf{o}(\tau)$. Določanje koeficientov matrike $\acute{\mathbf{A}}$ in vektorja $\acute{\mathbf{b}}$ izvedemo s uveljavljenim postopkom EM (angl. Expectation-Maximization) (Gales, 1997). Dobljeno transformacijo parametrov Gaussovih porazdelitev lahko enostavno preslikamo v linearno transformacijo vektorjev akustičnih značilk, kot je podano v spodnjem izrazu.

$$\widehat{\mathbf{o}(\tau)} = \acute{\mathbf{A}}^{-1}\mathbf{o}(\tau) + \acute{\mathbf{A}}^{-1}\acute{\mathbf{b}} = \mathbf{A}\mathbf{o}(\tau) + \mathbf{b}$$

Matriko $\mathbf{A}$ in vektor $\mathbf{b}$ ponavadi združimo v matriko $\mathbf{W} = [\mathbf{A}\mathbf{b}]$, ki nato enovito predstavlja iskano linearno

transformacijo. Koeficiente matrike W določamo iz govornih posnetkov. S tem pridemo do linearnih transformacij $\mathbf{W}(s)$, ki vektorje akustičnih značilk prilagodijo akustičnemu govornemu modelu. Matriko $\mathbf{A}$ (dimenzij 39×39) raztegnemo v vektor (dimenzij 1×1521), kateremu prilepimo tudi odklon $\mathbf{b}$ (dimenzij 1×39). Tako končni vzorec govora z oznakama vzbujeno in normalno stanje predstavlja vektor značilk dimenzij 1×1560.

3.2. Izpeljava "state-of-the-art" sklopa značilk

Danes je na voljo kar nekaj programskih orodji, ki lajšajo delo raziskovalcem na področjih obdelave govornih signalov in razpoznavanja govora. Prosto dostopno orodje OpenSmile so avtorji zasnovali z namenom združevanja obstoječih algoritmov izpeljav značilk, namenjenih razpoznavanju emocionalnih stanj iz govornega signala v eno samo orodje. V tem prispevku smo na emocionalni zbirki govornih posnetkov AvID izpeljali značilke, ki so bile tudi preizkušene na tuji zbirki emocionalnega govora (Schuller et al., 2009).

Orodje izpelje iz govornega posnetka t.i. nizko-nivojske akustične značilke ter značilke na podlagi statistične analize posameznega govornega posnetka. V postopku izpeljave značilk orodje sestavi vektor značilk, ki ga opisuje 16 nizkonivojskih akustičnih značilk. To so frekvenca spremembe predznaka (ang. zero-crossing-rate) v časovni predstavitvi govornega signala, energije govornega signala, osnovna frekvenca (ang. pitch frequency) normirana na 500Hz, razmerje zvenečenega proti nezvenečenemu delu govora (ang. harmonic-to-noise ratio) ter prvih 12 koeficientov melodičnega frekvenčnega kepstra (MFCC). Vsakemu kepstralnemu keoficientu so dodani dinamični delta koeficienti. Predhodnemu vektorju značilk se dodajo še značilke, ki imajo temeljijo na statistični analizi akustičnega signala. To so na primer, povprečna vrednost, standardna deviacija, tretji in četrti standardiziran moment (ang. skewness, kurtosis) govornega signala, maksimalna in minimalna vrednost signala, relativna pozicija ekstrema v signalu in območje ekstrema kot tudi linearni regresijski koeficienti ter srednjo kvadratno napako linearne regresije. Orodje tako vsakemu govornemu posnetku priredi 384 razsežni vzorec z oznako vzbujeno ali normalno stanje.

4. Postopki predobdelave

Kot smo že nakazali v drugem poglavju, zbirka emocionalnega govora AvID vsebuje posnetke, ki pa žal niso enakomerno porazdeljeni med oba razreda razpoznavanja. Kot smo navedli v drugem poglavju je priporočena za dosego bolj robustne ločilne meje med normalnim in vzbujenim emocionalnim stanjem govorca uporaba postopkov, ki omogočajo bolj enakomerno porazdeljeni moči obeh razredov razpoznavanj. Ena izmed možnosti je ta, da za vsakega govorca posebej zanemarimo primerno število govornih posnetkov, ki so označeni kot normalno emocionalno stanje. Ta način žal okrni govorni material ter v postopku učenja razvrščevalnika pripomore k manj splošni ločilni meji. Da bi vseeno obdržali realne vzorce normalnega stanja se raje opremo na uveljavljen algoritem SMOTE (Chawla et al., 2002), ki omogoča sintetično tvorbo vektorjev značilk vzorcev manj zastopanega razreda.

Vzorcem značilk, ki pripadajo emocionalnem razredu, z uporabo postopka SMOTE, dodamo desetkratno število umetno tvorjenjih vzorcev. Postopek tvori umetni vzorec tako, da vsakemu vektorju značilk realnih vzorcev manj zastopanega razreda razpoznavanja poišče izbrano število najbližjih sosedov. Med njimi naključno izbere enega. Nato tvori razliko med vrednostmi vektorja značilk izbranega najbližjega sosednega vzorca ter vrednostmi značilk obravnavanega vzorca. Razlike so nato pomnožene z naključnimi vrednostmi med 0 in 1. Vsaka naključno prirejene razlika je prišteta k obravnavanemu vzorcu, kar tvori nov umetno tvorjenji vzorec. Postopek ponavlja toliko časa, dokler ne pridobi želenega števila umetnih vzorcev. Algoritem omogoča naključno izbiro točke v območju med dvema vrednostma značilke obravnavanega vzorca ter njegovega najbližjega soseda. Tako umetno tvorjenji vzorci omogočajo v postopkih učenja razvrščevalnikov določitev bolj splošne ločilne meje med vzorci različno zastopanih razredov vzorcev.

5. Rezultati

Za realizacijo samodejnega razvrščevalnika emocionalnih stanj iz govora smo uporabili odprto-kodno programsko okolje WEKA. Programsko okolje predstavlja zbirko algoritmov strojnega učenja ter algoritmov za manipulacijo nad podatki. Zgrajeno je z namenom hitrega preizkušanja obstoječih algoritmov na novih podatkih. Ponuja tudi podporo za pred obdelavo podatkov, namenjenih kasnejšim poizkusom strojnega učenja, torej uporabo različnih filtrirnih postopkov, statistično evalvacijo učnih shem ter celo vizualizacijo vhodnih podatkov in rezultatov.

Po uspešnem združevanju posnetkov istih govorcev z enakimi oznakami razredov razpoznavanja ter po uspešni izpeljavi vzorcev vektorjev značilk smo naključno razdelil vzorce na pet sklopov z namenom navzkrižnega preverjanja rezultata udejanjenega razvrščevalnika. V posamezni iteraciji učenja in preverjanja je posamezen sklop predstavljal testno množico, ostali štirje pa učno množico. Na ta način smo poskrbeli, da so bili podatki vedno razdeljeni v učno in testno množico z razmerjem vzorcev $4:1$. Kasneje smo s pomočjo programska okolja WEKA v učni množici obeh tipov značilk s postopkom SMOTE pridobili umetno tvorjenje vzorce vzbujenega emocionalnega stanja. Tako smo uspeli moč množice razreda vzorcev z oznako vzbujenega emocionalnega stanja približati moči razreda vzorcev normalnega emocionalnega stanja. Kvantitativno predstavitev števila učne množice s privzeto testno množico prikazuje tabela 2.

Tabela 2: Predstavitev št. vzorcev vektorjev značilk v posameznem sklopu

Oznaka emo. stanja	učna m.	testna m.
Vzbujeno stanje	1408	31
Normalno stanje	1480	128
Skupaj	2888	159

Izbira metode podpornih vektorjev za udejanjanje razvrščevalnike emocionalnih stanj izhaja iz dejstva, da

so vektorji značilk, ki vpisujejo govorni signal, visoke dimenzije. Metoda podpornih vektorjev omogoča v visoko dimenzionalnem poiskati prostoru poiskati hiperravnino, ki ločuje vzorce posameznih razredov razvrščanja. Tako udejanjen razvrščevalnik omogoča tudi posplošeno razvrščanje vzorcev, ki niso enaki vzorcem v učni množici razvrščevalnika, saj hiper-ravnino postavi tako, da je na obeh straneh razdalja do najbližjega vzorca največja (Boser et al., 1992). Čeprav naj bi metoda omogočala dobro ločevanje tudi med številčno neenakomerno porazdeljenimi razredi razpoznavanja, smo pri eksperimentih opazili boljše rezultate z uporabo postopka SMOTE.

Za oba tipa izpeljave značilk smo s pomočjo orodja WEKA in dodatne knjižnice LibSVM[3] (Chang and Lin, 2001) udejanjili več razvrščevalnikov na osnovi metode podpornih vektorjev. Preizkušali smo različna jedra (ang. kernels) ter ugotovili, da za značilke MLLR najbolje ustreza simodialno jedro, medtem ko za skupek akustičnih značilk najboljše rezultate pridobimo z polinomskim jedrom tretje stopnje.

Pri sistemih za razpoznavanje čustev se je uveljavil preklic R (ang. recall) kot mera za primerjavo delovanja različnih sistemov. Ta ponazarja razmerje med pravilno razpoznanimi vzorci znotraj razreda, proti vsoti vseh vzorcev iz istega razreda. Velikokrat se v rezultatih poleg priklica pojavljajo tudi druge mere, ki omogočajo podrobnejši vpogled pri tolmačenju rezultata. Ena izmed takih mer je tudi natančnost P (ang. precision), ki opisuje razmerje med pravilno razpoznanimi vzorci znotraj razreda, proti vsoti pravilno razpoznanih vzorcev razreda istega razreda ter razvrščenih vzorcev drugih razredov v obravnavan razred (Schuller et al., 2009). Zaradi neuravnoteženosti testne množice podatkov in v želji dobrega razvrščanja v obeh razredih razpoznavanja podajamo rezultat v obliki neuteženega povprečnega priklica (AR) in neutežene povprečne natančnosti (AP).

Tabela 3: Predstavitev rezultatov razpoznavanja

Tip izpeljave značilk	AR [%]	AP [%]
OpenSmile	70,42	60,21
MLLR	59,12	52,43

6. Zaključek

Pogled na tabelo 3 opredeli skupek "state-of-the-art" značilk kot boljši za namen razpoznavanja spontanih emocionalnih stanj. Udejanjen razvrščevalnik na podlagi metode podpornih vektorjev bolje ločuje med vzorci z oznakami normalno in vzbujeno emocionalno stanje, kot v primeru značilk MLLR. Neuteženi povprečni priklic se razlikuje za več kot 11%. Rezultat razvrščanja z uporabo "state-of-the-art" skupka značilk je primerljiv z razvrščanjem emocionalnih stanj tudi na drugi emocionalni zbirki podatkov, kas so pokazali Schuller, Steidl in Batliner Schuller et al. (2009). Posledično je iz rezultata moč sklepati, da je emocionalna zbirka govora AvID označena korektno, čeprav delo označevanja še ni popolnoma zaključeno. Želja avtorjev je, da bi pridobili dodatne oznake emocionalnih stanj še vsaj dveh dodatnih označevalcev.

Ne glede na rezultate se ideja izpeljave značilk MLLR za razpoznavanje emocionalnih stanj govorca zdi še kako smiselna, saj te značilke nosijo informacijo o premiku značilnosti govorca iz normalnega v neko emocionalno (vzbujeno) stanje. Avtorji verjamemo, da se v MLLR parametrih nahaja veliko več informacije kot na to kažejo rezultati, vendar sama matrika $\mathbf{A}$ ter njen pripadajoči vektor $\mathbf{b}$ v smislu značilk ne omogočata udejanjenim razvrščevalnikom poiskati najboljše ločilne meje. V prihodnosti si želimo poiskati način, ki bi iz matrike $\mathbf{A}$ in vektorja $\mathbf{b}$ omogočal izpeljavo boljše ocene med razdaljami transformacij in ne le nad posameznimi elementi matrike $\mathbf{W} = [\mathbf{A}\mathbf{b}]$, kot smo to pokazali v tem prispevku. informacijo o premiku

7. Literatura

Bernhard E. Boser, Isabelle M. Guyon, in Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. V: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, str. 144–152. ACM Press.

C. C. Chang in C. J. Lin. 2001. Libsvm: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, in W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

F. Eyben, M. Wollmer, in B. Schuller. 2009. Openear — introducing the munich open-source emotion and affect recognition toolkit. V: *Proc. 3rd Int. Conf. Affective Computing and Intelligent Interaction and Workshops ACII 2009*, str. 1–6.

R. Gajšek, V. Štruc, S. Dobrišek, in F. Mihelič. 2009a. Emotion recognition using linear transformations in combination with video. V: *Speech and intelligence: proceedings of Interspeech 2009*, str. 1967–1970, Brighton, UK, September.

R. Gajšek, V. Štruc, F. Mihelič, A. Podlesek, L. Komidar, G. Sočan, in B. Bajec. 2009b. Multi-modal emotional database: Avid. *Informatica (Ljubljana)*, 33(1):101–106.

M. J. F. Gales. 1997. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, in Ian H. Witten. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

N. Pavešić. 2000. *Razpoznavanje vzorcev : uvod v analizo in razumevanje vidnih in slušnih signalov*. Fakulteta za elektrotehniko, Ljubljana.

Björn Schuller, Stefan Steidl, in Anton Batliner. 2009. The INTERSPEECH 2009 Emotion Challenge. V: ISCA, ur., *Proceedings of Interspeech 2009*, str. 312–315.

[3]http://www.csie.ntu.edu.tw/ cjlin/libsvm/

Prenova sistema dialoga Kolos za projekt UVID

Peter Holozan

Amebis, d. o. o.
Bakovnik 3, 1241 Kamnik
peter.holozan@amebis.si

Povzetek

V sklopu projekta Univerzalni vmesnik inteligentnega doma je bilo treba narediti navideznega sogovornika, ki pomaga pri upravljanju televizijskega sprejemnika. V ta namen bil sistem dialoga Kolos dopolnjen z novim stavčnim analizatorjem, uporabo lem in pomenov ter naprednimi zunanjimi referencami. Razvit je bil tudi modul za preprosto bazo znanja, ki odgovarja na vprašanja s pomočjo dejstev, naučenih iz korpusa FidaPLUS.

Update of the Kolos dialog system in the UVID project

Within the Universal interface for the intelligent home project (UVID), a virtual assistant had to be created to help operating the television set. To this end, the Kolos dialog system was upgraded with a new sentence analyzer, the information about lemmas and word senses, and with advanced external reference information. From the facts derived from the FidaPLUS corpus we developed a simple knowledge base module used for the question answering system.

1 Uvod

V Amebisu smo že pred časom razvili sistem dialoga Kolos, krmiljen s programskim jezikom K (Romih, 2004) in uporabljen pri sistemu za klepetanje Klepec[1] (Arhar, Romih, 2006).

Razvoj tega sistema je od leta 2004 nekoliko zastal, zdaj pa smo se za potrebe projekta UVID (Univerzalni vmesnik inteligentnega doma) odločili, da temeljito prenovimo sistem, nadgradimo programski jezik K (z dosedanje verzije 2.0 na 2.5) in dodamo še nekatere specializirane module, potrebne pri projektu UVID.

Najprej bo na kratko opisan projekt UVID.

Sledil bo opis sprememb sistema Kolos in jezika K.

Na koncu bodo predstavljeni še specializirani moduli, rezultati uporabe in načrti za prihodnost.

2 Projekt UVID

Projekt Univerzalni vmesnik inteligentnega doma (UVID) je skupni projekt Špice International[2], Instituta »Jožef Stefan«, Iskratela in Amebisa. Projekt je sofinanciran v okviru Javnega razpisa za spodbujanje raziskovalno razvojnih projektov razvoja e-vsebin in e-storitev v letih 2009 in 2010. Cilj projekta je razviti inteligentne e-storitve za upravljanje zabavne elektronike in inteligentnega doma nasploh (Stanovnik, 2009). Projekt je osredotočen na napredno upravljanje televizijskega sprejemnika, in sicer prek NTB[3].

Amebis med drugim v okviru tega projekta razvija modul za krmiljenje televizijskega sprejemnika v naravnem jeziku in splošno podajanje informacij (bodisi povezanih s televizijo (spored) bodisi različne splošne informacije), podoben modul razvija tudi IJS. Amebisov navidezni sogovornik se imenuje Micka, od IJS je Zvonko, skupaj pa njune odgovore sestavlja Miki (in s tem imenom se oba tudi predstavljata). Na ta način se poveča verjetnost, da bo sistem znal odgovoriti na zahtevo, ker Micka in Zvonko delujeta na različna načina, seveda pa se v precejšnjem delu tudi prekrivata (pri čemer se izbere odgovor, ki ima označeno večjo zanesljivost).

Za zdaj je mišljeno, da se bo uporabnik pogovarjal z navideznim sogovornikom s pomočjo tipkanja ukazov oz. vprašanj, z vključitvijo dovolj dobre razpoznave govora pa bo možna tudi govorna komunikacija (kar bo naredilo ti funkcionalnost še bistveno uporabnejšo).

3 Prenova sistema dialoga Kolos

Glavna sprememba v sistemu dialoga Kolos je uporaba Amebisovega stavčnega analizatorja, ki je bil razvit za potrebe Presisa[4] in Besane[5]. Kolos je do zdaj imel vgrajen preprost lematizator, ki pa je v teh letih postal zastarel in smo ga zato nadomestili, kar je prineslo precej sprememb tudi v internem zapisu razčlembe vprašanj.

Novi analizator pozna občutno več besed kot stari, pri neznanih besedah pa zna tudi ugotavljati, ali gre morda za zatipkano besedo (vendar Kolos to upošteva le pri iskanju po lemah, ne pa po samih besedah).

3.1 Jezik K

Jezik K temelji na vzorcih, ki jih išče v danem vprašanju, in seznamu možnih odzivov.

```
$ (5)
kako ti je ime
kdo si
kaj si
kdo si ti
povej mi ime
>
Micka, ali na kratko Mehansko Inteligentni Celostno Komunikativni Avtomat
Micka, vaša pomočnica pri uporabi televizije
```

Slika 1. Primer elementa v jeziku K.

[1] http://klepec.amebis.si/

[2] Špica International je vodilni slovenski ponudnik informacijskih sistemov in rešitev za obvladovanje časa in prostora

[3] Net-top-Box, majhen računalnik na PC arhitekturi (x86, Linux), na katerem teče napredni televizijski vmesnik, ki ga razvija Iskratel

[4] Amebisov strojni prevajalnik, http://presis.amebis.si

[5] Amebisov slovnični pregledovalnik, http://besana.amebis.si

Da ni treba pisati čisto vseh možnih vprašanj, vsebuje jezik K še elemente, ki lahko nadomeščajo poljubno besedo ali poljubno število besed, spremenljivke, s katerimi si lahko navidezni sogovornik zapomni podatke iz dosedanjega razgovora, funkcije za dostop do zunanjih spiskov besed in baz ipd. Pri vzorcih je možno določiti tudi prioriteto, ki pove, kateri vzorec naj se prvi uporabi, kadar jih ustreza več.

3.2 Nadgradnja jezika K

Spodaj so naštete najpomembnejše novosti jezika K2.5.

3.2.1 Uporaba lem in pomenov

Že K2.0 je delno podpiral uporabo lem, vendar ni bilo mogoče ločevati med npr. pasti (padem) in pasti (pasem). V K2.5 smo zato uporabili oznake iz leksikalne podatkovne zbirke Ases (Arhar, Holozan, 2009), kjer so dvoumne leme dodatno opisane (npr. »pasti (padem)«, »učen/5učiti«, »téma«).

Leme so v K2.5 zapisane med dvojnimi zavitimi oklepaji, dodan pa je še delček MSD, ki pomaga pri enolični identifikaciji leme (npr. »{{klop+Soz*}}«, »{{klop+Som*}}«, »{{žarek+P*}}«).

Dodani so bili še pomeni, ki so zapisani med dvojnimi oglatimi oklepaji (npr. »[[peta (noga){0:0:0}]]«). Tudi pomeni so določeni v zbirki Ases.

Dodane so bile tudi spremenljivke tipa lema oz. pomen (prej so bile spremenljivke lahko le nizi ali števila).

3.2.2 Globalne spremenljivke

Jezik K2.0 je sicer predvideval globalne spremenljivke, ki bi bile prenosljive med več moduli, vendar ta del ni bil nikoli izveden.

V K2.5 globalne spremenljivke delujejo, dodatno pa je razdelano tudi določanje privzete vrednosti, kjer imajo neničelne vrednosti prednosti pred ničlami (oz. praznimi nizi), kar še olajša kombiniranje različnih modulov v več projektih, pri čemer se obnašanje modulov lahko tudi malenkost razlikuje (tako je npr. Klepec lahko veliko bolj zafrkljiv kot Micka, čeprav uporabita isti modul).

3.2.3 Prioriteta odgovora

Statično prioriteto, ki jo pozna že K2.0, smo uporabili za določanje zanesljivosti odgovora Micke. Zanesljivost potrebuje Miki, in sicer za to, da se laže odloči, ali bo uporabil odgovor Micke ali Zvonka.

Zanesljivost je označena tako, da je na koncu odgovora zaporedje ##1 do ##6, kjer je ##1 najmanjša zanesljivost odgovora (tipično gre za mašilo, to je odgovor, ki se vrne, kadar vprašanje ni prepoznano; gre za odgovore tipa »Tega pa res ne vem.«, »Tega žal ne razumem, ali lahko poveste to kako drugače.«), ##5 največja zanesljivost odgovora, ##6 pa je uporabljen za primere, ko gre za odziv na uporabnikov odgovor na prejšnje vprašanje, in mora Miki obvezno uporabiti ta odgovor.

statična prioriteta	zanesljivost odgovora
1, 2	##1
3, 4	##2
5	##3
6, 7	##4
8	##5
9	##6

Tabela 1: Preslikava statične prioritete v zanesljivost.

3.2.4 Napredne zunanje reference

V K2.0 lahko z referencami v vzorcu zapišemo določeno množico besed, ki so npr. seznam besed v besedilni datoteki. Omejitev v K2.0 je bila, da so reference lahko le enobesedne, kar se je pokazalo kot zelo omejujoče, saj se ni dalo narediti niti seznama krajevnih imen. Druga pomanjkljivost je bila, da je bil rezultat reference lahko le najdeno/ni najdeno, kar je pomenilo, da se tega, kar je bilo poiskano, ni dalo enostavno uporabiti v rezultatu.

V K2.5 je bilo zato uvedeno, da lahko referenca uporabi poljubno število besed, hkrati pa ima tudi dostop do razčlembe besedila. Hkrati referenca vrne tudi besedilni rezultat, ki se lahko potem uporabi pri klicih funkcij (ali enostavno izpiše z ustrezno funkcijo).

Reference iz besedilnih datotek so bile razširjene še s podporo za uporabo lem in z več stolpci, kjer je mogoče kot parameter povedati, po katerem stolpcu so išče in v katerem stolpcu je rezultat, tako da je mogoče elegantno narediti šifrante in rezultat potem uporabiti pri klicu funkcije.

```
Slovenijo 1     SLO1
Slovenijo1      SLO1
SLO1            SLO1
SLO 1           SLO1
prvi program    SLO1
POP TV          POPTV
POPTV           POPTV
pop             POPTV
```

Slika 2. Delček šifranta z imeni programov.

Če je šifrant v taki obliki v datoteki »ref\prg.ref«, se lahko uporabi v K2.5 takole:

```
$
preklopi na @PreveriREF(__pot, »ref\\prg.ref«, 1, 2)
daj na @PreveriREF(__pot, »ref\\ prg.ref«, 1, 2)
daj @PreveriREF(__pot, »ref\\ prg.ref«, 1, 2)
hočem @PreveriREF(__pot, »ref\\ prg.ref«,1, 2)
prestavi na @PreveriREF(__pot, »ref\\ prg.ref«, 1,  2)
vklopi @PreveriREF(__pot, »ref\\ prg.ref«, 1, 2)
>
Kanal preklopljen na @Rezultat(#1).
\@\@tv\:command\=SetChannel\(@Rezultat(#1)\)
```

Slika 3. Uporaba šifranta.

Z uporabo referenc je tako lahko nadomeščeno veliko število vzorcev, hkrati pa je tudi olajšano dodajanje novih elementov, saj jih je treba dodati le na enem mestu (isti šifrant programov lahko uporabimo tako pri preklapljanju programov kot pri iskanju po sporedu).

Reference se v K2.5 tudi naložijo v pomnilnik ob prvi uporabi, tako da delujejo veliko hitreje kot prej, ko jih je program vsakič sproti bral z diska.

3.3 Preprosta baza znanja

Za hitro razširitev količine vprašanj, na katera zna odgovoriti Micka, smo uporabili zajemanje dejstev iz besedil, in sicer za Micko iz Fide+. Ideja je bila, da se uporabijo enostavčne povedi, ki jih uspe analizator uspešno razčleniti in določiti pomene, ter da zna Micka potem odgovoriti na vsa vprašanja v zvezi s tem stavkom.

To je hkrati tudi prvi Amebisov poskus na področju odgovarjanja na vprašanja (QA[6]) za slovenščino. To je področje računalniške lingvistike, ki se ukvarja s tem, kako dobiti pravilen odgovor na vprašanje, napisano v naravnem jeziku, iz zbirke besedil (Ledeneva, Sidorov, 2010).

Če bi bil stavek npr. »Irena je maja 2010 posadila rdečo ciklamo.«, bi morala Micka iz tega razbrati odgovore na naslednja vprašanja:

```
Kdo je posadil rdečo ciklamo maja 2010?
Kdaj je Irena posadila rdečo ciklamo?
Kaj je Irena posadila maja 2010?
Kaj je Irena posadila maja?
Kaj je Irena posadila 2010?
Kaj je posadila Irena?
Kakšno ciklamo je posadila Irena?
Kaj je bilo posajeno maja 2010?
```

Slika 4. Seznam možnih vprašanj.

Možnih variant vprašanj pa je seveda še več. Tako »lepih« (z enostavno enostavčno strukturo, po možnosti še brez zaimkov in odvisnosti od sobesedila) povedi sicer v realnih besedilih ni veliko, vendar se jih v veliki količini besedil v korpusu vseeno najde kar nekaj, tako da je bila domneva, da bo Micka iz tega znala odgovoriti na marsikatero splošno vprašanje.

Druga vrsta vprašanja, izhajajoča iz zgornjega primera, bi bila »Ali je Irena maja 2010 posadila rdečo ciklamo?«. Na prvi pogled je zavajajoče videti, da bi lahko na taka vprašanja Micka odgovarjala z »da« ali »ne«, vendar v resnici lahko dogovori le z »da« (oziroma »ne«, kadar je izvorni stavek zanikan), v nasprotnem primeru pa mora biti odgovor »ne vem« oz. ne sme dati nobenega odgovora.

Razširitev iskanja bi lahko bila še uporaba nadpomenk iz Asesa, s čimer bi program znal odgovoriti tudi na vprašanje »Kdo je posadil rastlino?« Pred to razširitvijo bo treba narediti novo verzijo izpeljane baze iz podatkovne baze Ases. Neposredna uporaba Asesa je namreč nekoliko počasna in baza je tudi precej obsežna, zato se za posamezne izdelke (Presis, Besana) naredi optimizirana izpeljana baza, ki je hitrejša in vsebuje le potrebne elemente. Nadpomenke oz. podpomenke so sicer implicitno uporabljene pri izdelavi baze za Presis, eksplicitno pa baza nima podatkov o njih in bo to treba dodati.

3.3.1 Zapis znanja

Baza znanja izhaja iz vmesnega jezika, ki ga naredi analizator. Vmesni jezik vsebuje vse potrebne slovnične podatke (stavčno analizo), besede so tudi razdvoumljene (ločeno vprašanje je seveda uspešnost razdvoumljanja), načeloma omogoča celo to, da so vprašanja v drugem jeziku, kot je vhodno besedilo (Cardeñosa, Gallardo, De la Villa, 2009), torej CLIR[7], vendar tega še nismo preizkusili (težava je, da dobijo pri sedanji izvedbi izpeljane baze pomeni (in tudi leme in oblikoskladenjske oznake) vsakič drugačne kode, tako da bazi za različna jezika nista primerljivi; morali bi razviti verzijo baze, ki bi hkrati podpirala več vhodnih jezikov ali pa uvesti stalne kode).

Stavek razbije na naslednje *delce* (stavčne člene): povedek (pomen glagola (glagolske predloge[8]) in dodatni podatki o glagolskem naklonu, času, dovršnosti in trdilnosti), elementi glagolske predloge (osebek, predmeti, lahko tudi predložni deli ali v nekaterih primerih prislovna določila), prislovna določila in členki (vezani na glagol). Kadar vsebuje stavek modalni glagol, je povedek sestavljen iz obeh glagolov, elementi glagolske predloge pa imajo dodan podatek, na kateri glagol se nanašajo.

Za primer »Irena je maja 2010 posadila rdečo ciklamo.« je zapis v vmesnem jeziku naslednji:

```
(-POV:(-STAg-npppdvt-------:(1OSB:(-SFR:(-DSF:(-
JED:(-SAMe:{2353ea;f146ec}[0]<dc0>))))),(*PVD:(-
GPO:[1])),(-PDOc:(-XLEa:{a,---,be,a-2010}[2,3])),
(0PVD:(-GGL:{4db851;3105c4}[4]<53ac>)),(2PR4:(-
SFR:(-DSF:(-PFR:(-DPF:(-
PRVo:{39919;16c9fd}[5]<528>))),(-JED:(-
SAMe:{4adc9;4b6dc3}[6]<50>))))),(-LOCkp:[7])))
```

Slika 5. Zapis stavka v vmesnem jeziku.

Vmesni jezik vsebuje nekatere podatke, ki so za potrebe zapisa znanja odvečni (oz. celo moteči), kot so kode lem besed, kode oblikoskladenjskih oznak besed in položaji besed v stavku. Zato vmesni jezik očistimo teh podatkov:

```
(-POV:(-STAg-npppdvt-------:(1OSB:(-SFR:(-DSF:(-
JED:(-SAMe:{;f146ec}[]<>))))),
(*PVD:(-GPO:[])),(-PDOc:(-XLEa:{a,---,be,a-
2010}[])),(0PVD:(-GGL:{;3105c4}[]<>)),(2PR4:(-SFR:(-
DSF:(-PFR:(-DPF:(-PRVo:{;16c9fd}[]<>))),(-JED:(-
SAMe:{;4b6dc3}[]<>))))),(-LOCkp:[])))
```

Slika 6. Prečiščeni zapis stavka v vmesnem jeziku.

Delci, ki jih ta primer vsebuje, so naslednji:

```
ppdt+3105c4
(-SFR:(-DSF:(-JED:(-SAMe:{;f146ec}[]<>))))
(-PDOc:(-XLEa:{a,---,be,a-2010}[]))
(-SFR:(-DSF:(-PFR:(-DPF:(-PRVo:{;16c9fd}[]<>))),(-
JED:(-SAMe:{;4b6dc3}[]<>))))

dodatni preprosti delec:
(-SFR:(-DSF:(-JED:(-SAMe:{;4b6dc3}[]<>))))
```

Slika 7. Najdeni delci.

[6] Question Answering

[7] Cross-language information retrieval

[8] glagolska predloga je element v Asesu, ki pove, na kakšen način se glagol uporablja v stavku, tj. s katerimi skloni predmetov in s katerimi predlogi se tipično lahko veže ipd. (primeri so npr. »[posaditi] PR4 KAM« , »[poslati] {PR3} PR4 KAM«, »[zaljubiti] se {v PR4}«, »imeti krompir«)

Vsak stavek, ki se ga da dodati v bazo, postane *dejstvo*. Stavek se tudi sam doda v bazo in je možen rezultat iskanja (ob posameznem delcu, kjer pa se rezultat naredi s pomočjo generatorja iz vmesnega jezika).

Na koncu se dodajo vse povezave med delci in dejstvom. Pri sestavljenih delcih (rdeča ciklama) se dodajo še preprosti delci (ciklama) in prav tako povežejo na dejstvo (s podatkom, da gre za drugotno povezavo), s čimer se poenostavi iskanje na račun večje baze. Enako je predvideno tudi za nadpomenke, kar bo sicer precej povečalo bazo, vendar so ti sistemi dialoga predvideni bolj za strežniško uporabo, zaradi česar velike podatkovne baze niso taka težava, kot bi to bilo pri distribuciji do posamičnih uporabnikov (pri strežniških sistemih, ki jih lahko hkrati uporablja veliko število ljudi, je zelo pomembna tudi hitrost delovanja).

Bazo smo izvedli z uporabo podatkovne baze SQL.

Pri iskanju vprašanja prevedemo v vmesni jezik, ga očistimo odvečnih podatkov in spet razbijemo na delce. Za vprašanje »Kdo je posadil rdečo ciklamo maja 2010?« tako dobimo naslednji rezultat:

```
ppdt+3105c4
(-SFR:(-DSF:(-JED:(-VPRse:{;26831a}[]<>))))
(-PDOc:(-XLEa:{a,---,be,a-2010}[]))
(-SFR:(-DSF:(-PFR:(-DPF:(-PRVo:{;16c9fd}[]<>))),(-
JED:(-SAMe:{;4b6dc3}[]<>))))
```

Slika 8. Najdeni delci za vprašanje.

Osebek vsebuje vprašalnico (element VPR, zgoraj odebejen), torej je to iskani element, zato poiščemo dejstvo, ki vsebuje vse preostale delce iz vprašanja, in delec, ki se pri tem dejstvu pojavlja na mestu osebka.

3.3.2 Luščenje znanja iz korpusa Fida+

Že pri zadnjem preizkusnem označevanju korpusa Fida+ (v okviru razvoja izboljšane in razširjene verzije korpusa, ki se bo imenovala KOS) smo kot stranski izdelek pripravili seznam vseh enostavčnih povedi (žal so se v seznamu znašle tudi neglagolske povedi, ki jih je bilo treba dodatno izločati), ki smo jih uspešno analizirali, kar je osnovni pogoj za zajem v bazo znanja. Na ta način programu za zajemanje ni treba analizirati celotnega besedila Fide+, ker bi vzelo zelo veliko procesorskega časa (v mesecih na posameznem računalniku), ampak ima za vhod bolj obvladljivo množico.

Kljub tej predhodni obdelavi traja zajem kar nekaj časa (okoli 30 dni), saj je v vhodnem spisku dobrih 10 milijonov povedi (v vsem korpusu je sicer dobrih 45 milijonov povedi). Od tega števila jih je program uspel zajeti dobrega 3,5 milijona, pri drugih pa kaj ni bilo v redu (ni bil glagolski stavek ali pa katera od besed ni imela določenega nobenega pomena).

4 Specializirani moduli

Poleg izboljšave nekaterih splošnih modulov, ki so bili že del Klepca (npr. sklanjanje besed, vprašanja o trenutnem času, zemljepisna vprašanja (položaji krajev v Sloveniji, razdalje med njimi ipd.)), smo posebej za Micko razvili specializirane module za projekt UVID.

4.1 Krmiljenje televizijskega sprejemnika

Vgradili smo podporo za prestavljanje programov (z uporabo naprednih zunanjih referenc) in nastavljanje glasnosti televizijskega sprejemnika. Ideja za prihodnost je še prižiganje/ugašanje televizorja, vendar te funkcije za zdaj NTB še ne podpira.

Glasneje!
Preklopi na POP TV!
Tiše!
Prestavi na Slovenijo 2!

Slika 9. Primeri ukazov za krmiljenje televizorja.

Podpora tem ukazom sama po sebi v tem trenutku ni posebno uporabna — uporaba običajnega daljinca je mnogo preprostejša — vendar je kljub temu dodana zaradi popolnosti, polna uporabnost pa se bo pokazala v primeru, če bo v prihodnosti v inteligentni dom dodana tudi razpoznava govora, s čimer bodo ti ukazi res postali uporabni.

Miki komunicira z NTB preko protokola HTTP, ukazi pa temeljijo na standardnem XBMC HTTP API[9].

4.2 Uporaba sporeda

NTB na zahtevo vrne podatke o sporedu v formatu XML EPG[10]. Dodatno smo se odločili podpreti še videoteko, za kar smo uporabili kar razširjeno referenčno datoteko.

koliko stane film iz videoteke
Odvisno od filma, večina med 20 in 25 žetonov, za kateri film pa vas zanima?##3
vojna svetov
Ta film stane 20 žetonov. Ga želite takoj pogledati?##6
da
Film Vojna svetov se bo takoj začel predvajati.##6

Slika 10. Primer dialoga o uporabi videoteke.

S pomočjo sporeda pa Micka odgovarja na vprašanja tipa: »Ali bo danes zvečer na sporedu kakšen film?«, kjer preišče sporede in pogleda, ali kaj ustreza pogoju, nato pa izpiše, na katerih programih bodo kateri filmi. Micka odgovarja na vnaprej določen seznam časovnih prislovnih določil (danes/jutri zvečer/popoldne, v petek zvečer ipd.).

bo danes popoldne kakšen nogomet
Oddaja SP v nogometu: Grčija - Argentina bo na SLO2 23. 06. ob 12:35.
Oddaja SP v nogometu: Slovenija - Anglija bo na SLO2 23. 06. ob 15:50.##4
bo jutri zvečer kakšen film
Oddaja Voditelj vrača udarec bo na AKANAL 24. 06. ob 20:00.
Oddaja Lov na zaklad bo na TV3 24. 06. ob 21:55.
Oddaja Sadamova hiša bo na HBO 24. 06. ob 19:00.##4

Slika 11. Primeri vprašanj o sporedu.

[9] http://wiki.xbmc.org/?title=WebServerHTTP-API
[10] electronic programming guide

5 Rezultati

Vse zgoraj opisane izboljšave so bile uspešno uvedene v sistem Klepec, Micka napisana v jeziku K2.5. Vse skupaj je bilo uspešno integrirano v prototip, ki je bil že predstavljen predstavnikom Ministrstva za visoko šolstvo, znanost in tehnologijo v okviru projekta UVID.

Nekateri primeri odzivov Micke so napisani že zgoraj, zato bodo na tem mestu bolj podrobno predstavljeni le rezultati baze znanja.

5.1 Rezultati baze znanja iz Fide+

Spodaj je nekaj primerov odgovorov baze znanje na zastavljena vprašanja. Kadar je več možnih odgovorov, je naključno izbran od njih (s čim daljšim iskanim delcem), tako da isto vprašanje običajno ne bo dalo istega odgovora v prvem poskusu.

vprašanje	odgovor
Kam gremo?	Gremo spet v brlog!
Kje je gorelo?	Gorelo je na dveh hektarjih zemlje.
Do česa prihaja?	S tem pa nastajajo nove težave in zapleti.
Kaj so napadli?	Leta 865 so neuspešno napadli Bizanc.
Kdo je bil izvoljen?	Za novega predsednika so izvolili predsednika ljubljanske vlade Marjana Vidmarja.
Kaj je bilo popravljeno?	Umetno jamo so popravili z novimi tehnikami.
Kdo se je zaletel?	V križišču sta trčila
Česa me je strah?	Pa saj jaz se ničesar ne bojim!
Koga poznam?	Delo babic v porodnem bloku poznam.
Kako ga je ubil?	Ubil ga je iz strahu.
Kje dežuje?	Pada v podzemni dvorani.
Kdo je pekel? Kaj je pekel?	Ivan Kastelic iz Novega mesta je pekel izredno dober ržen kruh. Odraščanje je pekel.

Tabela 2: Primeri odgovor iz baze znanje.

Na začetku so primeri, kjer so odgovori ustrezni (ustreznost v tem primeru pomeni, da najdeni stavek res lahko odgovor na zastavljeno vprašanje, čeprav ni nujno, da je to res pravi in najustreznejši odgovor). Pri dveh primerih (»Kdo je bil izvoljen?« in »Kaj je bilo pripravljeno?«) se vidi, kako analizator uspešno prehaja med tvornim in trpnim načinom. Naslednji primer (»Kdo se je zaletel?«) pa kaže, kako program najde rezultat, kadar je na isti pomen vezanih več sinonimov.

Zadnji štirje primeri prikazujejo različne tipe napačnih odgovorov. Pri primeru »Česa me je strah?« je težava v tem, da ne bi smel najti zanikanega glagola. Ker pa v vmesnem jeziku (po vzoru angleščine) ni dvojnega zanikanja, je v vmesnem jeziku ta trditev zapisa kot trdilna in jo baza znanja tako obravnava.

Primer »Koga poznam?« kaže, da vmesni jezik ne ločuje med osebami in drugimi pomeni in tako ne more ugotoviti, kateri pomeni ustrezajo kateri vrsti vprašalnih zaimkov.

Naslednji primer kaže na napačno analizirano besedilo. Prislovno določilo »iz strahu« je v analizi označeno kot prislovno določilo načina namesto vzroka. To težavo bomo odpravili s popravkom v Asesu, kjer bomo določili verjetnejšo kombinacijo predložnega in samostalniškega pomena.

Tudi predzadnji primer je napačno razdvoumljen, iz prejšnje povedi v korpusu se da ugotoviti, da se stavek nanaša na slap v podzemni jami.

Pri zadnjem primeru je težava to, da je stavek pravilno razdvoumljen, vprašanje pa je tako dvoumno, da program ne more ugotoviti pravega pomena brez sobesedila. In samo iz prejšnjega vprašanja se vidi, da je spraševalec najbrž imel v mislih pečenje, ne pekla.

6 Kako naprej

6.1 Spored

Uporabo sporeda bi bilo smiselno razširiti še s podatki o igralcih, režiserjih, dobitnikih filmskih nagrad, žanrih ipd., kar bi lahko dobili s pomočjo spletne podatkovne baze IMDb[11], precejšen del teh informacij pa zbirajo tudi že sami ponudniki digitalne televizije.

Tako bi lahko Micko vprašali, kdaj bo na sporedu film, v katerem igra določen igralec ali ga je režiral določen režiser. Taka povpraševanja bi si Micka lahko zapomnila in v primeru, ko bi bil kdaj drugič na sporedu kakšen film, ki bi ustrezal temu povpraševanju, posebej opozorila nanj.

6.2 Baza znanja

Za realni preizkus sem poiskal na spletni strani uciteljska.net nalogo iz bralnega razumevanja za 3. oz. 4. razred osnovne šole. Navodilo je, da učenec prebere besedilo in odgovori na vprašanja o besedilu. Besedilo naloge je naslednje:

NOVICA: SLON JE POBEGNIL

V TOREK JE IZ ŽIVALSKEGA VRTA V LJUBLJANI POBEGNIL SLON. IME MU JE JAKA. ODŠEL JE SKOZI VRATA, KI SO JIH POZABILI ZAPRETI. SPREHAJAL SE JE KAR PO CESTI. PRESTRAŠIL JE MNOGO LJUDI, KI SO POKLICALI POLICIJO. POLICIJA JE SLONA VARNO VRNILA V ŽIVALSKI VRT.

Slika 12. Besedilo naloge.

Besedilu sledi seznam vprašanj, ki v dani nalogi sicer podana v obliki kviza (učenec izbira med odgovori a, b in c), za preizkus baze znanja, naučene iz zgornjega besedila, pa sem uporabil le vprašanja:

1. Kaj je novica?
2. Kakšen je bil naslov novice, ki si jo poslušal?
3. Kako je bilo slonu ime?
4. Od kod je slon pobegnil?
5. Kako je slon pobegnil?
6. Kje se je slon sprehajal?
7. Kdo je slona vrnil nazaj v živalski vrt?

Slika 13. Vprašanja naloge.

[11] http://www.imdb.com/

Rezultat je bil, da program ni znal odgovoriti niti na eno od postavljenih vprašanj, kar kaže, da bo pred uporabo baze znanja v praksi treba narediti še marsikaj.

Analiza je pokazala naslednje razloge za neuspeh programa:

- pri 1. in 2. vprašanju vprašanje ni neposredno povezano z besedilom novice;
- pri 3. in 6. vprašanju je v stavkih, kjer je odgovor, osebni zaimek namesto besede slon;
- pri 3. vprašanju je vprašanje tudi v pretekliku, odgovor pa v stavku v sedanjiku;
- na 5. vprašanje ni neposrednega odgovora v besedilu;
- pri 7. vprašanju ni bilo odgovora, ker v besedilu piše, da je policija slona vrnila, vprašanje pa je, kdo ga je vrnil nazaj (če bilo obratno, torej »vrniti nazaj« v besedilu in le »vrniti« v vprašanju, bi program uspešno našel odgovor, vprašanja pa ne zna posplošiti);
- pri 4. vprašanju je prišlo do razlike pri razdvoumljanju glagola »pobegniti« med besedilom in vprašanjem; besedilo je bilo razdvoumljeno z »P:G:pobegniti iz PR2/NIOS KAM«, vprašanje pa z »P:G:uiti {PR3/OS} KAM«.

Kako torej nadgraditi bazo znanja, da bo znala odgovoriti na več vprašanj:

Prva nadgradnja baze znanja bo upoštevanje nadpomenk iz Asesa, kar bo precej povečalo število odgovorjenih vprašanj. Ob tem bi bilo morda smiselno upoštevati tudi to, da so osebni zaimki lahko nadpomenka pri osebah.

Druga nadgradnja, ki pa bo zahtevala temeljito dopolnitev analizatorja, bo razreševanje osebnih in kazalnih zaimkov (kar bo prišlo prav tudi pri Presisu, ker se kdaj spol zaimka v prevodu spremeni, česar Presis še ne zna upoštevati; poznavanje izvorne besede bo koristno tudi pri razdvoumljanju). Vse prevečkrat je rezultat povpraševanja v bazi znanja namreč osebni zaimek, kar brez sobesedila žal ne prinese dovolj informacije. Kot začasna rešitev se za odgovor uporabijo čim bolj zapleteni delci (kadar je več možnih odgovorov), kar zmanjša verjetnost, da bi bili odgovori osebni zaimki.

Tretja nadgradnja bi lahko bila časovna orientacija, da torej analizator ugotovi, na kateri datum se nanaša jutri, danes, včeraj ipd., ter na vprašanja odgovarja s perspektive današnjega dne. Če je tako npr. časopisna novica, da je bil včeraj v Mehiki potres, lahko iz datuma, kdaj je bila članek objavljen, sklepa, katerega dne je bil v resnici potres (vprašanje, ali je bil v Mehiki včeraj potres, čez eno leto v resnici nima istega odgovora kot na dan objave članka).

Še ena dopolnitev bi bila upoštevanje živosti pri odgovarjanju na vprašalne zaimke kdo/kaj. Ases sicer pri veliki večini pomenov ima podatek, ali gre za osebo, vendar to za zdaj ni zapisano v vmesnem jeziku, tako da iskanje tega ne zna uporabiti.

Veliko pa bo treba narediti tudi na kakovosti samega analizatorja (predvsem pri razdvoumljanju) in baze v Asesu (največ pri glagolskih predlogah; tako bo npr. treba dodati predlogo za »vrniti nazaj« in jo povezati na isti pomen kot »vrniti« (spotoma se lahko označi še, da uporaba »vrniti nazaj« slogovno ni najboljša in bi lahko Besana priporočala izbris besede »nazaj«)).

Najpomembnejša nadgradnja pa bo seveda razširitev na uporabo večstavčnih povedi, česar se do zdaj nismo dotaknili.

Vendar v analizator šele uvajamo višjo raven vmesnega jezika, ki bo opisala odnose med stavki (oz. polstavki in pastavki) v povedi, tako da tega za zdaj še ni na voljo.

In tudi sicer je smiselno najprej izboljšati bazo znanja iz enostavčnih povedi z zgoraj opisanimi nadgradnjami, šele potem pa pride na vrsto nov zapis znanja, ki bo primeren tudi za večstavčne povedi.

6.3 Nadaljnji razvoj jezika K

Nadaljnji razvoj jezika bo najbrž šel v smer uporabe vmesnega jezika pri opisu vzorcev in odgovorov, vendar bo kot temelj treba prej še izboljšati analizator in dodati analizo povedi. Smiselno je tudi, da bo uporaba novih funkcij le dopolnilo, saj vprašanja/zahteve uporabnikov v splošnem niso pisane slovnično pravilno, kar naredi analizatorju precej težav.

7 Sklep

S temeljito prenovo sistema dialoga Kolos je bilo mogoče za potrebe projekta Uvid vnesti odzive na marsikatere uporabnikove zahteve ali vprašanja, ki prej niso bili mogoči ali pa bi jih bilo zelo zapleteno napisati v jeziku K.

S tem se je odprla pot, da se sistem dialoga Kolos uporabi še v drugih projektih, za začetek pa predvidoma še prenovo sistema Klepec.

8 Zahvala

Prenova, opisana v članku, je bila izvedena v okviru projekta Univerzalni vmesnik inteligentnega doma, ki ga je sofinanciralo Ministrstvo za visoko šolstvo, znanost in tehnologijo v okviru Javnega razpisa za spodbujanje raziskovalno razvojnih projektov razvoja e-vsebin in e-storitev v letih 2009 in 2010.

9 Literatura

Arhar, Š., Holozan, P., 2009. ASES – leksikalna podatkovna zbirka za razvoj slovenskih jezikovnih tehnologij. V V. Mikolič (ur.), *Jezikovni korpusi v medkulturni komunikaciji*. Koper: Založba Annales.

Arhar, Š., Romih, M., 2006. Klepec: programirani sogovornik za slovenščino. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik 8. mednarodne multikonferece Informacijska družba IS 2006, Zvezek B, Jezikovne tehnologije*. Ljubljana: IJS.

Cardeñosa, J., Gallardo, C., De la Villa, M. A., 2009. *Interlingual Information Extraction as a Solution for Multilingual QA Systems*. *FQAS*, 500-511.

Ledeneva, Y., Sidorov, G., 2010. Recent Advances in Computational Linguistics. *Informatica*, 34:3-18.

Romih, M., 2004. Uporaba programskega jezika K2.0 v sistemih dialoga. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik 7. mednarodne multikonferece Informacijska družba IS 2004, Zvezek B, Jezikovne tehnologije*. Ljubljana: IJS.

Stanovnik, T., 2009. *Prijava projekta Univerzalni vmesnik inteligentnega doma*, Razpisni obrazec 3 – OPIS PROJEKTA.

Jezikovni viri projekta JOS

Tomaž Erjavec,[1] Darja Fišer,[2] Simon Krek,[1] Nina Ledinek[3]

[1] Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si, simon.krek@ijs.si
[2] Oddelek za prevajalstvo, Filozofska fakulteta univerze v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
darja.fiser@guest.arnes.si
[3] Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU
Novi trg 4, 1000 Ljubljana
NLedinek@zrc-sazu.si

Povzetek

Namen jezikovnih virov JOS je spodbuditi razvoj jezikovnih tehnologij in korpusnega jezikoslovja za slovenski jezik. Vire JOS sestavljajo oblikoskladenjske specifikacije, ki definirajo oblikoskladenjske lastnosti in nabor oznak za slovenščino, dva korpusa (jos100k in jos1M) ter dva spletna servisa (konkordančnik in orodje za oblikoskladenjsko označevanje besedil). V prispevku predstavimo posamezne vire, s poudarkom na jos100k, ki je enojezični vzorčeni in uravnoteženi korpus slovenskega jezika s 100.000 besedami in z ročno označenimi oz. pregledanimi oznakami za tri nivoje jezikoslovnega opisa. Na ravni oblikoskladnje je vsaki besedi pripisana njena oblikoskladenjska oznaka in lema. Na ravni skladnje so stavki v korpusu označeni z odvisnostnimi površinskoskladenjskimi povezavami med pojavnicami. Na pomenski ravni so vse pojavitve 100 najbolj pogostih samostalnikov v korpusu označene z njihovim pomenom glede na slovenski semantični leksikon sloWNet. Jezikovni viri JOS so označeni po mednarodnih standardih in priporočilih, oblikoskladenjske specifikacije skladno s sistemom MULTEXT-East V4, tako specifikacije kot korpusa pa skladno z navodili Text Encoding Initiative Guidelines, TEI P5. Vsi viri so dostopni za raziskovalne namene po licenci Creative Commons na naslovu http://nl.ijs.si/jos/.

The Language Resources of the JOS Project

The JOS language resources are meant to facilitate development of human language technologies and corpus linguistics for the Slovene language and consist of the morphosyntactic specifications, defining the Slovene morphosyntactic features and tagset; two annotated corpora (jos100k and jos1M); and two web services (a concordancer and text annotation tool). The paper introduces these components with a focus on jos100k, a 100,000 word sampled balanced monolingual Slovene corpus, manually annotated for three levels of linguistic description. On the morphosyntactic level, each word is annotated with its morphosyntactic description and lemma; on the syntactic level the sentences are annotated with dependency links; on the semantic level, all the occurrences of 100 top nouns in the corpus are annotated with their wordnet synset from the Slovene semantic lexicon sloWNet. The JOS corpora and specifications have a standardised encoding (Text Encoding Initiative Guidelines TEI P5) and are available for research from http://nl.ijs.si/jos/ under the Creative Commons licence.

1. Uvod

Jezikoslovno označeni korpusi predstavljajo osnovo za jezikovne tehnologije in korpusno jezikoslovje, vendar za mnoge jezike še vedno niso na voljo, še posebej v obliki zaključenih podatkovnih zbirk. Med njimi so temeljni vir predvsem ročno preverjeni besednovrstno oz. oblikoskladenjsko označeni korpusi, odvisnostne drevesnice in pomensko označeni korpusi.

V projektu »Jezikoslovno označevanje slovenščine«, JOS, katerega cilj je bil zapolniti praznino na tem področju za slovenski jezik, so bile izdelane oblikoskladenjske specifikacije, ki definirajo oblikoskladenjske lastnosti in nabor oznak za slovenščino in dva prosto dostopna ročno preverjena korpusa s standardiziranimi oznakami, med katerima je manjši označen na treh ravneh jezikoslovnega označevanja. O prvem koraku označevanja, pri katerem sta bila korpusa izdelana in oblikoskladenjsko označena, smo že poročali (Erjavec in Krek, 2008), tokrat pa poročamo o končnem rezultatu oblikoskladenjskega označevanja, vključno z dvema spletnima servisoma, in se osredotočimo na naslednja dva nivoja označevanja: skladenjsko in semantično.

2. Korpusa JOS

Korpusa projekta JOS sta jos100k s 100.000 besedami in jos1M z enim milijonom besed. Oba sta bila izdelana z vzorčenjem korpusa FidaPLUS (Arhar in Gorjanc, 2007), 600-milijonskega referenčnega korpusa slovenščine, ki je označen z avtomatsko pripisanimi in na podlagi sobesedila razdvoumljenimi oblikoskladenjskimi oznakami in lemami. Prvi korak na poti od korpusa FidaPLUS do korpusa JOS je bila pretvorba v format XML, da bi s tem dobili korpus v standardnem formatu in omogočili uporabo orodij za delo s XML, predvsem XSLT.

Korpusa jos100k in jos1M sta bila izdelana iz korpusa FidaPLUS z dvostopenjskim procesom filtriranja in vzorčenja z namenom, da pri končnem rezultatu dosežemo uravnoteženost in reprezentativnost, poskrbimo za kvaliteto besedil (zaključeni stavki, primerna dolžina stavka itd.) in zagotovimo varovanje avtorskih pravic.

Format korpusov je XML, s shemo, ki je narejena s parametrizacijo priporočil Text Encoding Initiative P5 (TEI Consortium, 2007). Shema uporablja TEI module za označevanje korpusov, preprosto jezikoslovno analizo, povezovanje in za lastnostne strukture.

```
<s xml:id="F0020003.557.2">
  <w xml:id="F0020003.557.2.1" lemma="ta" msd="Zk-sei">To</w><S/>
  <w xml:id="F0020003.557.2.2" lemma="biti" msd="Gp-ste-n">je</w><S/>
  <term type="sloWNet" sortKey="kraj" subtype="missing_hyponym" key="ENG20-08114200-n">
    <w xml:id="F0020003.557.2.3" lemma="turističen" msd="Ppnmein">turističen</w><S/>
    <w xml:id="F0020003.557.2.4" lemma="kraj" msd="Somei">kraj</w>
  </term>
  <c xml:id="F0020003.557.2.5">.</c><S/>
</s>
<linkGrp type="syntax" targFunc="head argument" corresp="#F0020003.557.2">
  <link type="ena" targets="#F0020003.557.2.2 #F0020003.557.2.1"/>
  <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.2"/>
  <link type="dol" targets="#F0020003.557.2.4 #F0020003.557.2.3"/>
  <link type="dol" targets="#F0020003.557.2.2 #F0020003.557.2.4"/>
  <link type="modra" targets="#F0020003.557.2 #F0020003.557.2.5"/>
</linkGrp>
```

Slika 1: Primer stavka iz jos100k: »To je turističen kraj.«

Shema uvaja tudi razširitve glede na TEI P5, predvsem preverjanje oblikoskladenjskih oznak, uporabljenih v korpusih, neposredno iz sheme XML. Korpusa vsebujeta tudi obsežne metapodatke, kjer so navedene bibliografske informacije o posameznih besedilih, taksonomija zvrstnosti besedil, nabor oblikoskladenjskih oznak, opisi posameznih značilk v naborih itd. Korpus jos1M je označen le oblikoskladenjsko, korpus jos100k v verziji 2.0 pa vsebuje tri ravni jezikoslovnega označevanja.

Slika 1 ilustrira oznake, uporabljene v treh ravneh jezikoslovnega opisu korpusa jos100k. Posameznim besedam so pripisane osnovne oblike oz. leme in oblikoskladenjske oznake. Skladenjske povezave so shranjene ločeno, pri čemer sta s pomočjo odvisnostne oznake medsebojno povezani po dve pojavnici. Stavčni identifikator je uporabljen za povezavo na koren drevesa. Semantične oznake iz slovenskega leksikona sloWNet (Fišer, 2009), ki so enake identifikatorjem sinsetov princetonskega WordNeta (PWN[1]), so pripisane posameznim terminološkim enotam. Pri vsaki od enot je označen tudi jedrni samostalnik in v nekaterih primerih podtip, ki označuje manjkajoči sinset (ali ne dovolj specifične hiponime) v PWN. Oblikoskladenjske in odvisnostne oznake so v korpusu zapisane v slovenskem jeziku, vendar jih je mogoče zamenjati tudi z angleškimi ustreznicami.

Element	*n*	Razlaga
div	248	vzorčeno besedilo korpusa FidaPLUS
p	1,599	odstavek
s	6,151	stavek
term	5,430	literal iz baze WordNet
w	100,003	besedna pojavnica
c	18,391	ločilo
S	98,890	presledek
linkGrp	5,961	skladenjska analiza stavka
link	112,442	skladenjsko odvisnostno -razmerje

Tabela 1: Število oznak XML v korpusu jos100k

[1] http://wordnet.princeton.edu/

V Tabeli 1 navajamo število TEI elementov, uporabljenih v korpusu jos100k: v korpusu so deli besedil iz skoraj 250-ih besedil s 1.600 vzorčenimi odstavki, ki vsebujejo 6.000 stavkov. Korpus vsebuje preko 5.000 semantičnih oznak in nekaj manj kot 6.000 stavkov z odvisnostnimi oznakami. Trenutno jih je približno 5 % brez skladenjske analize.

3. Oblikoskladenjsko označevanje

Oblikoskladenjske specifikacije JOS (verzija 1.1) predstavljajo osnovo za označevanje korpusov na ravni besed. Vsebujejo definicije nabora natanko 1.902 oblikoskladenjskih oznak (MSD), posamezne oznake so v specifikacijah tudi razčlenjene glede na svoje lastnosti. Specifikacije so tako kot korpusi formatirane skladno s TEI P5 in so kompatibilne s slovenskim delom večjezičnih oblikoskladenjskih specifikacij MULTEXT-EAST, verzija 4 (Erjavec, 2010). Specifikacije JOS so na voljo tako v slovenščini kot v angleščini, prav tako kot vse oznake (MSD) in njihove lastnosti. Poleg formata XML in iz njega izvedenega formata HTML, je nabor oznak pripravljen tudi v tabelarični obliki, ki je opremljena z pretvorbami med različnimi formati.

Ročno označevanje, ki ga je pod nadzorom izvajala skupina študentov, je vključevalo popravljanje MSD-jev in lem v obeh korpusih JOS. Izvorne oznake iz korpusa FidaPLUS so bile pretvorjene v skladu s specifikacijami JOS. Oznake v korpusu jos100k so bile ročno preverjene dvakrat s strani dveh različnih označevalcev in kjer so se rešitve razlikovale, je bila končna odločitev preverjena s strani tretjega. Korpus jos100k torej lahko služi kot ročno označeni referenčni korpus za označevanje slovenščine.

Korpus jos1M je prav tako oblikoskladenjsko označen, vendar so bili ročno preverjeni le »sumljivi« MSD-ji pri približno 190.000 besedah, ker finančne možnosti projekta niso omogočale ročnega preverjanja celotnega korpusa. Poskusi pa so pokazali, da kljub napakam korpus jos1M kot učni korpus proizvede boljše označevalne modele kot jos100k. Korpus jos1M ima torej predvsem vlogo učnega korpusa za učenje statističnih označevalnikov in lematizatorjev za slovenščino.

Spletna stran JOS vsebuje dva spletna servisa: konkordančnik in avtomatski označevalnik. Po obeh

korpusih je mogoče iskati s pomočjo spletnega vmesnika, katerega v zaledju podpira CQP (Christ, 1994). Spletni vmesnik omogoča prikazovanje in iskanje po besedah ali po njim pripisanih oznakah: lemah, MSD-jih in celo po posameznih oblikoskladenjskih lastnostih (podpira denimo iskalne pogoje, kot je [število="dvojina" & naslonskost="klitična"]), kar pomeni, da je omogočeno tudi podrobno raziskovanje slovničnih značilnosti korpusa.

Oblikoskladenjski označevalnik in lematizator (Erjavec in Džeroski, 2004) sta bila naučena na korpusu jos1M in sta del spletnega servisa JOS. Uporabniki lahko na spletno stran naložijo svoja besedila, servis pa jih vrne tokenizirana, označena in lematizirana. Izhodni format je kompatibilen s formatom za storitev Sketch Engine[2] (Kilgarriff in dr., 2004), tako da uporabniki, ki imajo dostop to tega servisa, obdelana besedila lahko naložijo v Sketch Engine in uporabljajo zmogljiva orodja za analizo svojih korpusov.

4. Površinskoskladenjsko razčlenjevanje

Odvisnostne drevesnice so eden od osnovnih jezikovnih virov, ki jih uporabljamo za preučevanje skladenjskih pojavov in kot učne ali testne podatkovne zbirke za statistične razčlenjevalnike. Prvi poskus oblikovanja odvisnostne drevesnice za slovenščino predstavlja SDT (*Slovene Dependency Treebank*) (Džeroski in dr., 2006), kjer je bil del korpusa MULTEXT-East (Erjavec, 2004) označen z analitičnimi odvisnostnimi strukturami po modelu praške odvisnostne drevesnice (*Prague Dependency Treebank*) (Hajič in dr., 2006). Pri ročnem razčlenjevanju tega korpusa pa se je izkazalo, da je model izjemno zapleten in zato študenti težko sledijo navodilom.

Zato je bil v okviru projekta JOS razvit nov odvisnostni model, ki je bistveno preprostejši, čeprav v osnovi še vedno temelji na praškem modelu. Pri njem je število odvisnostnih razmerij zmanjšano na 10, navodilom za ročno razčlenjevanje pa je relativno enostavno slediti. V prvem koraku je bil razvit sam model, napisana so bila navodila in korpus s 500 vzorčnimi stavki je bil natančno razčlenjen, da bi testirali model in zagotovili bazo zgledov za označevalce. V drugem koraku je skupina študentov pod strokovnim nadzorom obdelala celoten korpus jos100k. Vsi stavki so bili s pomočjo posebej prirejenega grafičnega orodja razčlenjeni s strani dveh označevalcev, kjer je prišlo do razlik pri odločitvah, je končno odločitev sprejel tretji označevalec. Rezultat je prvi ročno površinskoskladenjsko razčlenjen korpus slovenskega jezika.

V Tabeli 2 prikazujemo poimenovanja skladenjskih razmerij v slovenščini in število njihovih pojavitev v razčlenjenem korpusu jos100k. Podrobna razlaga njihove rabe pri razčlenjevanju presega okvire tega prispevka, kljub temu pa jih v nadaljevanju na kratko opišemo. »Modra« povezuje abstraktno vozlišče stavka ali povedi z elementi, ki tvorijo nadaljnje povezave v odvisnostnem drevesu. »Del« povezuje elemente brez odvisnostnega razmerja v običajnem pomenu jedro-določilo. Ta razmerja so posledično opredeljena le kot deli besedne zveze, tipično deli povedka. »Dol« povezuje jedra z njihovimi določili v besednih zvezah. »Ena«, »Dve«, »Tri« in »Štiri« povezujejo osebke, predmete in prislovna določila, vendar ta odvisnostna razmerja ne ustrezajo povsem definicijam v tradicionalnih slovenskih slovnicah. »Prir« povezuje dele prirednih struktur na besednozvezni ravni, »Vez« pa s povezavo na veznike skupaj s »prir« povezuje dve jedri v priredni strukturi v trikotnik. »Skup« povezuje pojavnice, ki kažejo močno tendenco po pojavljanju skupaj in tvorijo večbesedno zvezo brez prepoznavne funkcije in v notranji strukturi ne kažejo odvisnostnih razmerij.

Odvisnost	***n***
modra	32,912
del	7,879
dol	36,873
ena	5,641
dve	7,445
tri	2,762
stiri	6,827
prir	2,896
vez	8,858
skup	349

Tabela 2: Površinskoskladenjska razmerja v jos100k

Korpus bo služil kot učna in testna podatkovna zbirka za učenje odvisnostnih razčlenjevalnikov. Desetkratno prečno preverjanje je pri poskusih z razčlenjevalnikom MST[3] pokazalo 80 % natančnost razčlenjevanja, če gledamo tako povezavo kot odvisnostno oznako, oz. 84 %, če gledamo samo pravilnost povezave med pojavnicama.

5. Semantično označevanje

Semantično označeni korpusi so nepogrešljivi vir za razvoj sodobnih jezikovnih tehnologij. Za označevanje, ki je potekalo ročno, smo uporabili t.i. »slovarski model«, v okviru katerega označevalec za vsako pojavnico v korpusu, ki jo želi označiti, preveri njene pomene v slovarju, ki ga za označevanje uporablja, in glede na sobesedilo izbere najustreznejšega. Namesto klasičnega slovarja smo kot nabor pomenov uporabili semantični leksikon sloWNet, ki je bil izdelan s polavtomatskimi metodami iz že obstoječih večjezičnih virov, kot so dvojezični slovar, vzporedni korpusi in Wikipedija (Fišer in Sagot 2008). Trenutno sloWNet vsebuje približno 20.000 literalov, ki so razvrščeni v nekaj manj kot 17.000 sinsetov; vsebuje tako splošne kot tudi specifične pojme. Splošni večinoma prihajajo iz slovarja in vzporednega korpusa, specifični pa so bili pridobljeni predvsem iz Wikipedije. Trenutno v sloWNetu najdemo predvsem samostalniške pojme, nekaj pa je tudi glagolskih, pridevniških in prislovnih, ki so tako enobesedni kot večbesedni. Primerjava besedišča v sloWNetu in korpusu jos100k pokaže, da sloWNet vsebuje 30 % samostalnikov iz korpusa, pri čemer pokriva 90 % tistih, ki po pogostnosti sodijo v zgornjo tretjino.

[2] http://www.sketchengine.co.uk/

[3] MSTParser (McDonald in dr., 2006) je dosegel najboljše rezultate pri razčlenjevanju korpusa SDT na tekmovanju večjezičnih odvisnostnih razčlenjevalnikov CoNLL-X Shared Task 2006.

Za razliko od sekvenčnega označevanja, pri katerem označujemo celoten korpus besedo za besedo, smo se v tej raziskavi odločili za ciljno semantično označevanje (Miller in dr., 1994), kjer označujemo samo določene besede v korpusu. Da je ciljno označevanje učinkovitejše od sekvenčnega, poudarjajo številni avtorji (glej Kilgarriff 1998), saj na ta način semantične lastnosti določene besede obravnavamo hkrati, zaradi česar je označevanje bolj konsistentno. Ker je bil sloWNet izdelan polavtomatsko in sloni na tujejezičnem semantičnem leksikonu, smo poleg ciljnega označevanja v raziskavi uporabili koordiniran pristop (Agirre in dr., 2006), v skladu s katerim smo vzporedno z označevanjem preverjali in popravljali tudi sloWNet, s čimer smo zagotovili boljše ujemanje med pomeni v leksikonu in v korpusu.

Glede na to, da se s semantičnim označevanjem ukvarjamo prvič, smo se v raziskavi omejili na označevanje samostalnikov, saj je ravno določanje pomena samostalnikom najenostavnejše, prav tako pa so ti tudi najbolje zastopani v sloWNetu. Iz korpusa jos100k smo izluščili vse samostalnike, ki se v korpusu pojavljajo 30- ali večkrat in so hkrati tudi v sloWNetu, s čimer smo dobili 102 samostalnika. Najpogostejši samostalnik je *leto* s 346 pojavitvami, ostale besede so precej redkejše, saj se jih več kot 100-krat v korpusu pojavi le še šest (*dan, delo, čas, človek, država* in *svet*), seznam pa se konča s sedmimi besedami, ki se v korpusu pojavijo 30-krat (*besedilo, oče, pogled, predstavnik, projekt, razvoj* in *cesta*). Skupno število pojavitev samostalnikov, ki smo jih v korpusu označili, je 5.431 oziroma povprečno 53,2 pojavitev na besedo.

Iz korpusa smo izluščili konkordance za izbrane besede in jih shranili v ločene datoteke, po eno za vsako besedo. Čeprav so korpus označevali štirje različni označevalci, je bil za popravljanje sloWNeta in označevanje vseh pojavitev izbrane besede v korpusu vedno zadolžen isti označevalec. Med validacijo sloWNeta so označevalci pregledali vse sinsete, v katerih se njihova beseda pojavlja (vse pomene te besede), pa tudi vse večbesedne zveze, v katerih se njihova beseda v sloWNetu pojavlja (ponavadi, ne pa vedno, v vlogi podpomenke dodeljene besede). V primeru, da so v sinsetu odkrili napako, so napačen literal popravili (npr. napačno veliko začetnico v malo). Če so v sinsetu našli literal, ki tja ne sodi, so ga izbrisali, če pa so ugotovili, da v sinsetu nek literal manjka, so ga dodali. Pregledovanju sloWNeta je sledilo označevanje izbranih besed v korpusu.

Označevalci so v korpusu označili 5.431 pojavnic, ki so jim pripisali 517 različnih pomenov oz. povprečno 5,1 pomen na samostalnik. Največ (19,6 %) besed so označili s tremi različnimi pomeni. Sedmim samostalnikom so pripisali enega samega (*delavec, ministrstvo, minuta, muzej, odstotek, podjetje in sezona*), največ, 14, pomenov pa so pripisali besedama *čas* in *vrsta*. Več kot deset pomenov je bilo pripisanih še trem besedam: *prostor*, *konec* in *življenje*. 46 pojavnic je bilo označenih kot lastno ime, ki ga ni v sloWNetu, 25 pojavnic (0,1 %) pa je ostalo neoznačenih, saj označevalci zanje v sloWNetu niso našli nobenega ustreznega pomena. V večini teh primerov gre za kulturno-specifične pomene, ki jih bo potrebno naknadno dodati v sloWNet (npr. *voda na nekogaršnji mlin*).

Zanesljivost oznak smo preverili na vzorcu 500 naključno izbranih stavkov, ki so jih različni označevalci označili dvakrat. Dvojne oznake smo primerjali in na podlagi tega izračunali 66 % ujemanje med označevalci. Rezultati so sicer nekoliko nižji kot pri sorodnih eksperimentih v drugi jezikih, vendar je pri tem treba upoštevati dejstvo, da smo pri tem projektu označevali izključno najpogostejše samostalnike, ki ponavadi izkazujejo tudi najvišjo stopnjo večpomenskosti, zato je bilo delo naših označevalcev težje.

6. Zaključek

V prispevku smo predstavili rezultate projekta JOS, ki vsebujejo oblikoskladenjske specifikacije, dva označena korpusa in dva spletna servisa. Natančneje je opisan korpus jos100k, ki vsebuje ročno pripisane oblikoskladenjske oznake, leme, površinskoskladenjska odvisnostna razmerja in pri izbranih samostalnikih pomene iz leksikalne baze WordNet. V osnovi bo korpus služil kot učna in testna podatkovna zbirka za razvoj slovenskih oblikoskladenjskih označevalnikov, odvisnostnih razčlenjevalnikov in programov razreševanje večpomenskosti. Poleg rabe za namene razvoja jezikovnih tehnologij je korpus namenjen tudi jezikoslovcem, čeprav se sistem označevanja v določeni meri razlikuje od tradicionalnega pojmovanja jezikovnih pojavov.

Korpusi JOS in z njimi povezani viri so na voljo v formatu XML TEI P5, pa tudi v drugih, izvedenih formatih, ki so bolj primerni za specifične namene nadaljnjega procesiranja. Poleg jezikovnih virov sta rezultat projekta tudi dva spletna servisa, zanimiva za jezikoslovno rabo: spletni konkordančnik za oba korpusa in spletni označevalnik, ki omogoča tokenizacijo, lematizacijo in oblikoskladenjsko označevanje slovenskih besedil.

Viri projekta JOS so na voljo na spletni strani projekta[4], korpusa pod licenco Creative Commons Priznanje avtorstva-Nekomercialno 3.0.[5]

Predstavljena korpusa sta prvi tovrstni javno dostopni vir za slovenščino in bi zato lahko pomembno prispevala k razvoju jezikovnih tehnologij za slovenski jezik.

Delo na jezikovnih virih JOS za slovenski jezik se nadaljuje z dolgoročnim projektom »Sporazumevanje v slovenskem jeziku«[6], v okviru katerega je predvidena nadaljnja gradnja učnega korpusa in razvoj leksikona besednih oblik, milijardnega referenčnega korpusa, leksikalne baze ter drugih jezikovnih virov in priročnikov za slovenski jezik. V tem okviru bosta korpusa JOS nadgrajena v polmilijonski v celoti ročno preverjeni korpus z oblikoskladenjskimi oznakami, odvisnostnimi razmerji in ročno označenimi lastnimi imeni.

Zahvala

Delo, opisano v tem prispevku, sta omogočila projekt ARRS J2-9180 “Jezikoslovno označevanje slovenskega jezika: metode in viri” in projekt EU 6FP-033917 SMART “Statistical Multilingual Analysis for Retrieval and Translation”.

[4] http://nl.ijs.si/jos/
[5] http://creativecommons.org/licenses/by-nc/3.0/deed.sl
[6] http://www.slovenscina.eu/

Literatura

Eneko Agirre, in Philip Edmonds. 2006: Word Sense Disambiguation: Algorithms and Applications. Dordrecht: Springer.

Špela Arhar in Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega. *Jezik in slovstvo*, 52(2).

Oliver Christ. 1994. A Modular and Flexible Architecture for an Integrated Corpus Query System. V *Proceedings of COMPLEX '94* (str. 23–32). Budimpešta.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtsky in Andreja Žele. 2006. Towards a Slovene Dependency Treebank. V *Proceedings Fifth International Conference on Language Resources and Evaluation, LREC'06*, Paris. ELRA.

Tomaž Erjavec in Sašo Džeroski. 2004. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

Tomaž Erjavec in Simon Krek. 2008. Oblikoskladenjske specifikacije in označeni korpusi JOS. Zbornik Šeste konference Jezikovne tehnologije, Ljubljana.

Tomaž Erjavec. 2010. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC'10*, Paris. ELRA.

Darja Fišer. 2009. sloWNet - slovenski semantični leksikon. *Obdobja 28*. Ljubljana. str. 145-149.

Darja Fišer in Tomaž Erjavec. 2010. sloWNet: Construction and Corpus Annotation. V *Proceedings of Fifth International Conference of the Global WordNet Association (GWC'10)*, Mumbai.

Jan Hajič, Jarmila Panevová, Eva Hajičova, Petr Pajas, Petr Sgall, Jan Štěpánek, Jíří Havelka, in Marie Milkulová. 2006. *Prague Dependency Treebank 2.0*. Catalog Number LDC2006T01.

Adam Kilgarriff. 1998. Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. *Computer Speech and Language: Special Issue on Evaluation* 12 (4), 453–472.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. V *Proceedings of the 11th EURALEX International Congress,* str. 105–116, Lorient, France.

Nina Ledinek, Tomaž Erjavec: Odvisnostno površinskoskladenjsko označevanje slovenščine: specifikacije in označeni korpusi. Zbornik Simpozija Obdobja: Infrastruktura slovenščine in slovenistike, Ljubljana, 2009.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser. V *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, Robert G. Thomas. 1994. Using a semantic concordance for sense identification. *Proceedings of the workshop on Human Language Technology*.

TEI Consortium, editor. 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Strojno prevajanje in slovenščina

Jernej Vičič*

*Primorski Inštitut za Naravoslovje in Tehnologijo, Univerza na Primorskem
6000 Koper
jernej.vicic@upr.si

Povzetek

Članek predstavlja pregled strojnih prevajalnih sistemov, ki omogočajo prevajanje v slovenski jezik ali iz slovenskega jezika. Osnovo članka predstavlja obsežna primerjava vseh sistemov, ki podpirajo prevode v ali iz slovenskega jezika. Uporabljene metodologije evalvacije kakovosti prevodov so bile: samodejni objektivni metodi na osnovi metrik BLEU in METEOR ter subjektivni ročni metodi, prva temelječa na metriki WRR (Word-Recognition Rate) ter metoda, temelječa na smernicah LDC. Primerjani sistemi so: Google translate, Microsoft Bing translator, Amebis Presis ter GUAT. Rezultati raziskave so pokazali velik napredek sistemov temelječih na statističnih metodah, vendar ti sistemi kažejo sistematske napake, ki jih bo z opisanimi metodami težko odpraviti.

Machine Translation and Slovenian Language

The paper presents an overview of the machine translation systems that deal with Slovenian language. It is based on a comprehensive evaluation and comparison of all the publicly available machine translation systems that include Slovenian language in translation pairs. Evaluation methodologies used in the experiment were: automatic evaluation based on metrics BLEU and METEOR and subjective assisted evaluation based on metric WRR (Word-Recognition Rate) and LDC guidelines. The systems included in the comparison were: Google translate, Microsoft Bing translator, Amebis Presis and GUAT.

1. Uvod

Članek predstavlja pregled strojnih prevajalnih sistemov, ki omogočajo prevajanje v slovenski jezik ali iz slovenskega jezika. Kakovost strojnega prevajanja je v zadnjih leti zelo napredovala, s prihodom velikih korporacij pa je tudi dostopnost prevajalnih sistemov olajšana. Članek v osnovi poskuša odgovoriti na jasno vprašanje: kateri prevajalni sistem je najboljši, vendar se je med izvedbo eksperimentov pokazalo, da odgovor na to vprašanje ni nedvoumen. Predstavljeni so vsi sistemi, ki med podprtimi jezikovnimi pari ponujajo slovenščino. Pri primerjavi sistemov so bile uporabljene samodejne ter ročne metodologije evalvacije kakovosti prevodov, rezultati kažejo, da je korelacija med samodejnimi metodami ter ročnim testiranjem pri nekaterih sistemih zelo nizka.

Izdelan je bil povedno poravnan korpus besedil, ki je bil uporabljen v testne namene pri vseh metodologijah evalvacije.

Nadaljevanje članka je razdeljeno, kot sledi: v Razdelku 2. so prikazani razlogi za izbiro jezikovnih parov, uporabljenih v eksperimentu, v podrazdelkih je vsak jezikovni par obširneje predstavljen z vidika strojnega prevajanja. Razdelek 3. predstavlja izbiro prevajalnih sistemov ter obširne opise uporabljenih tehnologij pri posameznih sistemih. Sledi Razdelek 4., ki predstavlja metodologijo testiranja ter samo testiranje in rezultate testiranja. V Razdelku 5. so predstavljeni osnovni izsledki, nakazano je tudi možno nadaljnje delo.

2. Jezikovni pari

Evalvacijo smo razdelili na dve skupini; na testiranje sistemov s prosto izbiro jezikovnih parov ter na sisteme sorodnih jezikov. Za jezikovne pare prve skupine smo izbrali skupni jezikovni par slovenščina-angleščina, ki je v povezavi s slovenskim jezikom tudi največkrat uporabljan. Med sorodnimi jezikovnimi pari smo izbrali jezikovni par slovenščina-srbščina, ki edini omogoča primerjavo med sistemi, saj edini slovenski sistem podpira le ta jezikovni par.

2.1. Slovenščina-angleščina

Slovenščina in angleščina nista sorodna jezika, najpomembnejše razlike so na morfološkem ter skladenjskem nivoju. Angleščina je jezik z dokaj omejeno morfologijo ter relativno fiksnim besednim vrstnim redom. Z vidika strojnega prevajanja (Machine Translation - MT) je pomembno, da lahko angleščina kaže večjo stopnjo dvoumnosti posameznih besednih oblik v odvisnosti od njihovih morfoloških oznak, še posebej besednih vrst; ista besedna oblika je lahko samostalnik, glagol ali kaj drugega. To se lahko zgodi tudi v slovenščini, vendar v veliko manjšem obsegu. Druga lastnost angleščine, ki vpliva na kakovost prevodov, je njena sposobnost za gradnjo zapletenih samostalniških zloženk (npr. long term car park courtesy pickup vehicle), medtem ko slovenščina (in tudi drugi slovanski jeziki) veliko pogosteje uporabljajo kombinacijo pridevnikov in samostalnikov (dostavno vozilo dolgoročnega parkirišča), kar pomeni, da je običajno potrebno prevesti pridevnik kot samostalnik ali obratno, kar je seveda izziv za vsak strojni prevajalni sistem.

2.2. Slovenščina-srbščina

Tako slovenščina kot srbščina pripadata skupini južnoslovanskih jezikov, ki jih govorijo večinoma prebivalci z območja bivše Jugoslavije. Srbščina je najbolj razširjena v Srbiji, slovenščina v Sloveniji. Jezika si delita skupne ko-

renine ter še pomembneje skupno polpreteklo zgodovino; oba jezika sta bila uradna jezika skupne države, celo predavana v šolah kot materina jezika ter jezika okolja.
Obstaja niz pomembnih razlogov za postavitev prevajalnega sistema za predstavljen jezikovni par, omenimo le najpomembnejša: ekonomiji obeh novih držav, Slovenije ter Srbije, sta še vedno tesno povezani; mlajše generacije imajo jezikovne težave v medsebojni komunikaciji.
Oba jezika sta visoko pregibna ter morfološko bogata.

2.3. Izbira in priprava testnega gradiva

Začetni načrt testiranja je predvideval uporabo naključno izbranih povedi vzporedno poravnanega večjezičnega dela korpusa MULTEXT-EAST (Dimitrova et al., 1998), romana 1984 Georga Orwella. Ta korpus vsebuje poravnane povedi v vseh treh jezikih, ki smo jih nameravali uporabiti pri testiranju.
Prvi krog testiranj je pokazal, da je Google zelo verjetno uporabil korpus (Dimitrova et al., 1998) pri učenju svojega sistema. Podobnost prevodov sistema z referenčnimi prevodi je bila več kot le naključna.
Odločili smo se, da bomo zamenjali testno gradivo. Uporabili bi lahko druge večjezične korpuse kot je na primer Opus (Tiedemann, 2009). Izključili smo vse korpuse, ki so dostopni prek interneta, saj pri teh korpusih obstaja možnost, da so bili uporabljeni kot učno gradivo pri izdelavi prevajalnih sistemov.
Pripravili smo novo učno množico, ki je bila zgrajena iz podnapisov filma, izbrali smo Matrix. Ta film je bil, in je še, zelo priljubljen med piratskimi uporabniki, tako smo lahko pričakovali prevode v vseh želenih jezikih (slovenščina, angleščina ter srbščina) in tudi dovolj kakovostne prevode, saj skupnost večih uporabnikov omogoča preverjanje ter popravljanje napak v prevodih ter ocenjevanje posameznih prevodov. Izbrali smo različice prevodov z najboljšimi ocenami ter največ prenosi ter jih še ročno pregledali.
Podnapisi filmov že vsebujejo informacijo o času prikaza na zaslonu, kar omogoča enostavno izdelavo poravnanih korpusov. Izdelali smo povedno poravnan korpus dveh jezikovnih parov z enim skupnim jezikom. Pri poravnavi smo uporabili večino metod, opisanih v (Tiedemann, 2007). Segmente, kjer je bilo časovno ujemanje podnapisov dvoumno, smo izpustili. Podnapisi filmov imajo v povprečju krajše, posledično strukturno manj zapletene povedi. To dejstvo smo omejili z izločitvijo najkrajših povedi, mejo smo postavili empirično pri 20 znakih. Tako smo dobili korpus z osnovnimi lastnostmi, predstavljenimi v Tabeli 1

Tabela 1: Osnovni podatki o testnem korpusu

jezikovni par	jezik	št. povedi	št. besed
sl-en	slovenščina	760	4814
sl-en	angleščina	760	5646
sl-sr	slovenščina	742	4624
sl-sr	srbščina	742	4460

3. Pregled sistemov

Po pregledu spleta ter pogovoru s strokovnjaki področja smo se odločili za naslednji izbor prevajalnih sistemov: Google Translate, Microsoft BING translator, Amebis Presis ter GUAT. Vrstni red sistemov je naključen. Vse izbrane sisteme lahko uvrstimo v dve paradigmi strojnega prevajanja.
Google Translate ter Microsoft BING translator spadata v paradigmo sistemov statističnega strojnega prevajanja (Statistical Machine Translation - SMT). Takšni sistemi so osnovani na parametričnih statističnih modelih, ki so naučeni na poravnanih dvojezičnih korpusih (učnih primerih). Namesto razdeljevanja stavkov po slovničnih pravilih iščemo splošne vzorce, ki se porajajo pri uporabi jezika. Glavna prednost statističnega pristopa je, da so metode neodvisne od jezika (čeprav uporabnejše za določene jezike, med temi ni slovenščine). Glavna pomanjkljivost sistemov strojnega prevajanja na osnovi pravil je slab pregled nad delovanjem sistema, sistematske napake je zelo težko odpraviti, uvajanje lingvističnega znanja je le delno mogoče oziroma celo nemogoče.
Amebis Presis ter GUAT spadata med sisteme strojnega prevajanja temelječe na pravilih (Rule-Based Machijne Translation - RBMT). Način zapisa pravil se razlikuje med sistemi, veže pa jih skupno dejstvo, da je postavitev takšnega sistema dolgotrajno opravilo. V to skupino sodi večina današnjih komercialnih prevajalnih sistemov, čeprav se pri gradnji poslužujejo nekaterih manj standardnih prijemov. Sistemi te paradigme izvorno besedilo najprej morfološko ter skladenjsko analizirajo ter izdelajo predstavitev vhodnega besedila, ponavadi v obliki skladenjskega drevesa izpeljave. Ta predstavitev se še dodatno abstrahira s poudarkom na zahtevah strojnega prevajanja. Proces transferja prevede abstraktno predstavitev vhodnega besedila v izvornem jeziku v podobno predstavitev v ciljnem jeziku, to predstavitev sistem uporabi kot osnovo za generacijo besedila v ciljnem jeziku, v bistvu uporabi inverzne metode prvega dela na ciljnem jeziku.

3.1. Google Translate

Google Translate[1] je tipični pripadnik sistemov statističnega strojnega prevajanja (Statistical Machine Translation - SMT), ki je predstavljena v Razdelku 3.. Sistem ne uporablja dodatnega jezikovnega znanja, zanaša se samo na korelacijo med znanimi pari že prevedenih vzporednih besedil. Statistične metode zahtevajo ogromne količine besedil ter veliko računalniške moči za obdelavo teh besedil. Google ima oboje, besedila, nabrana za izdelavo iskalnika, ter veliki grozdi računalnikov omogočajo hitro izdelavo sistemov za strojno prevajanje z zavidljivo kakovostjo.
Opisane lastnosti so omogočile Googlu izdelavo prevajalnih sistemov za kar 58 svetovnih jezikov[2], torej za kar $58 * 58 = 3364$ jezikovnih parov.

O natančnem delovanju prevajalnega sistema Google Translate ni veliko znanega, znane so le osnovne metode. Pri uporabi sistema za prevode ter pri izvajanju evalvacije prevajalnega sistema smo prišli do podobnih zaključkov kot nekaj avtorjev, ki je svoja mnenja predstavilo na odprti debati na dopisnem seznamu Corpora (List, 2010).

[1] Google Translate: `http://translate.google.com`
[2] `http://translate.google.com`, podatek z dne: 20.6.2010

Za večje jezike ter pogostejše jezikovne pare, kot so angleščina-nemščina, angleščina-francoščina ali angleščina-kitajščina, so sistemi naučeni posebej, medtem ko so manjši jeziki, mednje sodi tudi slovenščina, prevajani prek vmesnega jezika. To pomeni, da prevajanje v slovenščino ne poteka neposredno iz izvornega jezika, ampak se najprej besedilo prevede v angleščino (ponovno špekulacija) ter šele nato v slovenščino. Kakovost prevodov je slabša kot pri večjih jezikovnih parih, kar potrjujejo tudi rezultati testiranja, prikazani v Razdelku 4. Predstavljeno špekulacijo lahko potrdimo le z empiričnimi primeri prevodov, kot je primer prikazan na Sliki 1

```
Razumeš?
Гот ит?
Romanizirano:  Got it?
```

Slika 1: Primer prevoda iz slovenščine v srbščino, kjer se v prevodu pojavljajo angleške besede. Primer kaže na verjetno uporabo angleščine kot vmesnega jezika pri prevajanju.

3.2. Microsoft BING Translator

Bing Translator [3] (Quirk et al., 2005) je hibridni sistem za strojno prevajanje naravnih jezikov. Sistem temelji na statističnem strojnem prevajalniku, obširneje je predstavljen v Razdelku 3., ki uporablja tudi pravila, ki so odvisna od jezika ter določeno mero analize izvornega besedila. Microsoft imenuje ta sistem kot "jezikovno obveščeno statistično strojno prevajanje"(Linguistically informed statistical machine translation).
Sistem je v osnovi statistični sistem za strojno prevajanje na osnovi fraz, ki vključuje jezikovno odvisno analizo besedila, drevesa odvisnosti (dependency trees) ter drevesa izpeljave (parse trees) in pravila za poravnavo besed (word alignment rules) za generalizacijo naučenih fraz.

3.3. Amebis Presis

Prevajalni sistem Presis podjetja Amebis (Romih in Holožan, 2002) je bil prvi sistem za strojno prevajanje, ki je med prevajalnimi jezikovnimi pari vseboval slovenski jezik.
Sistem sodi v paradigmo strojnih prevajalnih sistemov na osnovi pravil (Rule-Based Machine Translation - RBMT), natančneje je predstavljena v Razdelku 3., Presis analizira vsako poved v izvornem jeziku v slovnične komponente, kot so osebek, predmet, povedek in atributi ustreznih semantičnih kategorij. Na osnovi analiziranega izvornega besedila izbere pripravljena pravila, ki omogočajo prevod analiziranih komponent v ciljni jezik, nato sintetizira poved v ciljnem jeziku.
Ena glavna odlik sistema Presis je možnost prilagoditve prevajalni domeni z vključitvijo dodatnih slovarjev vendar primerjava tako spremernjenega sistema ne bi bila objektivna, zato smo to možnost opustili.

[3]Microsoft Bing Translator: `http://www.microsofttranslator.com/`

Pri testiranju smo uporabili spletno različico prevajalnega sistema Amebis Presis 2.0-pre37 [4]

3.4. GUAT

GUAT [5] je rezultat akademskih eksperimentov metod za hitro postavitev prevajalnih sistemov za sorodne jezike. Izdelava tega sistema je opisana v (Vičič, 2009) in (Vičič in Homola, 2010). Osnova sistema je odprtokodno ogrodje Apertium (Corbi-Bellot et al., 2005). Apertium je ogrodje za izdeavo za sistemov za strojno prevajanje na osnovi pravil plitkega prenosa (Shallow Transfer Rule-Based Machine Translation - RBMT). Okolje je posebej primerno za sorodne jezike, saj omogoča le morfološko analizo ter sintezo besedil, ne vsebuje pa orodij za popolno analizo izvornih besedil. Prevajalni sistemi za skladenjsko različne jezike, kot sta slovenščina in angleščina bi dosegali slabše rezultate.

3.5. Ostala orodja

Pri pregledu obstoječih sistemov smo naleteli še na dva prevajalna sistema, ki vsebujeta slovenščino med prevajalnimi pari: VoiceTran[6] (Žganec Gros et al., 2005) ter Prevajalnik.net[7].
Prvi je glasovni komunikator in je najbolj uporaben za podajanje povelj ter postavljanje preprostih vprašanj. Omogoča prepoznavanje govora ter ročne pisave. Del sistema VoiceTran je tudi sistem za prevajanje besedila. Zaradi ožje namembnosti sistema bi bila smiselnost takšnega testiranja vprašljiva. Sistem temelji na posebni različici orodja Presis, ki je opisano v Razdelku 3.3.
Prevajalnik.net je spletna storitev, ki omogoča prevode iz slovenščine ter v slovenščino za več kot 40 svetovnih jezikov. Storitev uporablja Googlovem prevajalnik, ki je opisan v Razdelku 3.1. Za testiranje tega sistema se nismo odločili, saj bi bili rezultati identični Googlovim.

4. Metodologija vrednotenja ter rezultati

Testni primeri, uporabljeni pri testiranju so predstavljeni v Razdelku 2.3., za ponovljivost rezultatov raziskave so vsa gradiva na voljo na naslovu: `http://jt.upr.si/mt_v_sloveniji/`. Pri vrednotenju so bile uporabljene štiri metode:

1. Samodejno vrednotenje s pomočjo metrike BLEU (Papineni et al., 2001).
2. Samodejno vrednotenje s pomočjo metrike METEOR (Lavie in Agarwal, 2007).
3. Ročno vrednotenje, temelječe na metriki WRR (Word-Recognition Rate)
4. Ročno vrednotenje, temelječe na smernicah LDC (LDC, 2005).

Metriki BLEU ter METEOR sta najbolj razširjeni samodejni metriki za evalvacijo strojnega prevajanja, odločili

[4]Presis, `http://presis.amebis.si/`
[5]Prevajalni sistem GUAT: `http://jt.upr.si/guat/`
[6]VoiceTRAN: `http://www.voicetran.org/`
[7]VoiceTRAN: `http://prevajalnik.net/`

smo se za uporabo obeh, čeprav se mnogi avtorji strinjajo, da metrika BLEU ni primerna za primerjavo različnih sistemov (Callison-Burch et al., 2006). Pri uporabi samodejnih metrik smo izvorne povedi poravnanega korpusa predstavljenega v Razdelku 1 prevedli s prevajalnim sistemom. Prevode sistemov in referenčne prevode, ciljne povedi testnega korpusa, smo primerjali z obema metrikama. Uporabili smo vse razpoložljive testne primere, osnovni podatki o testnem korpusu so zapisani v Razdelku 1.

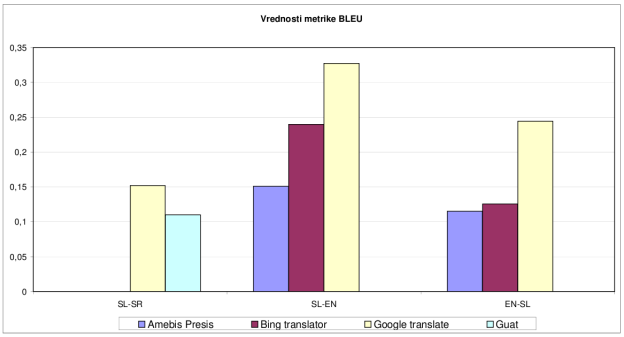
Obe metodi ročnega vrednotenja sta bili izvedeni na podoben način: evalvacija je obsegala naključno izbiro 100 povedi iz obsega vseh testnih povedi, prevod teh povedi s prevajalnim sistemom in izvedbo ročne evalvacije. Testiranje smo opravili s po dvema ocenjevalcema za vsak ciljni jezik. Slovenščina je bila obema ocenjevalcema materin jezik, za angleščino pa smo uporabili enega ocenjevalca, ki mu je angleščina materin jezik in enega, ki aktivno obvlada ta jezik.

4.0.1. Samodejno vrednotenje s pomočjo metrike BLEU

Bilingual Evaluation Understudy - BLEU (Papineni et al., 2001) je bila prva in še vedno najbolj razširjena metrika za evalvacijo kakovosti prevodov sistemov strojnega prevajanja. Kakovost prevodov je predstavljena kot natančnost ujemanja prevodov sistema za strojno prevajanje z referenčnimi prevodi poklicnih prevajalcev. Vrednosti so izračunane za posamezne prevedene odseke, ponavadi povedi, ter povprečene za celoten testni korpus. Berljivost ter slovnična pravilnost nista upoštevani.

BLEU uporablja spremenjeno različico preciznost (precision), ki za razred predstavlja število pravilno klasificiranih elementov (true positives), za primerjavo kandidata za prevod z enim ali več referenčnimi prevodi. Sprememba z osnovno preciznost naj bi poskrbela za lastnost sistemov strojnega prevajanja, ki težijo k daljšim prevodom.

Uporabili smo javno dostopno implementacijo metrike BLEU (NIST, 2008), različico v11b. Rezultati so prikazani na Sliki 2, večje vrednosti predstavljajo boljše rezultate. Rezultati testiranja z metriko BLEU kažejo na ve-



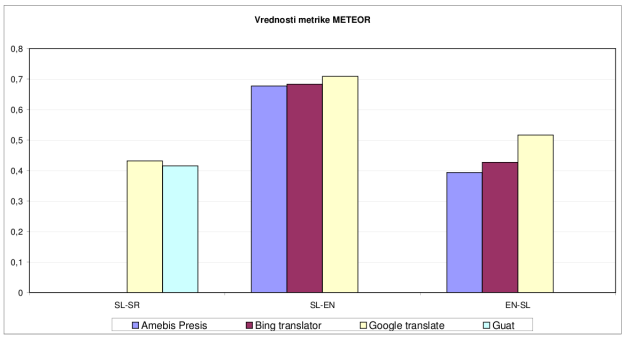
Slika 2: Rezultati testiranja z metriko BLEU. Sistem Google dosega najboljše rezultate.

liko prednost sistema Google translate. Pomisleki avtorjev glede predstavljenih rezultatov metrike BLEU so predstavljeni v Razdelku 5.

4.0.2. Samodejno vrednotenje s pomočjo metrike METEOR

Uporabili smo javno dostopno implementacijo metrike METEOR (Lavie in Agarwal, 2007), različico 0.6. Metrika temelji na harmonični sredini natančnosti ter priklica unigramov (unigram precision and recall), kjer je priklic močneje utežen kot natančnost. Vsebuje še več metod jezikovnih tehnologij, ki niso prisotne pri ostalih samodejnih metrikah strojnega prevajanja kot so krnjenje in ujemanje sinonimov kot pomoč pri iskanju ujemanja besed. Krnjenje je predvsem primerno za visiko pregibne jezike saj omejuje vpliv napačne uporabe pregibanja; na primer napačne uporabe sklona pri samostalnikih.

Samo orodje nima priložene kode za krnjenje srbskega jezika, uporabili smo orodje, ki je bilo izdelano za vrednotenje rezultatov metod, opisanih v (Vičič in Homola, 2010). Rezultati so predstavljeni na Sliki 3, večje vrednosti predstavljajo boljše rezultate.

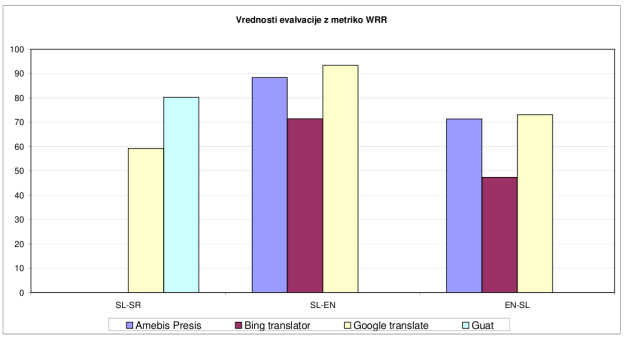


Slika 3: Rezultati metrike METEOR.

4.0.3. Ročno vrednotenje, temelječe na metriki WRR (Word-Recognition Rate)

Metrika temelječa na utežena Levenshteinovi razdalji (weighted Levenshtein edit-distance) (Levenshtein, 1965), poznana tudi kot Word Error Rate (WER) izračuna najmanjše število sprememb, ki jih moramo narediti za izdelavo *pravilne* povedi v ciljnem jeziku iz samodejno izdelane povedi (prevoda ocenjevanega sistema). Število sprememb še utežimo z dolžino povedi. Dovoljene spremembe so vstavitev, brisanje ter zamenjava besede.

Sama izvedba evalvacije je bila sestavljena iz samodejnega prevoda testnih primerov ter preštevanje števila sprememb, ki jih moramo opraviti za izdelavo dovolj dobrega prevoda. Definicija dovolj dobrega prevoda, ki smo jo uporabili pri tem eksperimentu, je prevod, ki je sintaktično pravilen ter izraža poln pomen izvorne povedi. Rezultati na Sliki 4 predstavljajo WRR, Word Recognition rate (1 - WER), ki predstavlja uspešnost prevajalnega sistema in ne velikosti napake sistema, torej večje vrednosti so boljše.



Slika 4: Rezultati evalvacije z metriko Word Recognition Rate - WRR. Pri jezikovnem paru s srbščino dosega GUAT signifikantno boljše rezultate kot Google, pri jezikovnih parih z angleščino so signifikantno slabši rezultati sistem Bing, razlika med ostalima sistemoma ni signifikantna.

4.0.4. Ročno vrednotenje, temelječe na smernicah LDC

Subjektivno ročno ocenjevanje kakovosti prevajalnih sistemov smo izvedli po smernicah določenih na delavnici NIST Machine Translation Evaluation Workshop konzorcija Linguistic Data Consortium - LDC (LDC, 2005). Ta metodologija je najbolj razširjena pri ročni evalvaciji sistemov strojnega prevajanja.

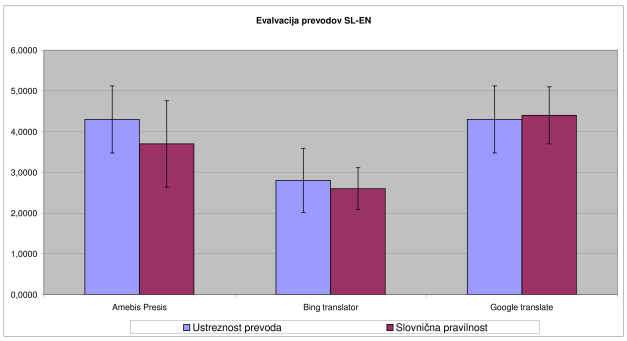
Uporabljeni sta bili dve pet-točkovni lestvici, ena opisuje ustreznost prevodov (adequacy), druga pa slovnično pravilnost prevodov (fluency). Lestvica ustreznosti prevodov (adequacy) kaže, koliko pomena iz referenčne povedi je izraženega v hipotetičnem prevodu: 5 = Vse/All, 4 = Večina/Most, 3 = Veliko/Much, 2 = Malo/Little, 1 = Nič/None.

Druga lestvica kaže slovnično pravilnost prevodov in je vsebina ne zanima. Vrednosti lestvice ustrezajo: 5 = Slovnično popoln prevod/Flawless translation, 4 = Slovnično dobro besedilo/Good target language, 3 = Besedilo ni materin jezik/Non-native target language, 2 = Slovnično nepravilno besedilo/Disfluent target language, 1 = Nerazumljivo besedilo/Incomprehensible text.

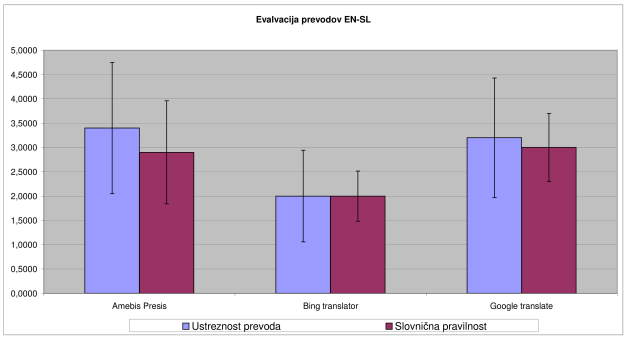
Ločeni lestvici za ustreznost prevodov ter slovnično pravilnost sta bili razviti pod predpostavko, da je lahko prevod slovnično nepravilen, pa še vedno vsebuje ves pomen izvornega besedila in v določenih primerih to že zadošča.

Rezultati so predstavljeni na Slikah 5, 6, 7 Vsaka slika kaže primerjavo sistemov za določen jezikovni par. Standardna deviacija kaže odstopanja vrednotenja posameznih povedi. Večje vrednosti predstavljajo boljše rezultate.

Tabela 2 kaže zadostno mero ujemanja do zelo veliko mero ujemanja med ocenjevalci (satisfactory to very high inter-rater agreement) po Cohenovem koeficientu kappa (Cohen, 1960).



Slika 5: Rezultati evalvacije sistemov jezikovnega para SL-EN po smernicah (LDC, 2005). Sistem Presis dosega rezultate primerljive z Googlom, sistem Bing dosega signifikantno slabše rezultate.



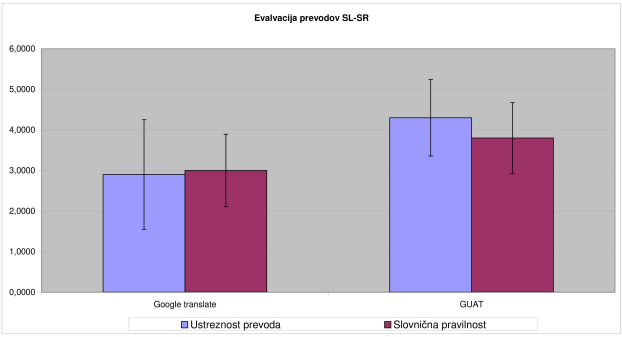
Slika 6: Rezultati evalvacije sistemov jezikovnega para EN-SL po smernicah (LDC, 2005). Sistem Presis dosega rezultate primerljive z Googlom, sistem Bing dosega signifikantno slabše rezultate.

5. Zaključek in nadaljnje delo

Začetni načrt evalvacije je obsegal vse opisane metode evalvacije, čeprav smo se zavedali neprimernosti metrike BLEU (Papineni et al., 2001) za primerjavo različnih prevajalnih sistemov. Veliko avtorjev se strinja, da metrika BLEU sistematično zapostavlja sisteme RBMT, kot na primer (Callison-Burch et al., 2006; Labaka et al., 2007), prav tako metrika ni primerna za visoko pregibne jezike. Dva od predstavljenih sistemov spadata v skupino prevajalnih sistemov paradigme RBMT in sta se pri tej metriki zelo slabo odrezala. Evalvacijo z BLEU metriko smo vseeno uvrstili v eksperiment zaradi zgodovinske primerljivosti rezultatov. Druga uporabljena samodejna metrika METEOR je po priporočilu avtorjev veliko primernejša za visoko pregibne jezike in odpravlja največje pomanjkljivosti metrike BLEU, rezultati še vedno postavljajo Googlov prevajalni sistem na prvo mesto, vendar z manjšo razliko.

Ročni metodologiji, ki veliko bolje odražata dejansko kakovost prevodov, kažeta veliko boljše rezultate sistemov RBMT, vendar moramo upoštevati manjše število testnih primerov ter na možno neobjektivnost evalvatorjev, čeprav je bilo ujemanje med evalvatorji zadovoljivo do odlično.

Primerjava sistemov jezikovnih kombinacij SL-EN ter EN-SL kaže na veliko slabšo kakovost prevodov v iz angleščine



Slika 7: Rezultati evalvacije sistemov jezikovnega para SL-SR po smernicah (LDC, 2005). Sistem GUAT dosega signifikantno boljše rezultate kot Google.

Tabela 2: Cohenov koeficient kappa (Cohen, 1960) za posamezne sisteme kaže zadostno mero ujemanja do zelo veliko mero ujemanja med ocenjevalci. Prikazani so samo podatki primerjave ujemanja ocenjevalcev sistemov jezikovnega para sl-en

	Presis	Bing	Google
kappa	0,86	0,69	0,78
pričak. naklj. ujemanje	0,300	0,317	0,305
št. primerov	100	100	100

v slovenščino v primerjavi s prevodi iz slovenščine v angleščino. Podajamo možno razlago, ki pa pojasni le del problema: statistični prevajalni sistemi za izbiro končnih kandidatov uporabljajo jezikovne modele ciljnih jezikov, jezikovni model za slovenščino je naučen na veliko manjši učni množici kot jezikovni model za angleščino, učno množico pogojuje število dostopnih besedil v elektronski obliki. Ta razlaga je brezpredmetna pri sistemu Presis, ki temelji na pravilih.

Pregled osnovnih napak sistemov: glavna napaka statističnih sistemov je problem lokalnega ujemanja besed v leksikalnih kategorijah. Kot primer navedimo ujemanje samostalnikov ter pridevnikov v spolu, sklonu ter številu. Ta problem je še posebej izrazen v slovanskih jezikih. Pri sistemih paradigme RBMT je opazno manjše besedišče. Ta problem je še posebej prisoten pri sistemu GUAT, ki ima dvojezični slovar ter slovar ciljnega jezika (srbščine) izluščen iz relativno majhnega korpusa. Sistemi za strojno prevajanje se nenehno izboljšujejo, primerjava prikazana v tem članku bo ob rednih podobitvah dostopna na naslovu: `http://jt.upr.si/mt_v_sloveniji/`.

6. Literatura

Chris Callison-Burch, Miles Osborne, in Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. V: *Proceedings of EACL*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Antonio M. Corbi-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gemma Ramirez-Sanchez, Felipe Sanchez-Martinez, Inaki Alegria, Aingeru Mayor, in Kepa Sarasola. 2005. An open-source shallow-transfer machine translation engine for the Romance languages of Spain. V: *Proceedings EAMT conference*, str. 79–86, May.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevič, Tomaž Erjavec, Heiki Jaan Kaalep, in Dan Tufis. 1998. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. V: *COLING-ACL*, str. 315–319.

Gorka Labaka, Nicholas Stroppa, Andy Way, in Kepa Sarasola. 2007. Comparing rule-based and data-driven approaches to Spanish-to-Basque machine translation. V: *Proceedings of the Machine Translation Summit XI*, str. 41–48.

A. Lavie in A. Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. V: *Proceedings of Workshop on SMT at the ACL conference*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Tehnično poročilo, LDC.

V. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, str. 845–848.

Corpora List. 2010. Corpora list.

NIST. 2008. Evaluation software.

Kishore Papineni, Salim Roukos, Todd Ward, in Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Tehnično poročilo, IBM.

Chris Quirk, Arul Menezes, in Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. V: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*.

Miro Romih in Peter Holožan. 2002. A slovenian-english translation system. V: *Proceedings of the 3rd Language Technologies Conference*, str. 167.

Jorg Tiedemann. 2007. Improved sentence alignment for movie subtitles. V: *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'07)*, Borovets, Bulgaria.

Jörg Tiedemann. 2009. News from opus - a collection of multilingual parallel corpora with tools and interfaces. *Recent Advances in Natural Language Processing*, 5:237–248.

Jernej Vičič in Petr Homola. 2010. Speeding up the implementation process of a shallow transfer machine translation system. V: *Proceedings of the 14th EAMT Conference*, str. 261–268, Saint Raphael, France. European Association for Machine Translation.

Jernej Vičič, 2009. *Metode hitre izdelave gradiv za prevajalne sisteme plitkega prenosa za visoko pregibne jezike*, str. 133–153. Znanstveno-raziskovalno središče, Založba Annales.

Jerneja Žganec Gros, France Mihelič, Tomaž Erjavec, in Špela Vintar, 2005. *The VoiceTRAN Speech-to-Speech Communicator*, str. 379–384. Springer Berlin / Heidelberg.

Uporaba wordneta za boljše razdvoumljanje pri strojnem prevajanju

Darja Fišer, Špela Vintar

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@guest.arnes.si
spela.vintar@ff.uni-lj.si

Povzetek

V prispevku predstavljamo rezultate raziskave, v kateri smo s pomočjo slovenskega in angleškega wordneta skušali izboljšati strojno prevajanje na leksikalni ravni, kjer je problematično predvsem ustrezno prevajanje večpomenskih besed in večbesednih zvez. Za potrebe raziskave smo najprej izdelali vzporedni korpus novic EU, ki smo ga avtomatsko semantično označili in razdvoumili s pomočjo angleškega in slovenskega wordneta. Ker ta dva semantična leksikona uporabljata iste identifikacijske kode za koncepte, ju je mogoče uporabiti tudi za iskanje prevodnih ustreznic. Neodvisno od semantičnega označevanja smo korpus prevedli s strojnim prevajalnikom, prevode, ki jih je predlagal, pa primerjali s prevodnimi ustreznicami, pridobljenimi iz wordneta, pri čemer smo za referenčni prevod uporabili človeški prevod iz vzporednega korpusa. Raziskavo smo izvedli za obe jezikovni kombinaciji in pri tem uporabili dva različna strojna prevajalnika. Rezultati so spodbudni, saj smo s pomočjo wordneta v obeh smereh in pri obeh uporabljenih prevajalnikih izboljšali precej prevodov, kar še posebej velja za prevajanje iz angleščine v slovenščino s prevajalnikom Presis.

Using wordnet for better word-sense disambiguation in machine translation

This paper presents the results of an experiment in which we tried to improve machine translation at lexical level with wordnets for Slovene and English. Mistranslations often arise due to inadequate word-sense disambiguation of polysemous words and detection of multi-word expressions, and we believe that wordnet can help with both. First we created a parallel corpus of EU news, which was then automatically annotated with disambiguated wordnet ids. Because Slovene and English wordnets use the same synset ids, they can be used for finding translation equivalents. Independently of word-sense disambiguation, we machine-translated the corpus and then compared the translations with translation equivalents obtained from wordnet while human translation from the parallel corpus was used as a reference. The experiment was conducted in both language directions and with two different machine translation systems. The results are encouraging as we were able to improve the translations in all experimental settings, especially in translation from English into Slovene with the machine translation system Presis.

1. Uvod

Z naraščanjem količine medjezikovne komunikacije besedil narašča tudi potreba po strojnem prevajanju. Predvsem za kombinacije z večjimi svetovnimi jeziki sodobni strojni prevajalniki že omogočajo prevode, ki so povsem sprejemljivi v okoliščinah, kjer prevajamo besedila z zelo omejenega področja in so napisana v nadzorovanem jeziku, zadovoljivo uporabno vrednost pa imajo tudi prevodi splošnejše narave, za katere natančni prevod niti ni potreben ali pa si ga zaradi finančnih, kadrovskih ali časovnih omejitev preprosto ne moremo privoščiti. Kljub temu pa strojni prevodi še zdaleč niso popolni; poleg neustreznih slovničnih konstrukcij v generiranih prevodih pogosto nastopajo težave na leksikalni ravni, kjer je uporabljen sicer legitimen prevod neke besede, vendar v napačnem pomenu, kjer so prevedeni posamezni deli večbesednih zvez namesto cele zveze, kjer besede ostajajo neprevedene, ker manjkajo v prevajalnikovem leksikonu, in podobno.

S tem problemom se ukvarjamo v pričujoči raziskavi, s katero želimo pokazati, da je mogoče rezultate strojnega prevajanja bistveno izboljšati z uporabo wordneta, in sicer predvsem pri razdvoumljanju večpomenskih besed, pa tudi pri iskanju prevodov za večbesedne ustreznice in kot dodaten dvojezični vir za iskanje prevodnih ustreznic, predvsem strokovno specifičnih izrazov. Wordnet so kot spodnji del ontologije učinkovito uporabili v strojnem prevajalniku PANGLOSS (Knight 1993). Pri strojnem prevajanju korejskih besedil so se s pomočjo latentne semantične analize in semantične podobnosti besed v wordnetu spopadli s prevajanjem kolokacij (Yuseop idr. 2002), za razdvoumljanje večpomenskih izrazov v izvirniku in izbiro ustreznega prevoda pa so ga uporabili pri statističnem prevajanju iz angleščine v bengalščino (Salam idr. 2009). Wordnet so uporabili tudi za prevajanje samostalniški zvez iz arabščine v angleščino (Ali idr. 2009). Te so pri prevajanju zelo pomembne, saj nosijo večino informacij v stavku. Prevodne ustreznice zanje so poiskali s semantičnim razdvoumljanjem, ki temelji na iskanju najustreznejšega semantičnega grafa za prevajani stavek na zelo podoben način, kot to počnemo v naši raziskavi, ki pa ni omejena zgolj na samostalnike.

Razdvoumljanje za potrebe strojnega prevajanja se razlikuje od klasičnega avtomatskega razdvoumljanja, saj za prevajanje zadošča stopnja razdvoumljanja, ki v ciljnem jeziku zahteva različne prevode, na kar so opozorili že Vickrey idr. (2005). Nesporen doprinos razdvoumljanja in leksikalne semantike h kvaliteti strojnega prevajanja pa sta z empiričnimi raziskavami strojnega prevajanja fraz in ovrednotenjem rezultatov dokazala Carpuat in Wu (2007), s čimer sta dokončno ovrgla trditve nekaterih avtorjev, ki so trdili nasprotno. Na leksikalni ravni ostajamo tudi v tej raziskavi, pri čemer je naš namen pokazati učinkovitost vključevanja wordneta in semantičnega razdvoumljanja v proces strojnega prevajanja, ne pa razvoj strojnih prevajalnikov samih, zato tudi ne generiramo prevoda celotnega stavka.

V prispevku najprej predstavljamo korpusne in slovarske vire, ki smo jih uporabili v raziskavi. Sledi opis semantičnega označevanja korpusa in razdvoumljanja, nato pa še strojnega prevajanja. V petem razdelku analiziramo in ovrednotimo rezultate raziskave, prisevek pa sklenemo z razpravo in načrti za prihodnje delo.

2. Priprava virov

V pričujoči raziskavi smo uporabili angleško-slovenski vzporedni korpus novic EU, ki smo ga sestavili in obdelali sami, ter že obstoječa angleški in slovenski wordnet, ki smo ju zgolj preoblikovali v ustrezen format za potrebe semantičnega razdvoumljanja.

2.1. Vzporedni korpus

Raziskavo smo opravili na vzporednem slovensko-angleškem korpusu, ki smo ga sestavili iz novic s portala EU[1]. Korpus vsebuje nekaj več kot 500 novic o najrazličnejših temah, kot so okolje, kultura, znanost in druge (glej Tabelo 1), ter obsega približno 120.000 pojvnic v slovenščini in 140.000 pojavnic v angleščini.

Tema	Št. novic
okolje	83
delovanje EU	64
zaposlovanje	50
kultura	48
ekonomija	47
mednarodni odnosi	46
znanost	33
pravo	32
podjetništvo	26
transport	21
kmetijstvo	20
energija	20
regionalna politika	16
skupaj:	506

Tabela 1: Število novic v korpusu glede na tematiko.

Stavčno poravnavo korpusa smo opravili z orodjem WinAlign[2]. Slovenski del korpusa smo tokenizirali, oblikoskladenjsko označili in lematizirali s pomočjo spletnega servisa JOS (Erjavec idr. 2010), angleškega pa z orodjem ToTaLe (Erjavec idr. 2005). Nato smo iz korpusa izločili vsa ločila in slovnične besede, da so v njem ostale zgolj leme samostalnikov, glagolov, pridevnikov in prislovov s pripisano besedno vrsto. Pojavnice smo v vsaki novici posebej oštevilčili in jim pripisali oznako 1, če obstajajo v korpusu in jih želimo semantično označiti, oz. oznako 0, če temu ni tako. V tej fazi smo v korpusu označili tudi večbesedne zveze, ki jih najdemo tudi v wordnetu. Slika 1 vsebuje izsek iz slovenskega dela korpusa, ki je pripravljen za semantično označevanje.

```
ctx_061211
pregledati#v#74#1 biti#v#75#1 znesek#n#76#0
dodeljen#a#77#0 leto#n#78#1 oceniti#v#79#1
napredek#n#80#1 eu#n#81#1 izpolnjevanje#n#82#1
cilj#n#83#1 gospodarski_rast#n#84#1 nov#a#85#1
deloven_mesto#n#86#1
```

Slika1. Primer iz korpusa.

2.2. Wordnet

Za določanje pomena besedam v korpusu smo uporabili wordnet za angleški in slovenski jezik. Wordnet je leksikalna podatkovna zbirka, ki vsebuje samostalnike, glagole, pridevnike in prislove. Zbirka je zasnovana pojmovno, kar pomeni, da so v njej vse besede, ki označujejo isti pojem, združene v sopomenske nize oziroma sinsete (npr. *luč* in *svetilka*), ti pa so nato med seboj povezani s semantičnimi relacijami, kot so nad- in podpomenskost, holo- in meronimija, protipomenskost in druge.

Posamezno sopomenko v sinsetu imenujemo literal, ki se v različnih pomenih lahko pojavlja v več sinsetih (npr. *jezik* kot sredstvo komunikacije, *jezik* kot organ, *jezik* kot del čevlja). Poleg literalov in semantičnih relacij vsak sinset vsebuje tudi razlago in primere rabe, številni pa še področno oznako in druge informacije.

Angleški wordnet[3] je prvi tovrsten semantični leksikon, ki je začel nastajati pred dobrima dvema desetletjema na Univerzi v Princetonu (Fellbaum 1998). Princeton WordNet (PWN) je še danes največji, saj vsebuje približno 155.000 različnih literalov oz. 117.000 sinsetov.

Slovenski wordnet[4] (sloWNet) je veliko mlajši in manjši, izdelan pa je bil s polavtomatskimi metodami (Fišer 2009) na podlagi PWN in trenutno vsebuje petino angleških sinsetov, kar številčno pomeni približno 17.000 sinsetov oz. 20.000 literalov. Slovenski wordnet za pojme v semantični mreži uporablja iste identifikacijske kode kot angleški, zato ju lahko uporabljamo tudi kot dvojezični vir, ki omogoča iskanje prevodnih ustreznic v obeh jezikih, kar je ključno za našo raziskavo.

Kot je razvidno iz Tabele 2, primerjava besedišča v slovenskem delu korpusa in sloWNetu pokaže, da jih od nekaj več kot 116.000 pojavnic v korpusu skoraj 85.000 oz. 73 % najdemo tudi v sloWNetu (med temi je okoli 3.300 različnic), med katerimi je tudi nekaj manj kot 2.500 večbesednih leksemov. Več kot en pomen v sloWNetu ima približno 56.700 iztočnic, pri čemer so skoraj vse enobesedne.

V angleškem delu korpusa, ki šteje nekaj nad 143.000 pojavnic, se s Princeton WordNetom prekriva skoraj 100.000 oz. 70 % pojavnic (med temi je nekaj več kot 8.000 različnic), med katerimi je okoli 4.600 večbesednih leksemov. Večpomenskih je 82.300 različnic, med katerimi je prav tako večina enobesednih.

	ang	slo
št. pojavnic	99.809	84.905
št. različnic	8.015	3.289
št. večbesednih	4.651	2.478
št. večpomenskih	82.269	56.697
št. enobesednih večpomenskih	80.865	56.574

Tabela 2: Prekrivanje korpusa in wordnetov.

[1] http://ec.europa.eu/news/
[2] http://www.translationzone.com/en/
[3] http://wordnet.princeton.edu/
[4] http://nl.ijs.si/slownet/

Za potrebe semantičnega označevanja smo wordneta preoblikovali v bazo znanja v predpisanem formatu, pri čemer smo normalizirali literale v male tiskane črke in lematizirali večbesedne literale, da jih bo mogoče prepoznati v prav tako lematiziranem korpusu.

3. Semantično označevanje korpusa in razdvoumljanje

Cilj semantičnega označevanja korpusa je, da večpomenskim pojavnicam v korpus pripišemo ustrezen pomen, na podlagi katerega bo nato prek medjezikovne povezave v slovenskem in angleškem wordnetu mogoče najti njene prevodne ustreznice. Ker želimo razviti metodologijo, primerno za vse vrste besedil in različne jezikovne kombinacije, smo za vsak jezik posebej pojavnicam v korpusu pomene iz wordneta pripisali avtomatsko.

Za to smo uporabili prosto dostopno orodje za razreševanje večpomenskosti UKB (Agirre in Soroa 2009), ki za določanje pomena besede glede na sobesedilo, v katerem se pojavlja, upošteva relacije med sinseti v wordnetu. Iz množice grafov, ki se za besedo in njeno sobesedilo v semantični mreži izdelajo, program izbere tisti pomen, ki je na podlagi števila in moči povezav med pojmi v mreži ocenjen z najvišjo stopnjo verjetnosti. S tem orodjem je mogoče razdvoumljati eno- in večbesedne samostalnike, glagole, pridevnike in prislove, ki jih wordnet vsebuje. Poleg tega, da je UKB mogoče uporabiti za vse jezike, za katere obstajajo semantični leksikoni tipa wordnet, je njegova prednost tudi v tem, da pomene vseh besed v istem sobesedilu razreši hkrati, zato je postopek pripisovanja pomenov zelo hiter.

Rezultat označevanja je seznam pojavnic iz korpusa, ki so jim pripisane identifikacijske kode izbranih sinsetov. Za lažjo primerjavo s strojnimi prevodi v naslednjem koraku smo za pripisane semantične oznake izpisali še ustrezne literale iz slovenskega in angleškega wordneta. Primer semantično razdvoumljenega korpusa prikazuje Slika 2.

```
UKB: ctx_090529 1 oznaka ENG20-06824154-n !!
        SWN: oznaka, znak, znamenje
        PWN: marker, marking, mark
UKB: ctx_090529 2 živilo ENG20-00018827-n !!
        SWN: hranilo, živilo
        PWN: food, nutrient
UKB: ctx_090529 7 prihodnost ENG20-14265057-n !!
        SWN: bodočnost, prihodnost
        PWN: future, hereafter, futurity, time to come
```

Slika 2. Semantično razdvoumljene pojavnice v korpusu s slovenskimi in angleškimi literali.

4. Strojno prevajanje korpusa

Povsem neodvisno od semantičnega označevanja smo korpus strojno prevedli. Ker namen te raziskave ni ocenjevanje kvalitete slovenskega wordneta ali avtomatskega semantičnega označevanja za slovenščino, temveč razvoj metodologije za izboljšavo strojnih prevodov s pomočjo wordneta na splošno, smo se odločili postopek preizkusiti v obeh smereh. Zato smo korpus prevedli iz angleščine v slovenščino in obratno.

Za objektivnejšo primerjavo rezultatov smo uporabili dva različna strojna prevajalnika: Presis,[5] ki temelji na ročno napisanih pravilih, in statistični prevajalnik GoogleTranslate,[6] ki se jezikovnih modelov uči iz velike količine besedil, dostopnih na svetovnem spletu. Presis izvorni stavek najprej analizira in prevede v vmesni jezik, nato pa na polagi analize generira stavek v ciljnem jeziku. Pri tem uporablja ročno napisana pravila, dvojezični leksikon in druge jezikovne informacije iz zbirke Ases, ki vključuje tudi ločevanje med posameznimi pomeni besed (Holozan 2008). Presisov prevod celotnega korpusa v obe smeri je prijazno zagotovilo podjetje Amebis.

Googlov prevajalnik uporablja vzporedna besedila na spletu, denimo nekdanje dokumente Združenih narodov in Evropske komisije, v katerih je »identična vsebina profesionalno prevedena v mnogo jezikov« (Softky 2007). Osnova za razvoj statističnega prevajalnika je bil šestjezični korpus z 20 milijardami besed, s pomočjo katerega je Google izpilil statistične algoritme za učenje jezikovnih modelov, pri izboljševanju prevoda – denimo besednega reda ali izbiri ustreznice glede na kontekst – pa se opira še na gigakorpuse besedil v ciljnem jeziku. Googlov prevajalnik ima vgrajeno avtomatsko vrednotenje prevoda, pomaga pa si še s funkcijo »Prispevaj boljši prevod«, kjer skuša jezikovne modele izboljševati s pomočjo povratnih informacij uporabnikov.

Strojno prevedena besedila smo nato lematizirali in jih na podlagi oznak za segmente, ki smo jih pridobili pri stavčni poravnavi korpusa, primerjali s semantično razdvoumljenim izvirnikom in prevodnimi ustreznicami, pridobljenimi preko medjezikovnih povezav iz wordneta. Pri primerjavi nas je zanimalo, ali sta strojni prevod in prevod s pomočjo wordneta ustrezna. To smo ugotavljali s primerjavo z referenčnim človeškim prevodom iz vzporednega korpusa.

5. Rezultati in vrednotenje

5.1. Prevajanje iz slovenščine v angleščino

V slovenskem delu korpusa smo zaznali nekaj čez 38.000 večpomenskih pojavnic, kar je približno 4.000 večpomenskih različnic. Primerjava strojnega, wordnetovega in referenčnega prevoda iz slovenščine v angleščino pokaže, da se v približno 40 % primerov wordnetov predlog ujema z referenčnim prevodom. Od tega se Presis pri 1.558 primerih (434 različnih besedah) ne ujema z referenčnim prevodom in bi ga lahko s pomočjo wordneta in semantičnega razdvoumljanja izboljšali. Google je po teh merilih malce boljši, saj se z referenčnim prevodom ne ujema 867 primerov (368 različnih besed), kjer bi strojni prevod lahko izboljšali.

Po en primer napačnega angleškega prevoda obeh prevajalnikov, ki bi ju s pomočjo wordneta bilo mogoče izboljšati, vsebuje Slika 3. V prvem primeru je besedo *koza* Presis prevedel z napačnim pomenom *smallpox*, medtem ko smo z avtomatskim razdvoumljanjem s pomočjo wordneta prišli do pravilnega prevoda *goat*, ki ga vsebuje tudi referenčni človeški prevod iz korpusa. Drugi primer je za besedo *napoved*, ki jo je narobe prevedel Google, in sicer z izrazom *announcement*, medtem ko bi z uporabo wordneta besedo pravilno prevedli v *forecast.*

[5] http://presis.amebis.si/prevajanje/

[6] http://translate.google.com/

	WORDNET = KORPUS	WORDNET ≠ KORPUS
WORDNET = GOOGLE	1.527	1.194
WORDNET ≠ GOOGLE	368	1.544
WORDNET = PRESIS	1.343	1.211
WORDNET ≠ PRESIS	434	1.082
ujemanje v %	40,90	59,10

Tabela 3: Ujemanje med prevodi različnic za smer slovenščina-angleščina.

WORDNET: 081114 koza → **goat**, caprine animal
PRESIS: Almost 360 million of pigs, sheep, a **smallpox** and cattle and more billion of poultry execute every year in European Union because of meat.
KORPUS: Every year nearly 360 million pigs, sheep, **goats** and cattle and several billion poultry are killed for their meat in the EU.

WORDNET: 090119 napoved → prognosis, **forecast**
GOOGLE: But because of the sharp economic slowdown, the interim **announcement** has been expanded.
KORPUS: But in light of the sharp economic slowdown, the current interim **forecast** has been expanded.

Slika 3. Primer uspešne uporabe wordneta za izboljšanje strojnega prevajanja iz slovenščine v angleščino.

Ročni pregled naključno izbranih 200 primerov, kjer se prevod z wordnetom ne ujema niti z referenčnim niti s strojnim prevodom, pokaže, da je med njimi 73 takih primerov, ki so pravilno razdvoumljeni, le 127 pa takih, kjer je pojavnicam v korpusu glede na sobesedilo pripisan napačen pomeni. To se pravi, da je med negativnimi primeri dobra tretjina še vedno pravilno razdvoumljena. Razlogi za neujemanje so številni: v prevodu je uporabljen sinonim, v prevodu je uporabljen isti besedni koren v drugi besedni vrsti, v prevodu je opazovana beseda izpuščena, prevod je svobodnejši. Taka sta tudi primera v spodnji sliki; prvi ponazarja izpust besede *vrsta* v referenčnem prevodu, drugi pa svobodnejši prevod za *spomin*:

WORDNET: 080729 vrsta → species
PRESIS: Their task was to make impossible immoderate catch of Atlantic cod, that cold waters of this sea were full of her sometimes, this kind is disappearing quickly now.
KORPUS: They were on a mission to stop the overfishing of cod, once plentiful in the icy waters, but now fast disappearing.

WORDNET: 090423 spomin → memory, retention
PRESIS: Prelov fish is le yet moved away memory.
KORPUS: Rampant overfishing is a thing of the past.

Slika 4. Primer razhajanja med wordnetovim, strojnim in referenčnim prevodom zaradi ohlapnejšega prevoda.

Zanimivo je, da je v ročno pregledanem vzorcu samo ena beseda razdvoumljena dvakrat različno: *meso → meat* in *meso → flesh, pulp*, pa še to obakrat narobe; prvič je bilo meso v zvezi *goveje meso*, za katero se pravilni prevod glasi *beef*, drugič gre za opis revščine, ko si družina ne more privoščiti mesa.

Wordnet lahko pomembno izboljša tudi prevajanje večbesednih enot. V korpusu smo našli kar 166 primerov, kjer je bil strojni prevod pomanjkljiv zaradi neustrezne obravnave stalne besedne zveze, medtem ko je wordnet predlagal pravilno ustreznico; npr. *biotska raznovrstnost – biotic diversity* namesto *biodiversity*; *vezani les – tied wood* namesto *plywood*.

5.2. Prevajanje iz angleščine v slovenščino

V angleškem delu korpusa smo zaznali približno 48.000 večpomenskih pojavnic oz. okoli 6.000 večpomenskih različnic, ki hkrati obstajajo tudi v slovenščini, kar je nekoliko več kot v slovenskem delu korpusa. Zanimivo je, da je skupni odstotek ujemanja wordnetovega prevoda z referenčnim nižji – okrog 32%. Vendar je po drugi strani več primerov, ki jih lahko z uporabo wordneta izboljšamo: pri Presisu je takšnih kar 3.730 primerov (865 različnih besed), pri Googlu pa 901 primerov (485 različnih).

	WORDNET = KORPUS	WORDNET ≠ KORPUS
WORDNET = GOOGLE	1.491	1.099
WORDNET ≠ GOOGLE	485	3.014
WORDNET = PRESIS	1.277	1.403
WORDNET ≠ PRESIS	865	2.454
ujemanje v %	32,45	67,55

Tabela 4: Ujemanje med prevodi različnic za smer angleščina-slovenščina.

Ročno vrednotenje naključno izbranih 200 primerov pokaže, da je med negativnimi primeri kar polovica pravilno razdvoumljenih (99 pravilno : 101 napačno). Ta rezultat je precej boljši od razdvoumljanja v slovenščini, kar je zanimivo, glede na to, da smo pri obeh jezikih uporabili isto semantično mrežo. Če primere pogledamo pobliže, opazimo, da so glagoli veliko slabše razdvoumljeni od samostalnikov (npr. dosledna napaka *be → živeti*, namesto *biti*), vendar je pri tem treba poudariti, da je razdvoumljanje glagolov veliko težje kot razdvoumljanje samostalnikov. Poleg tega je bil sloWNet izdelan avtomatsko, po končani gradnji pa so bili samostalniki pregledani in popravljeni, glagoli pa ne, zato je mogoče, da so vzrok za te napake tudi neustrezno prevedeni sinseti oz. manjkajoči pomeni za nekatere večpomenske glagole.

Primera, v katerih je bilo razdvoumljanje in prevajanje v slovenščino s pomočjo wordneta uspešnejše od strojnega, vsebuje Slika 5. V prvem primeru je Presis z besedo *prikuha* narobe prevedel izraz *vegetable*, medtem ko smo z wordnetom dobili ustrezen predlog *zelenjava*. V drugem primeru pa Presis za glagol *wait*, ki smo ga z wordnetovo pomočjo uspeli prevesti pravilno s *čakati*, sploh ni predlagal nobenega prevoda in je glagol pustil v izvirniku, najverjetneje zato, ker ni uspela analiza izvornega stavka.

WORDNET: 090529 vegetable → **zelenjava**
PRESIS: on biti že obvezen za nekaj hrana prodati v evropski unija , including neobdelan govedina , perutnina , sadje , **prikuha** , jajce , med , vino in oliven olje .
KORPUS: oznaka z kraj pridelava biti v eu za nekateri živilo že obvezen , deti za goveji meso , perutnina , sadje , **zelenjava** , jajce , med , vino , oliven olje .

WORDNET: 090423 wait → počakati, **čakati**
PRESIS: normalno sedanji politika - zadnji overhauled v 2002 - biti ne izteči se za recenzijski do 2012. ampak biti položaj postajati preveč negotov do **wait** that dolgo .
KORPUS: sicer biti biti naslednji pregled zdajšnji politik - zadnji pregled biti biti leto 2002 - predviden šele leto 2012 , toda stanje v ribištvo biti tako kritičen , da ne smeti veliko **čakati** .

Slika 5. Primer uspešne uporabe wordneta za izboljšanje strojnega prevajanja iz angleščine v slovenščino.

6. Zaključek

V prispevku smo predstavili poskus izboljšave strojnih prevodov na leksikalni ravni s pomočjo semantičnega razdvoumljanja in iskanja prevodnih ustreznic v angleškem in slovenskem wordnetu. Analiza in vrednotenje rezultatov sta pokazala, da je z uporabo že obstoječih jezikovnih virov in tehnologij na enostaven in hiter način v številnih primerih mogoče učinkovito ugotoviti pravi pomen, s tem pa tudi prevod večpomenskih izrazov, s čimer imajo strojni prevajalniki dandanes še precej težav, ne glede na to, ali za prevajanje uporabimo sisteme, ki temeljijo na ročno napisanih pravilih, ali statistične prevajalnike.

Čeprav smo se v eksperimentu spopadali z napakami v prevodih, za katere razlog tiči v napačnem oblikoskladenjskem označevanju in lematizaciji ali napačnem avomatskem razdvoumljanju večpomenskih besed, smo s pomočjo wordneta zabležili izboljšave tako pri uporabi Presisa kot Googlovega prevajalnika v obeh jezikovnih kombinacijah, še največ jih je bilo pri Presisovem prevodu v slovenščino.

Dejstvo, da je Google Translate v obeh smereh dosegel boljše rezultate, bi lahko utemeljili s tem, da smo v korpus vključili vzporedna besedila, objavljena na svetovnem spletu, in to s področja delovanja EU. Ker Googlov prevajalnik svoje jezikovne modele gradi prav s pomočjo tovrstnih vzporednih besedil, lahko upravičeno pričakujemo, da se bo pri njihovem prevajanju zelo dobro odrezal.

Pri vrednotenju rezultatov se je izkazalo, da je ustreznost prevodov zelo težko vrednotiti avtomatsko, kljub temu, da smo imeli na voljo človeški referenčni prevod. Ta je namreč velikokrat svobodnejši in namesto sicer pravilno razdvoumljenega izraza vsebuje njegovo približno sopomenko, izraz v spremenjeni besedni vrsti, izraz preprosto izpusti, ali pa je bolj idiomatski od izvirnika. Zato je pri tovrstnih preizkusih nujen tudi ročni pregled vsaj vzorca dobljenih rezultatov, s čimer dobimo natančnejši vsebinski vpogled v prevode.

V okviru bilateralnega projekta z madžarsko akademijo znanosti nameravamo v prihodnje raziskavo razširiti še na jezikovni par angleščina-madžarščina in postopek preizkusiti na prevajalnem sistemu MorphoLogic. S tem bomo dobili objektivnejšo potrditev učinkovitosti in jezikovne neodvisnosti metode, predstavljene v prispevku.

7. Literatura

Ali, Ola Mohammad, Mahmoud Gad Alla in Mohammad Said Abdelwahab (2009): Improving machine translation using hybrid dictionary-graph based word sense disambiguation with semantic and statistical methods. International Journal of Computer and Electrical Engineering, 1/5.

Carpuat, Marine in Dekai Wu (2007): Improving statistical machine translation wusing word sense disambiguation. Zbornik konference Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).

Eneko Agirre in Aitor Soroa (2009): Personalizing PageRank for Word Sense Disambiguation. Zbornik 12. konference European chapter of the Association for Computational Linguistics (EACL'09).

Erjavec, Tomaž, Darja Fišer, Simon Krek in Nina Ledinek (2010): The JOS Linguistically Tagged Corpus of Slovene. Zbornik 7. mednarodne konference Language Resources and Evaluation (LREC'10), Valletta, Malta.

Fellbaum, Christiane (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Fišer, Darja (2009): Pristopi za avtomatizirano gradnjo semantičnih zbirk. V: Nina Ledinek, Mojca Žagar Karer, Marjeta Humar (ur.) Terminologija in sodobna terminografija. Ljubljana: Založba ZRC, ZRC SAZU.

Holozan, Peter (2008): Samodejno luščenje slovarja iz vzporednega korpusa s pomočjo vmesnega jezika in pomenskega razdvoumljanja. Zbornik Šeste konference Jezikovne tehnologije.

Knight, Kevin (1993): Building a large ontology for machine translation. Zbornik delavnice ARPA Human Language Technology Workshop.

Salam, Khan Md. Anwarus, Mumit Khan in Tetsuro Nishino (2009): Example based English-Bengali machine translation using wordnet. Zbornik konference TriSA'09.

Softky, Bill (2007): How Google translates without understanding.http://www.theregister.co.uk/2007/05/15/google_translation/page2.html (23. 6. 2010)

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis in Daniel Varga (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Zbornik 5. mednarodne konference Language Resources and Evaluation (LREC'06).

Vickrey, David, Luke Biewald, Marc Teyssier in Daphne Koller (2005): Word-Sense Disambiguation for Machine Translation." Zbornik konference Empirical Methods in Natural Language Processing (EMNLP).

Yuseop, Kim, Jeong-Ho Chang in Byoung-Tak Zhang (2002): Target Word Selection Using WordNet and Data-Driven Models in Machine Translation. Zbornik konference PRICAI'02: Trends in Artificial Intelligence.

Ontologije ali semantične mreže kot obogatitev terminologije

Jasna Belc

Generalni sekretariat Vlade Republike Slovenije
Sektor za prevajanje; Gregorčičeva 20, Ljubljana
Jasna.Belc@gov.si

Povzetek

Članek govori o zamisli projekta uvedbe in oblikovanja slovenskih (strokovnih: pravnih) ontologij ali semantičnih mrež kot vira, obogatitve ali nadgradnje terminološke zbirke pravnih izrazov. Sestavljanje ontologij nasploh je v zadnjem času zelo razširjena disciplina v evropskih in svetovnih projektih računalniškega jezikoslovja. Prikaz T. Vealeja (14) je obraten proces bogatenja ontologij, tj. taksonomsko urejenih sistemov besednih/konceptualnih kategorij, predvsem iz slovarskih in enciklopedičnih virov (npr. wikipedia). Slovenščina nima na voljo dovolj slovarskih, enciklopedičnih, tezaverskih virov v elektronski obliki s prostim dostopom, če jo primerjamo z angleščino in nekaterimi evropskimi in drugimi svetovnimi jeziki. Naš projekt želi sestaviti osnovno zbirko pravnih ontologij (osnovnih pravnih konceptov) v smislu ontološke vede, glede na razmerja, v kakršnih ti pojmi dejansko obstajajo v slovenskem pravnem jeziku. Opira se zlasti na obstoječe (večjezične) elektronske in delno na tiskane vire pravnih in drugih ontologij evropskega prava (npr. JurWordNet, Legal Taxonomy Syllabus ipd.), kakor tudi na programe za gradnjo le-teh ter ustrezne že obstoječe besedilne vire (Eur-lex, slovenska zakonodaja, SVEZ-IJS Acquis Corpus, Evrokorpus ipd.).
Ključni izrazi: ontologije, semantične mreže, leksikalne taksonomske zbirke, terminologija, korpusi področnih besedil (pravni strokovni jezik), računalniško podprto sestavljanje.

Abstract

The Article discusses the start of a project to set up and develop Slovene ontologies or semantic webs (in the field of law) in order to source terms for and enrich and upgrade legal terminological databases. Recently, ontology building has been a very popular discipline in computational linguistics projects Europe- and world-wide. The aim of the present project takes the opposite approach to T. Veale's (14), in which it is primarily lexical and encyclopaedic sources (such as Wikipedia) that constitute a major source in the process of building up ontologies or taxonomically ordered systems of lexical/conceptual categories. Compared with English or major European or world languages, Slovene has few large dictionaries, encyclopaedias, thesauruses, etc., accessible in electronic form online. In the first phase, our project is aimed at creating a basic ontological database (of fundamental legal concepts) in the scientific sense, with reference to the actual relationships between the concepts (terms) in Slovene legal language. The project is supported by existing (multilingual) electronic or printed legal sources and other European law ontologies (e.g. JurWordNet, LOIS WordNet, Legal Taxonomy Syllabus, etc.), as well as existing programs for building ontologies (Protégé etc.) and semantic webs and relevant existing text sources (Eur-Lex, Slovenian legislation, SVEZ-IJS Acquis Corpus, Evrokorpus, etc.)
Keywords: ontologies, semantic web(s), lexical taxonomic databases of type WordNet, terminological databases, corpora of domain-specific texts (legal language), computer-assisted database (ontology) building.

1. Uvod

1.1. Leksikalne ontologije

Leksikalne ontologije ali semantične mreže so sistemi kategorij med besednimi in pojmovnimi sistemi (med urejenimi glosarskimi, slovarskimi, tezaverskimi, enciklopedičnimi in terminološkimi zbirkami), ki obe vrsti sistemov povezujejo[1] in se naslanjajo predvsem na taksonomsko organizirano zgradbo pojmov nekega področja.

Zasnova ontologij ali semantičnih mrež na konkretnem strokovnem področju (npr. pravo, medicina, računalništvo, genetika itd.) zajema znanje stroke in semantična razmerja (vodoravna: sinonimija, antonimija in navpična: hiperonimija, hiponimija) med pojmi, predstavljenimi z besedami preučevanega jezika[2].

Predlagani projekt (o)bogatitve terminoloških virov (nasprotni proces kot pri T. Vealeju[3]) z ontologijami zajema podatke iz znanstvenih, strokovnih in poljudnoznanstvenih besedil (najbolje v obliki (označenih) korpusov; na področju prava npr. neoznačeni: Eur-lex, slovenska zakonodaja, besedila iz strokovnih revij, dvo- in večjezični korpusi: SVEZ-IJS Acquis Corpus – označeni, Evrokorpus ipd.), ki omogočajo 'ročno' ali 'računalniško obdelavo' (s pripomočki za rudarjenje podatkov, sestavljanje semantičnih oz. ontoloških mrež in njihovo ponazoritev). Oblikovanje slovenskih (pravnih) ontologij oz. semantične mreže se tako opira na že izoblikovane ontologije oz. sematične in besedne mreže (JurWordNet, Semantic Web, SloWNet, EuroWordNets, ProtégéOntoLT, Polaris, Legal Taxonomy Syllabus, NeON idr.), dosegljive v elektronski obliki, ali tudi na slovarske in enciklopedične vire v tiskani obliki.

Vizualizacija pojmov in razmerij med njimi je odvisna od predstavitvenega in oblikovnega programa ter od množice podprogramov, ki pomagajo pri iskanju pojmov, ki nastopajo v sobesedilu v značilnih relacijskih razmerjih, ki jih predhodno definiramo za vsako strokovno področje posebej (največkrat ob sodelovanju jezikoslovca in strokovnjaka). Pri sestavljanju ontologij je mogoče izhajati tudi iz strokovnih (znanstvenih ali drugih) besedil in izločati strokovne pojme z računalniškimi prijemi[4] ter jih nato umeščati med ontološke gradnike z vnaprej predvidenimi ali pa tudi računalniško posnetimi (računalniško učenje ali modeliranje) razmerji med pojmi

1 Veale, T. (2007), Enriched Lexical Ontologies, sklop predavanj na ESSLLI 2007, Dublin.
2 Kunze C., Lemnitzer L. (2005), Computational Semantics, sklop predavanj na ESSLLI 2005, Edinburg.
3 Veale, T., ibid.
4 Cimiano P., 2006.

in v njiohovem opisu. Sestavljanje ontologij ali semantičnih mrež lahko poteka 'ročno', lahko pa tudi delno ali popolnoma 'računalniško podprto'. Pri tem so nam v pomoč nekateri že obstoječi prosto dostopni programi na internetu. Dober premislek o množici izbranih kategorij utegne sestavljavcu ontologije prihraniti čas ob nujnem ročnem popravljanju, če se izkaže, da je treba dodati ali izbrisati nekatere med njimi. Vendar je ob naslonitvi na že obstoječe ontološko premišljene leksikalne in pojmovne zbirke nekaj teh skrbi že skorajda odveč.

Uporabnost in namen sestavljanja ontologij je poleg bogatitve terminoloških ali kakih drugih slovarskih zbirk nasploh, širjenje znanja o (izbrani: npr. pravni) stroki (kot enciklopedičnega ali nujnega znanja pri usvajanju strokovnih vsebin), pa tudi kot pomoč strokovnemu prevajalcu, specializiranemu za konkretno področje; gradnja zbirk znanja, ki je dana na voljo za nadaljnjo računalniško obdelavo, bodisi na ravni razčlembe drugih besedil in njihove logično-semantične obdelave (glede na vgrajene logično-semantične funkcije vzpostavljanja inferenc), bodisi kot bogatejša leksikalna in terminološka zbirka, ki omogoča vizualizacijo pojmov in njihovih razmerij ter daje tako uporabniku širši vpogled v razumevanje posameznih strokovnih pojmov.

1.2. Ontologije nasploh – kaj so ontologije?

a) Ontologija v filozofiji:

Je znanost o značilnostih stvari, njihovem obstoju oziroma realnosti nasploh, kakor tudi o temeljnih kategorijah bitnosti (stvari) in njihovih medsebojnih razmerjih[5]. Tradicionalno velja ontologija za vejo filozofije in je bila dolgo znana pod imenom metafizika, ukvarja se z vprašanji t. i. entitet, ki obstajajo ali veljajo za obstoječe; kako se te entitete združujejo v večje razrede; kako so znotraj njih razdeljene hierarhično v smislu podobnosti in razlik.

Za ta članek zgodovinska filozofska razglabljanja niso toliko pomembna, je pa že pri definicijah v pravu (bolje: konkretni zakonodaji) zaznati, da se tudi pravo, ko mora konkretno opredeliti pojme in stvari kot npr. premičnine ali nepremičnine, ki so lahko definirane kot lastnina/posest posameznika (gl. Stvarnopravni zakonik, SPZ, členi: 16., 17. in 21.), kjer se pri mereoloških kategorijah, kot so *'je del* (nečesa), *spada k*' (pri opredelitvah pojmov, kot so 'sestavina', 'pritiklina', 'zbirna stvar' …), zateka k modifikacijam ali postavljanju omejitev človekovega razmišljanja in uvaja pogoj *'po splošnem pojmovanju*' (ubesedeno dob. *'kar po splošnem pojmovanju šteje (za)*'), sklicujoč se na človeka nasploh, ki ima ob splošni razgledanosti o stvareh tudi izkušnjo, kakšna je narava teh stvari v njihovem običajnem stanju. Pri preučevanju pravnih terminov se pokaže ravno ta prehod od najbolj splošnih terminov h konkretnejšim. Splošnejši termini tako prevladujejo v splošnejših pravnih aktih (ustavi ali okvirnih aktih za določeno pravno področje), medtem ko so podrobnejši termini, povezani s splošnejšimi, najpogosteje uporabljenimi v izvedbenih aktih (uredbah, odločbah, sklepih itd.). Z razvojem znanosti in tehnologij se spreminja tudi vedenje o stvareh… Tako je v omenjenem SPZ kot predmet posesti navedena tudi *energija* (v opredelitvi temeljnih pojmov – 'stvar', 15.(2) člen: 'Za stvar se štejejo tudi različne oblike energije in valovanja, ki jih človek lahko obvladuje'). Vendar pa se v Slovenskem pravnem leksikonu avtorji (Apovnik in dr., 1999, str. 144) obenem sprašujejo tudi, ali so sestavni del *pravnega reda* tudi splošne mnenjske predstave občanov o pravu, državi in družbi, in menijo, da je to sporno.

b) Ontologija v računalništvu in računalniškem jezikoslovju:

Izraz 'ontologija' izhaja iz filozofije in se je skozi zgodovino uporabljal na različne načine. V svojem bistvu je definicijski pomen 'ontologij' v računalništvu, logiki in ontološki vedi kot novi disciplini '(formalni) model za opisovanje sveta', ki sestoji iz množice vrst/tipov stvari, lastnosti in razmerij med vrstami stvari (tipi). V splošnem se domneva, da obstaja tesna zveza in podobnost med dejanskim svetom in lastnostmi modela v posamezni ontologiji.

Ontologija v računalništvu in logiki predstavlja formalni prikaz (predstavitev) množice pojmov znotraj nekega strokovnega področja in razmerij med temi pojmi. Uporablja se kot opis razlag in definicij lastnosti tega področja, zato je zanimiva njihova uporaba v jezikoslovnih (leksikalna semantika) in terminoloških raziskavah prav s stališča teh opredelitev (definicij) pojavov in razmerij med njimi znotraj tega strokovnega področja.

V formalni ontologiji (teoriji ontologij) je ontologija formalni eksplicitni opis pojavnosti po zgledu opisov semantičnih polj (jezikoslovnih teorij iz 60. let 20. stoletja) in logičnih opisov (razmerij med pojavi ali entitetami) iz teorije logike (prav tako iz druge polovice 20. stoletja). Oboje temelji tako na teoriji množic kot tudi na razširjeni teoriji predikatnega računa.

Tako v filozofiji kot v računalništvu je ontologija po definiciji predstavitev entitet, idej in dogodkov, skupaj z njihovimi lastnostmi in medsebojnimi razmerji – glede na izbrani sistem kategorij. Glede na zadnje se obe disciplini srečata pred vprašanjem relativnosti ontologij (glede na nek konsolidirani arbitrarno sprejeti opisni sistem), najpomembnejši avtorji na obeh področjih pa so: Kripke in Quine (filozofija) ter Sowa in Guarino (računalništvo), podobno tudi sodobni projekti s tega področja (Cyc, na področju umetne inteligence, DOLCE, projekt temeljnih ontologij). Razlika med obema je predvsem v splošni osrediščenosti obeh znanosti[6].

c) Ontologija kot samostojna disciplina:

Po Corazzonu[7] ponuja ontologija merila, ki razlikujejo med seboj različne vrste stvari (konkretne od abstraktnih, obstoječe od neobstoječih, realne od idealnih, neodvisne od odvisnih …) in različne vrste povezav med njimi (razmerja ali relacije, odvisnosti in predikatna razmerja v smislu propozicij itd.). V računalniškem jezikoslovju in umetni inteligenci (sistemi znanj) se razvija zlasti logično formalizirana komponenta na eni strani in vizualizacija ontologij na drugi.

Ontologija, pojmovana kot nova disciplina, je še vedno oprta na svoj izvor v filozofiji in zlasti fenomenologiji. Formalizirana ontologija si prizadeva biti aksiomatska veda in rigorozno uporabljati matematično metodo moderne simbolne logike in aksiomatskih sistemov, hkrati pa pri obravnavi svojega predmeta uporablja kot svojo

5 Wikipedia: geslo Ontologies (Philosophy)

6 en.wikipedia.org/wiki/Ontology_(computer_science)

7 www.formalontology.it

razpoznavno metodo tudi intuitivno preučevanje in pojmovanje temeljnih lastnosti, načinov in vidikov bivanja in entitet nasploh (Guarino, 2005 in kasneje, o temeljni ontologiji DOLCE).

Najbolj prepoznavne analitične kategorije so: stvari, stanje stvari, deli, celote (mereologija) ter razmerja med deli in celotami, zapisana v smislu odvisnostnih zakonov. Analitična opisna in formalizabilna ontologija se ne ukvarja s problemom razmerij med formalno in materialno ontologijo (konkretno ontologijo nekega področja), njen predmet je le strog logični formalni sistem za opisovanje kategorij (stvari, entitet in njihovih razmerij). Nasprotno pa opora v formalizabilnem sistemu daje podlago za računalniško podprto gradnjo ontologij kot zbirk znanj različnih strokovnih področij, pa tudi njihovo združevanje v enovitejši sistem, ki omogoča iskanje entitet, razmerij, podatkov, informacij in njihovo strojno učenje ter izdelavo računalniških modelov strok ter generira specifična strokovna razmišljanja, sklepanja (na logično-formalnem sistemu inferenc) v njih.

2. Predstavitev projekta

Članek prikazuje potek projekta z vzpostavitvijo semantične mreže oz. leksikalnih ontologij na področju pravne terminologije. Prvi korak pri tej vzpostavitvi mreže sloni na gradivu, ki je po obliki podobno 'miselnim vzorcem', ali t. i. preglednicah (brošurah, tabelaričnih in povzemalnih prikazih, ki spominjajo na ontološke prikaze pojmov), ki jih profesorji na pravni fakulteti uporabljajo pri svojem pedagoškem delu. Pravo kot veda (teorija prava) in pravo kot praktična disciplina (t. i. procesno ali izvršilno ali izvedbeno pravo) obravnavata širok spekter pravnih (prepletajočih se in hkrati ločenih) disciplin in definirata temeljne pravne institute, pravni kategorialni aparat, skupaj s pravnimi mnenji itd. V smislu pravne taksonomije in pravne terminologije je treba upoštevati vsako pravno disciplino posebej, hkrati pa jih tudi povezovati in postopoma graditi 'taksonomsko pravno nadstavbo', ki si jo v okviru jezikovnega preučevanja lahko predstavljamo kot semantično mrežo hierarhično in vzporedno razvrščenih pojmov, ki jih definiramo v okviru semantičnega pristopa k pravni terminologiji in vedenju o pravu (taksonomije), v ontološkem okviru jih lahko predstavimo v obliki t. i. WordNets (pojmovnih ali taksonomskih mrež nekega strokovnega področja), le-te pa potem uporabimo v nekem ontološkem računalniškem programu, ki upošteva standarde sestavljanja ontologij (npr. *NeON*, *101 Ontology* ali *Protégé*).

V konkretnem strokovnem področju, tj. posamezni pravni disciplini, najprej vzpostavimo ob poznavanju stroke tudi semantična razmerja med pravnimi pojmi, ki jih sproti dopolnjujemo in gradimo širšo semantično mrežo pravnih pojmov (v *JurWordNet* oz. *LOIS WordNet*). Ponuja se nam primerjava z obstoječimi sistemi semantičnih mrež (v drugih jezikih), ki so kot ontologije definirane za konkretni pravni jezik. Vsak pravni jezik gradi lastne pravne ontologije ali semantične mreže/taksonomije in ima podlago v veljavnem pravnem sistemu oz. pozitivnem pravu na ozemlju neke države (jezikovne in državne meje si seveda vedno ne ustrezajo, npr. nemški pravni jezik v Švici ali Nemčiji ter francoski pravni jezik v Franciji, Kanadi ali Luksemburgu itd.). Večjezični sistemi povezanih ontologij praviloma niso neposredne preslikave iz enega pravnega jezika (terminologije, ontologij) v drugi[8], ampak so razmerja zapletenejša, čeprav je sama primerjava možna tam, kjer so pravne ureditve med seboj podobne (npr. zakonodaja EU in zakonodaje posameznih držav članic, ali pa zakonodaja srednjevevropskih držav z obsežnim vplivom nemške zakonodajne prakse, zakonodaja držav z romanskim uradnim jezikom tudi kot primer 'kontinentalne zakonodajne prakse' z velikim poudarkom na izhodiščnem rimskem pravu, z ustaljenimi leksikalnimi (latinskimi) pravnimi formulacijami ter britanska zakonodaja, ki izhaja bolj iz same pravne prakse t. i. case law. Slovenska zakonodajna praksa se od leta 1990 – ob nastanku prvih zakonov – ustave kot prvega državnega akta novonastale oz. takrat nastajajoče, države, v večji meri zgleduje po nemški, predvsem pri sestavljanju okvirnih zakonov, in nekaterih drugih sorodnih zakonodajnih praksah (pri pisanju in sprejemanju zakonov), saj so bili slovenski odvetniki in pravniki nasploh 'vajeni' branja in tolmačenja zakonov, napisanih v nemščini v času Habsburške monarhije, nekateri pravni izrazi so bili že takrat oblikovani bodisi kot kalki ali kot (po smislu) prevodne ustreznice nemških pravnih izrazov)[9] veliko novosti pa so v slovensko zakonodajo prinesli predvsem danes vodilni slovenski pravni strokovnjaki.

Predlagani projekt obogatitve terminoloških virov z ontologijami na področju prava zajema tako iz različnih virov, ki so na voljo na spletu (npr. Eur-Lex, slovenska zakonodaja – na različnih spletnih straneh, med drugim v Registru pravnih predpisov Slovenije (ki ga ureja Služba vlade za zakonodajo): www.rps.si, na www.zakonodaja.si ali na straneh Državnega zbora www.dz-rs.si, kjer so na vpogled tudi zakoni v pripravi; besedila iz strokovnih revij, komentarji k publiciranim pravnim aktom različnih založb, zlasti založb Uradni list in IUS-INFO ipd.) ter predvsem iz že vzpostavljenih semantičnih oz. ontoloških mrež zunaj slovenskih spletnih strani (JurWordNet, LOIS Legal WordNet, Legal Taxonomy Syllabus, EuroWordNets, SloWNet, ProtégéOntoLT, Polaris, NeON, Cyc[10] ter drugi ontološki, urejeni strokovni in enciklopedični viri). Po predhodno opravljeni strokovni in jezikovni raziskavi definiranja pravnih strokovnih pojmov ter testiranju obstoječega urejevalnika ontologij (101 Ontology, NeON Toolkit, ProtégéOntoLT), ustreznega prikazovalnika razmerij med pojmi in njihovo razčlembo (ta faza je v teku) in zgradbene razčlembe obstoječega modela ontologij za druge evropske jezike (kot stranski produkt pretvorb formata XML (kot jezikovnotehnološkega in dokumentalističnega standarda tako za strokovna besedila kot tudi za jezikovne zbirke, včasih tudi ontološke zbirke, npr. LOIS Legal WordNet) v format, značilen za urejanje in shranjevanje ontoloških zbirk OWL (web ontology language). Program za urejanje, gradnjo ontologij je običajno večmodulski, vsebuje več podprogramov, ki omogočajo vnos, urejanje, preverjanje konsistentnosti, določanja razmerij ter vizualizacije in iskanja izrazov, ki se nahajajo v okviru definiranih semantičnih razmerij (med obstoječimi programi se je kot testni program za urejanje slov. (pravnih) ontologij izkazal kot dosleden, zmogljiv,

8 Gruntar Jermol, A. (2005), str. 175–177.

9 Jemec-Tomažin, M. (2005).

10 http://www.cyc.com/, načela zgradb iskalnikov po spletnih straneh.

nadgradljiv, standarden in dovolj prijazen za uporabnika – program NeON Toolkit).

Pri omenjenem projektu je nujno sodelovanje pravnih, jezikoslovnih in računalniških strokovnjakov, sestavljanje oz. gradnja ontologij sprva poteka sprva 'ročno' s sprejemanjem standardov za definiranje temeljnih pravnih pojmov in razmerij med njimi, kasneje pa tudi 'računalniško podprto'. Pri tem so na voljo številna že obstoječa orodja, ki jih je treba morda le še nekoliko prilagoditi. Od podrobnosti semantičnega opisa (definicij) in rabe pravnih pojmov, ki so v nastajajočih ontologijah prisotni, pa je odvisno, ali bodo ontologije bolj površinske, bližje Wordnets (besednim mrežam), ali pa bolj goste (z gostejšo mrežo razmerij in opisa posameznega pojma), bližje dejanskim semantičnim mrežam ustreznega področja. Med obstoječimi pravnimi ontološkimi zbirkami sta predvsem dve, ki bosta služili kot zgradbena podlaga za gradnjo slovenskih ontologij, v prvi fazi projekta pa kot primer za bogatitev terminoloških zbirk, zlasti spletnih (npr. Evroterm), kjer je treba ponatančiti zlasti opise pravnih terminov in jih postaviti v povezave (semantična razmerja: nadpomenskost, podpomenskost, sopomenskost, protipomenskost ipd.), kar prevajalcu ali drugemu uporabniku bistveno olajša delo.

2.1. Prva faza

Predstavitev uporabe že izdelanih ontologij (LOIS Wordnets, Taxonomy Syllabus) kot podlago k zamisli za oblikovanje slovenskih pravnih ontologij – ter hkratno dopolnjevanje terminološke zbirke, zlasti z mikroelementi[11] ontoloških prikazov, kot so definicije pojmov, raba izrazov (ustrezne kolokacije v pravnem jeziku), razmerja med izrazi/pojmi in njihova pojmovna razčlenitev , ponatančitev.

Uvodne težave – razpoložljivi viri: pičlost dosegljivih pravnih virov, obdelanih primerno za jezikoslovno, pomenoslovno (leksikalno, pojmovno), ontološko obravnavo. S pravnega vidika so na voljo vsi zbrani zakoni, podzakonski akti in ustava Republike Slovenije, kakor so bili objavljeni v tiskanem Uradnem listu Republike Slovenije (www.uradni-list.si), tudi v spletni obliki; druga oblika istih virov je proti plačilu naročnine dosegljiva za uporabnike pod imenom zbirke IUS-INFO (vsebuje dodatne komentarje k zakonom, zbrane iz strokovnih revij ali posebnih izdaj…). V zadnjem času GV založba ob izdajanju dveh pravnih strokovnih revij (nista dosegljivi v elektronski obliki, razen na portalu IUS-INFO), Pravnika in Pravne prakse, izdaja tudi predstavitveni, definicijski material pravne teorije za interesente (študente prava in druge) pod imenom Pravne preglednice (katerih avtorji so trenutni nosilci predavanj posameznih pravnih disciplin na ljubljanski pravni fakulteti). V sklopu pobud študentov in profesorjev prava nastajajo tudi geselski članki v prostodostopni spletni enciklopediji Wikipediji v slovenskem jeziku (sl.wikipedia.org/wiki/pravo), nekaj t. i. študijskega gradiva pa postavljajo na splet tudi študentje prava in njihovi profesorji (prof. dr. J. Čebulj, nekdanji ustavni sodnik, definicije pojmov iz javnega prava, iz predavanj za štud. leto 2008-09).

Definicije pravnih pojmov se da iskati tudi na drugačen način, predvsem na Googlovem spletnem naslovu (www.google.com) s posebnimi iskalnimi prijemi (npr. Term:definition, kjer Term pomeni kateri koli iskani izraz), vendar dobimo tako večinoma le tujejezične, največkrat angleške izraze, ki nam zaradi drugačnega pristopa k zakonodaji in pisanju zakonov ter pravne prakse in pravnega sistema pri slovenski pravni terminologiji ne pomagajo prav veliko, opozorijo nas pa lahko, na katere tuje izraze in pojme smo še lahko pozorni, te seveda preverimo tudi s približnimi ustreznicami v dvo- in večjezičnih pravnih ali splošnih slovarjih, večinoma dostopnih na spletu ali v elektronski obliki (plačljivo), namestljivi kot programska oprema.

Že v italijanskem govornem prostoru je očitna razlika pri ukvarjanju z računalniško dostopnimi pravnimi viri in z računalniškimi pristopi za njihovo obdelavo. Poleg projekta inštituta ITTIG iz Firenc (LOIS Legal WordNet) obstaja še cela plejada poletnih šol, konferenc, inštitutov v različnih krajih Italije (od Rima, Milana itd., do Trenta, npr. Jurix idr.), ki se ukvarja bodisi s standardizacijo zapisov pravnih besedil v elektronski obliki, bodisi s samim sestavljanjem ontologij in standardov zanje (za semantične jezikoslovne raziskave ali raziskave na področju razvoja računalniškega orodja za spremljanje, ekstrakcijo podatkov in ugotavljanje razmerij med njimi, npr. pravnimi pojmi v ontoloških zbirkah, ter programi za sklepanje o pomenu in/ali t. i. ugotavljanju inferenc iz strokovnih besedil). V slednjem se računalniško področje umetne inteligence z uporabo različnih vrst logik povezuje s področjem prava, ki postaja v svetu trend, v slovenskem prostoru pa ostaja predvsem zanimivost, ki bi bila sicer vredna preučevanja in napora.

2.2. Druga faza

Gradnja ontološke pravne zbirke s prikazovalnokom razmerij med pojmi je kljub obstoječim računalniškim orodjem dolgotrajna naloga: zahteva sodelovanje pravnikov, jezikoslovcev in glede delovanja programov občasno tudi računalnikarjev.

Ta faza se nanaša predvsem na raziskavo formalnih zgradbenih elementov ontologij, po eni strani na pojmovni ravni (pravna zgradba pojmov, definiranih v splošnejših pravnih aktih, t. i. zakonikih), po drugi strani pa na zgradbo v tehničnem smislu, ki se kaže predvsem v rabi konkretnega normiranega, standardiziranega zapisa pravnih besedil, ki omogoča tudi njihovo izmenjavo, samodejno računalniško branje ipd. in t. i. jezika ontologij. V zadnjem času je bilo več poskusov standardizacije obeh, tako se kot jeziki omenjajo označevalni jeziki (angl. *mark-up languages*), kot 'osnovna zgradba' za zapis ontologij pa je tako v terminologiji kot v ontologijah uveljavljen standardiziran zapis v formatu XML (ki je imel več predhodnih in vzporednih oblik, npr. SGML, HTML, PAT, RTF z oblikovnimi formati SIM, OPC, CDD, EnAct…, nekateri formati se uporabljajo v določenih državah, večinoma zunaj EU, npr. CHLexML v Švici[12] ali PAT v Avstraliji – na Tasmaniji). Večina računalniških konzorcijev, ki se ukvarjajo s sestavljanjem in razvojem ontologij, podpira

11 izraz iz: Gruntar Jermol, A. (2005), Lässt sich Rechtssprache visualisieren?

12 http://www.rechtsinformation.admin.ch/copiur/index.html, Copiur, Zvezni urad za elektronske pravne dokumente sodišča, Bern.

že omenjena standarda (npr. konzorcij W3C[13] ali LOIS – ITTIG[14]), XML/RDF in OWL (Ontology Web Language) z njunimi različicami, ustvarjenimi za različne namene (npr. za spletne uporabnike, ki želijo prejeti le besedila zakonov v primerni obliki za tisk itd. – OWL-S in njegov predhodnik DAML-S). Spletni ontološki jezik (OWL, ali tudi Web OWL) ima prav tako nekaj različic ali 'podjezikov', ki so namenjeni različnim nalogam: OWL Lite, OWL DL in OWL Full.

Za ilustracijo si oglejmo jeziko(slo)vne parametre, ki jih postavljajo v pravni informatiki predvsem strokovnjaki za semantični splet in pravne ontologije. Računalnikarji jih imenujejo metapodatki, za jezikoslovce in pravnike so to bistveni vsebinski podatki. V evropskem projektu Estrella in drugje so jih definirali v t. i. shemi MetaLex OWL. V grobem okvirju opisujejo predvsem dogodke (events, hipne, tudi situacije) in dogajanja/ukrepe (actions), tretja kategorija so transakcije (transactions, odškodnine, nadomestilča, povračila, plačila itd.). Vsaka kategorija v pravu opredeli tudi udeležence (participants) teh dogodkov nasploh. Pri tem si MetaLex pomaga s klasifikacijo udeleženskih vlog (thematic roles) kot jezikoslovnim pojmom, te opredeli kot semantično razmerje med glagolom in argumentom (samostalniška zveza) povedi. V pravni informatiki pri tem uporabljajo predvsem t. i. standarde CEN (ki so bili definirani že 1991). Udeleženci v dogodku (ki je zanimiv s pravnega stališča) so lahko[15]:

– neposredni in/ali določljivi (določujoči) udeleženec (pasivni oz. aktivni udeleženec);

– vir in/ali izdelek (prvi je prisoten le na začetku, slednji je nujno prisoten na koncu dogodka/dogajanja);

– vršilec je določljivi udeleženec ali skupina udeležencev, vir je vzvod nekega dogodka/dogajanja (samo slednji imajo aktivne vršilce);

– sredstvo je neposredno določljivo na kraju dogodka/dogajanja in se med tem ne spreminja npr. zakon, zakonski predpis – velja od nekega datuma do nekega drugega kasnejšega datuma);

– prizadeti je neposredni udeleženec in utrpi neko dogajanje;

– izdelek dogajanja je produkt dogajanja in utrpi nekatere zgradbene spremembe in predstavlja rezultat tega dogajanja

– prejemnik je določljivi udeleženec dogajanja in hkrati izid dogajanja (pri transakcijah);

– izid (rezultat) dogajanja je določljivi predmet dogajanja (stvar, ki je bila izdelana, namerno ali kot splet okoliščin);

– datum je neposredni parameter (lastnost), je predmet zakonodaje – kdaj se je kateri dogodek/dogajanje zgodilo/dogajalo.

Generični pojmi v pravu (kot sistemu pravnih predpisov, zakonodaji) so:

– nastanek pravnega predpisa;

– pobuda za nastanek predpisa (pobudnik: pristojni (zakonodajni ali izvršilni) organ);

– pritožba kot instrument (pri pristojnem organu) za uporabo ali spodbijanje veljavnega prepisa.

```
<LITERAL LEMMA="contratto" SENSE="1">
  <EXAMPLES />
     </LITERAL>
     </VARIANTS>
- <INTERNAL_LINKS>
- <RELATION ID="110" TYPE="fuzzynym">
  <TARGET_WM ID="544" PART_OF_SPEECH="N" />
     </RELATION>
- <RELATION ID="111" TYPE="has_hyperonym">
  <TARGET_WM ID="4" PART_OF_SPEECH="N" />
     </RELATION>
- <RELATION ID="113" TYPE="co_result_instrument">
  <TARGET_WM ID="99" PART_OF_SPEECH="N" />
     </RELATION>
     </INTERNAL_LINKS>
- <EQ_LINKS>
- <RELATION ID="160000" TYPE="eq_near_synonym">
  <TARGET_WM ID="32" PART_OF_SPEECH="N" />
     </RELATION>
     </EQ_LINKS>
     </WORD_MEANING>
```

(primer zapisa ontoloških gesel z uporabo standardov)

MetaLex in CEN OWL[16] vsebujeta poleg navedenih še druge opisne parametre za opisovanje pravnih predpisov (zapisane v formatu RDF ali XML), zlasti besedilne modifikatorje, medsebojne (zakonske, normativne) reference, citate ipd. z določenimi semantičnimi vrednostmi, npr. področje delovanja instrumenta (ukrepa) – npr. finančni, šolski, kmetijski sektor; območje delovanja (prostor – neka država ali njen del); časovno obdobje; spremembe predpisov, ki vplivajo na veljavnost predpisa ali ga v določenih komponentah omejujejo ali tudi razširijo njegove pristojnosti; ter drugi atributi in kazalci (staro besedilo, novo besedilo predpisa, prečiščene različice ipd.).

V tej fazi je nujna opredelitev pojmovnih kategorij, ki jih želimo vnesti v ontologije, zamejitev pravnega področja in vzpostavljanje razmerij ter gradnja pojmovnih mrež (dreves), v nekaterih programih je samodejna. Iz tega izhaja tudi potreba po sodelovanju različnih strokovnjakov, predvsem tistih, ki sodelujejo pri vsebinskem delu projekta, uporabljajo pa delno že nekatere navedene formalizirane standarde.

2.3. Tretja faza – samodejno sestavljanje ontoloških zbirk

Najprej je treba preučiti obstoječe programe, ki bi bili primerni za sestavljanja ontoloških zbirk.

Trenutno je v fazi testiranja program NeON Toolkit, ki omogoča vnašanje in primerjavo med vnesenimi pojmi na podlagi uvrščanja v razrede (pojmi – word-senses, literals – leksikalni izrazi, variants – njihovi sinonimi), številčenje ali indeksiranje pojmov in leksikalnih enot, jezikovno opredelitev slovničnega razreda (PoS – besedna zveza), opredelitev razmerij do drugih pojmov (generični pojem, vsebuje podpomenko, nadpomenko, sinonimni izraz, antonim ipd.). Določene logične dele pojmov in razmerja je mogoče vnesti z urejevalnikom ontologij, za njihovo prikazovanje (vizualizacijo) v obliki 'semantičnih mrež' poskrbijo, npr.:

- risalniki ontoloških vzorcev (Ontology Design Patterns), poleg the pa smo ali bomo še potrebovali:

13 spletna stran konzorcija W3C: www.w3.org;
14 konzorcij LOIS na Inštitutu ITTIG v Firencah: izdelal ontološko zbirko LOIS Legal WordNet: spletna stran: www.ittig.cnr.it;
15 Palmirani, M. in dr. (2006-07).

16 glej projekt Estrella (projekt EU), ki definira standarde za pravne in zakonodajne vire (Legal Informative Resources): IST-2004-027655 (European project for Standardised Transparent Representation in order to Extend Legal Accessibility)., 22. 1. 2007, CNIPA-MRIPA.

- programe za pretvorbo formatov (med XML, XSD, RDF in OWL – format converters);
- programe za pomenoslovno (semantično) označevanje (semantic annotation programmes), običajno se pri zajemanju strokovnega izrazja iz strokovnih besedil ustrezne korpuse označi s pomenoslovnimi oznakami, ki jih želimo imeti pozneje v ontološki zbirki za morebitne pomenoslovne razčlembe;
- programe za ontološko označevanje – označevanje razmerij, pomembnih v ontologijah ali semantičnih mrežah (ontological annotation);
- programe za indeksacijo jezikov in označevanje relacij med njimi (indexation programmes), ti so lahko že vsebovani v obstoječem programu;
- vpisovalnike/urejevalnike semantične mreže ali ontologije (ontology/semantic web/net editor);
- ogledovalnik ontoloških dreves/mrež oz. prikazov (ontological database viewer/periscope);
- iskalnik besedilnih ustreznic, pojmov (information retrieval system, term extraction).

Samodejno sestavljanje sledi poskusnemu ročnemu vstavljanju in preverjanju lokacij ontoloških enot (pojmov) v ontologiji, predvsem je nujen program, ki oblikuje drevesne mreže in omogoča širjenje le-teh po izbiri sestavljavca. Zato je nujno, da je program čim bolj elastičen in prilagodljiv potrebam vsebinskega dela sestavljanja osnovnih pravnih ontologij. Iz standardnih formatov XML je mogoče zapisane vsebine kot elemente ontologij na primeren način pretvoriti v program vizualizacije le-teh. Pri tem je potrebna zgradbena in miselna pretvorba med obema zapisoma (neke vrste oklepajnim in drevesnim, po zgradbi).

Nujno potrebno je sodelovanje jezikoslovcev, pravnikov ali poznavalcev pravne terminologije in računalnikarjev. V tem smislu se tudi določi vrstni red, količina, vrsta končnega izdelka in njegova uporaba.

3. Zaključek

Projekt sestaviti osnovno zbirko pravnih ontologij v smislu ontološke vede, glede na razmerja, v kakršnih ti pojmi dejansko obstajajo v slovenskem pravnem jeziku je obsežen in dolgotrajen projekt. Opora na obstoječe (večjezične) elektronske in delno na tiskane vire pravnih in drugih ontologij evropskega prava (npr. JurWordNet, Legal Taxonomy Syllabus ipd.), kakor tudi na programe za gradnjo le-teh ter ustrezne že obstoječe besedilne vire (Eur-lex, slovenska zakonodaja, SVEZ-IJS Acquis Corpus, Evrokorpus ipd.) je v začetku koristna, med izvajanjem samim pa projekt razvija tudi lastno metodologijo in išče lastne rešitve. Potrebno je vsekakor sodelovanje več ustreznih znanstvenih strok. Kot uvodni stranski produkt gradnje ontologij je lahko pri zamisli gradnje ontologije koristno predvsem dopolnjevanje terminološke zbirke z mikroelementi. V nadaljevanju je po vzpostvitvi ožje ali širše zasnovane ontološke (pravne) zbirke z ustreznim prikazom mreže razmerij med pojmi potreben razmislek bodisi o združevanju s terminološko zbirko (obratna pretvorba formatov: OWL v XML), prelitjem ontologije v tako zbirko (kar tako zbirko tudi notranje obogati), bodisi kot možnost zunanjega navezovanja med enakimi izrazi (pojmi) v terminološki in ontološki zbirki v obliki mrežnega prikakovalnika (vizualizacija).

4. Literatura

Borgo, S., Guarino, N., Vieu L., 2005. Formal Ontologies for Semanticists, *skop predavanj na ESSLLI 2005* (ESSLLI-05 reader), Edinburg.

Buitelaar, P., Cimiano P., 2007. Ontologies and Lexical Semantics in Natural Language Understanding, *sklop predavanj na ESSLLI 2007* (ESSLLI-07 reader), Dublin (Trinity College Dublin).

Cimiano P., 2006. Ontology Learning and Population from Text: Algorithms. Evaluation and Applications, Springer, Springer Science and Business, Berlin.

Erjavec, T., 2005. Annotation of language resources: XML, TEI, OWL, *sklop predavanj na ESSLLI 2005* (ESSLLI-05 reader), Edinburg.

Fellbaum, C., 1998. WordNet, An Electronic Lexical Database, The MIT Press, Cambridge, Massachusetts.

Fišer, D., 2005. Pristop k izdelavi leksikalnih podatkovnih zbirk, *Jezik in slovstvo*, let. 50 (2005), št. 6. <http://www.ceid.upatras.gr/Balkanet> [1.09.2009].

Gruntar Jermol, Ada, 2005. Lässt sich Rechtssprache visualisieren?, *Zagreber germanistische Beiträge*, letnik 14, str. 175–189, grafični prikazi.

Janevski S., 2009. Varstvo potrošnikov, Prodaja in nakup potrošniškega blaga in storitev s sodno prakso, primeri iz prakse in obrazci, Založba Uradni list, Ljubljana. (kazalo na: http://www.uradni-list.si/uploads/kazalo.pdf)

Jemec, M., 2007. Slovensko pravno izrazje od Habsburške monarhije do Evropske unije ali 'ius est ars'. V: *Irena Orel (ur.), Obdobja 24 – Metode in zvrsti. Razvoj slovenskega strokovnega jezika*. Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, Ljubljana.

Kohler, A. in Schulte am Walde, S., 2007. Introduction to corpus-based computational semantics, *sklop predavanj na ESSLLI 2007*, Dublin.

Kunze C., Lemnitzer L. (2005). Computational Lexicology, *sklop predavanja na ESSLLI 2005* (ESSLLI-05 reader), Edinburg.

Müller, J., 2002. Slovenski večjezični pravni slovarji. *Filologija*, knjiga 38–39, 83–90.

Palmirani, M. in dr. (2006-07), Estrella 6 Project (r. 3.1).

Selic, B., Benjamins, R., Casanovas, P., Gangemi, A., 2005. Law and the Semantic Web: Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications, Springer-Verlag Berlin in Heidelberg.

Veale, T., 2007, Enriched Lexical Ontologies, *sklop predavanj na ESSLLI 2007* (ESSLLI-07 reader), Dublin.

nekateri elektronski ontološki ali semantični računalniški viri:

www.eulawtaxonomy.org
www.ittig.cnr.it/Ricerca/
www.illc.uva.nl/EuroWordNet/
wordnet.princeton.edu
lojze.lugos.si/~darja/slownet.html

elektronski pravni ali pravnoteoretični viri:

ec.europa.eu/prelex/apcnet.cfm?CL=sl
www.uradni-list.si/1/objava.jsp?urlid=200287&stevilka=4360 (Stvarnopravni zakonik)
it.wikipedia.org/wiki/Pagina_principale
it.wikiversity.org/wiki/Pagina_principale

Media Analysis through Contrast Pattern Mining

Senja Pollak

Faculty of Arts, University of Ljubljana, Slovenia
senja.pollak@ff.uni-lj.si;

Abstract

Data mining aims at constructing models or finding interesting patterns. This paper presents a selection of data and text mining approaches for finding contrasting patterns in newspaper text corpora. This study analyzes a corpus of articles covering the 2007 Kenyan elections and post-election crisis, attempting to capture the differences between local (Kenyan) and Western (British and US) newspaper articles. We have first applied a contrast set mining method aimed at extracting contrasting patterns in articles of different news sources. We have also applied a tool for semi-automated topic ontology construction Ontogen, enforcing the separation into clusters of local and Western documents, and analyzed the keywords proposed by the Support Vector Machine classifier that best distinguish between the two classes. Next we used the same tool for finding the topics by k-means clustering of the entire document corpus and interpreted the contrasting words between local and Western views on each of these topics. The most interesting contrasting patterns reveal that, in contrast to local media, the Western press interpreted the elections from the ethnic perspective.

1. Introduction

Data mining aims at constructing models from large collections of data, or at finding interesting patterns in the data (Witten and Frank, 2005). Data mining can be viewed as a central step in the process of *knowledge discovery in databases* (KDD), where knowledge discovery refers to the overall process of discovering useful knowledge from data (Fayyad et al., 1996). The knowledge discovery process is a cooperative effort of humans and computers: humans select the data to be explored, define analysis problems, set goals and interpret the results, while computers search through the data looking for patterns and models that meet the human-defined goals. *Text mining* (Feldman and Sanger, 2007) is a variant of data mining in which models and patterns are extracted from unstructured natural language text.

Media analysis has been a topic of several studies, including the application of statistical and machine learning approaches to daily monitoring of news from different media in the Europe Media Monitor research and development project of the European Joint Research Center (http://emm.newsbrief.eu/overview.html) that gathers reports from news portals world-wide in 43 languages, classifies the articles, analyses the news texts by extracting information from them, aggregates the information, issues alerts and produces intuitive visual presentations of the information found. Text classification methods prove to be a useful vehicle e.g., for different newspaper article classification tasks, such as article genre, topic or author classification. The closest to our work is a media analysis study of Fortuna et al. (2008).

This paper explores a specific media analysis task, aimed at finding patterns that distinguish different document classes, such as journal articles published by different media when reporting about the 2007 Kenyan election crisis. We aim at finding differences between two selected classes in the corpus: local and Western journals. Two main methods for finding contrasting patterns are adopted: contrast set mining (Bay and Pazzani, 2001) and semi-automatic topic ontology construction (Fortuna et al., 2007).

The starting hypothesis is that news coverage of Kenyan events is not the same in the Kenyan local and in the US and British Western press. We expect differences due to addressing different audiences, different ways of reporting, as well as some ideology-related differences, since reporting is never neutral: newsmakers always select the news, present and interpret the events, frequently from a particular ideological position (Fairclough, 1995). They write and report addressing their expected audiences and offer a frame of interpretation of events. Media power is generally symbolic and persuasive, in the sense that the media primarily have the potential to control to some extent the minds of readers or viewers, but not directly their actions (Van Dijk, 1995).

In the theory of linguistic pragmatics, using language is seen as a process of meaning generation characterized by the constant making of choices, where all (conscious and unconscious) lexical, syntactic or discursive choices, are considered to be significant and could have ideological implications (Verschueren 2008).

In this research, we show how text and data mining techniques can be used for detecting contrasting choices. In previous work (Pollak, 2009) we have proven that differences in local and Western reporting can be approached as data mining classification task, generating classification models with 90% accuracy and providing interesting insight into the data. In current paper we use contrast set mining and text mining techniques to continue in this line of research.

We are aware that data and text mining models do not take into consideration the contexts in which the articles were originally produced and interpreted and therefore accentuate that these techniques are an interesting view of a corpus, which should be combined with further qualitative interpretations.

The paper is structured as follows. Section 2 presents the data set, followed by applying a simple tag clouds approach, aimed at analyzing differences between the two groups of articles. Section 3 addresses a contrast set mining task where the goal is to find differences between groups of instances (Western and local articles). Section 4 addresses the same problem by a topic ontology construction approach, supported by the OntoGen tool.

2. Corpus description and exploration

This media analysis study concerns Kenyan presidential and parliamentary elections, held on December 27, 2007, and the crisis following the elections. Two main election candidates were incumbent president Mwai Kibaki and the opposition presidential candidate Raila Odinga. Odinga's party Orange Democratic Movement (ODM) won the parliamentary election against Kibaki's Party of National Unity (PNU). Kibaki is a member of the traditionally dominant Kikuyu ethnic group and Odinga is a member of the Luo ethnic group. Kibaki was declared the winner of presidential elections and swore in, despite Odinga's claims of victory. The election was followed by violence, riots and conflicts. The crisis ended in a power-sharing agreement, signed end of February 2008: Kibaki became the president, and Odinga the prime minister who swore in April 2008.

The original corpus was collected as part of the *Intertextuality and Flows of Information* project, led by the Center of Pragmatics, University of Antwerp. For our experiments we selected 464 articles (about 320,000 words) from six different daily newspapers in English, covering the time period from December 22, 2007 to February 29, 2008. The British and US press *(The Independent, The Times, The New York Times* and *The Washington Post*) that we label "Western" (WE) published 232 articles on the topic. In order to have a symmetrical corpus, we randomly selected 232 articles from the Kenyan, "local" newspapers (LO) *Daily Nation* and *The Standard*.

Since our aim is to better understand the way of reporting on the same events by different newspapers, we had to remove all information that could be distinctive for the two document classes, but not significant for our work. To illustrate, newspapers have normally only few journalists covering Kenya events; if their names were not removed, the author's name could easily be selected as a distinguishing feature.

Therefore, we removed meta-information such as newspaper source, authors of articles, dates of publication, names of photographers, mails of authors, types of articles, etc., and used only the remaining relevant data for document analysis (titles, text and photo descriptions).

In order to get an initial data understanding we used the most straightforward method for getting a "feel" of the contents of the data: the tag cloud representation of documents (http://tagcrowd.com/). As our goal was to find the differences between the local and Western articles, we constructed two separate tag cloud representations shown in Figure 2, one for each of the two classes (local and Western). Since we are interested in contrasting patterns, it is particularly interesting to pay attention to the words that are present in the cloud of only one of the classes. These distinguishing words are listed below:

Local:	added, commission, community, constitution, eck, house, Kenyans, mediation, meeting, members, minister, mp, national, ODM, office, peace, PNU, security, solution, state, support, team.
Western:	africa, areas, days, ethnic, kikuyu, live, luo, mwai, opposition, supporters, town, tribal, tribe, united, valley, week, Western, years.

Figure 1: Contrasting words

From a comparison between two tag clouds (Figures 2a and 2b) we can see that many key words are the same (which correspond to the expectation concerning the fact that they speak of the same topic). However, gap analysis (Figure 1) can lead to understanding of differences which show different perspectives on the reported events.

The most interesting observations are that in the local press we do not find any references to the ethnicity, while in the Western press the words *ethnic*, *tribe, Kikuyu* and *Luo* are part of the tag cloud (i.e. between the most frequent words). Secondly, words appearing only in local press are more of political frame, e.g. *ODM* and *PNU* (the main two political parties), *eck* (standing for Electoral Commission of Kenya) *mp, commission, community, constitution, mediation, meeting, peace, security, solution*.

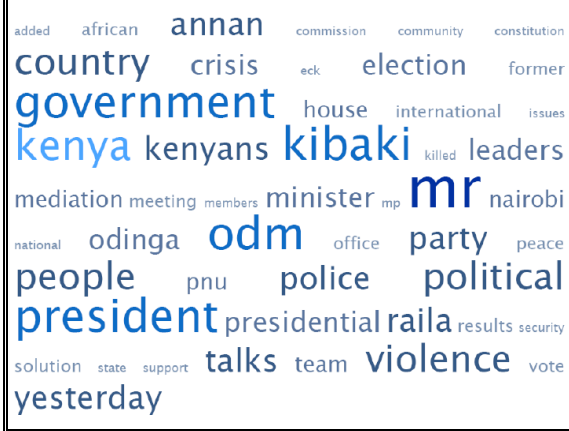


Figure 2a. Tag cloud representation of local articles.

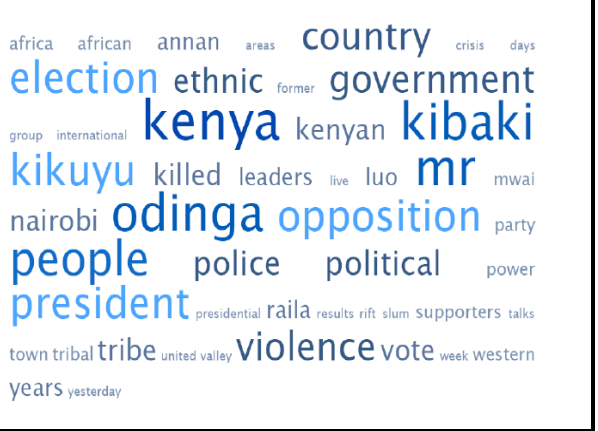


Figure 2b.Tag cloud representation of Western articles.

Figure 2: Tag cloud interpretation

3. Contrast Pattern Mining

Finding contrasting patterns is one of fundamental tasks in data analysis as it aims at finding differences between instances of different groups. This descriptive data mining task, where groups of labeled examples are given and the goal is to find differences between the groups, is known as *contrast set mining* (CSM) (Bay and Pazzani, 2001). Specifically, CSM is defined as finding "conjunctions of attributes and values that differ meaningfully in their distributions across groups." In this paper, a *subgroup discovery* (SD) method was applied to discover contrasting descriptions of groups of documents of the given class (LO or WE), as it was shown in (Kralj Novak et al., 2009) that each CSM task can be adequately translated to a SD task. A SD task aims at constructing rules describing a group of instances of the target class, which is as large as possible and has the most unusual statistical distribution of the target class (Wrobel, 1997).

For the task of contrast set mining, we used SD algorithm (Gamberger and Lavrač, 2002) implemented as Orange[1] web service (Podpečan et al., 2010). For data representation, we built feature vectors of 500 attributes (word unigrams) ranked by chi square test. The feature values are binary (1 - word is present in the document, and 0 – word is not present in the document). A selection of best contrasting patterns (subgroups) for each class is shown in Figure 3 (all rules have high coverage of positive examples of the target class only).

1) kikuyu = 1 & odm = 0 => WE (TP =114, FP = 0)
2) kikuyu = 1 & post-election = 0 => WE (TP = 96, FP =0)
3) kikuyu = 1 & mr = 0 => class = WE (TP = 89, FP = 0)
4) dr =1 & seems = 0 & odm = 1 => LO (TP = 51, FP = 0)

Figure 3: Selection of contrasting rules. TP (true positives) and FP (false positives) show the number of positive and negative examples respectively.

The analysis shows that all the contrasting patterns discovered for class WE involve the word *Kikuyu*, and are characterized by the absence of words: *ODM, post-election* and *Mr* respectively. Rule 1 covering a lot of WE examples says that Western press uses word *Kikuyu* but not the abbreviation *ODM*. The simplest interpretation can be done in terms of different backgrounds of addressed readers: Orange Democratic Movement party's abbreviation for Western readers bears no meaning and was avoided by Western journalists, whereas explaining who belongs to *Kikuyu* is not needed for Kenyan readers. However, in the corpus we discover that the name of the party is not a very unbalanced feature only in its abbreviated form, but also when named *Orange Democratic Movement*. This leads to a hypothesis that Western media prefer to present the Kenyan elections in terms of conflicts between different ethnic groups and not in terms of political parties (cf. tag cloud representations in Figure 2).

Rule 3 indicates that when Western press focuses on the main ethnic group (*Kikuyu*), it does not mention many protagonists, which could be introduced by title *Mr*. The interpretation of events as a *post-election* crisis or violence is specific to local press (cf. second part of Rule 2).

[1] http://www.ailab.si/orange/

Based on Rule 4 it would be interesting to explore the way of attributing the sentences to different sources when reporting. It is possible that the presence of *Dr* and the absence of *seems* indicate that local media use more often reported speech, attributing opinions to different sources (we list some examples from the local part of the corpus: *Dr Kofi Annan said; Dr Condoleezza Rice cleared the air over; Dr Alfred Mutua said*), while Western media often use sentences without mentioning the exact source (some examples from Western press: *It all seems likely to get worse; Mr. Odinga seems different; the election seems to have tapped into an atavistic vein of tribal tension*).

4. Contrasting keyword detection

In computer science the term *ontology* denotes a formal representation of a set of concepts of a domain and the relationships among these concepts. Ontologies are organized hierarchically: a concept is divided into a set of sub-concepts. A concept representing a set of documents can be also described by the main topics addressed in the documents. Accordingly, a *topic ontology* (Fortuna, 2007) is a hierarchical organization of documents' topics and their sub-topics.

Semi-automatic topic ontology construction tool OntoGen (http://ontogen.ijs.si/) (Fortuna, 2007) is mainly used for building topic ontologies from unlabeled data, but can be used also for document classification, search, etc. OntoGen is a semi-automatic tool, because it actively supports the user in the ontology construction process.

In OntoGen, hierarchical decomposition of a given set of documents into document subsets is performed by *k*-means clustering and each sub-domain (sub-concepts) is described by the main topics that the documents cover.

OntoGen offers two different ways of getting the topic descriptions. Using the first one (Keywords), the list of keywords is composed of most descriptive words for the document cluster: i.e., *n* most frequent keywords describing the centroid of the document cluster. Using the second one (SVM Keywords), the list of keywords is composed of most contrasting words, best distinguishing the selected concept (document cluster) from its sibling concepts in the hierarchy, where the distinguishing keywords are extracted by the Support Vector Machine (SVM) classifier.

This section describes how OntoGen tool (Fortuna et al. 2007) was used for detecting the differences between the local and the Western press.

The goal of the first experiment was to see which are the subtopics of the local and Western coverage of Kenyan election and post-election crisis. We first created the local and Western category and categorized the documents inside these two categories. We observe the distinguishing keywords, selected by the SVM algorithm implemented in OntoGen, for the local and Western class, respectively. The SVM-based contrasting keywords, best distinguishing between the articles of the two classes, are presented below:

Local:	odm, mp, team, mr, pnu, odm_leader, president_kibaki, dr, media, statement
Western:	kikuyu, mr_kibaki, opposition, mr_odinga, luo, tribe, tribalism, opposition_leader, odinga, ethnic

Figure 4: SVM keywords

These patterns show the difference in a way of referring to main protagonists: local articles use *ODM leader* and *President Kibaki,* while Western media call them *Mr Kibaki, Mr Odinga* and *opposition leader*. Local media mention political parties and functions: *ODM, PNU* and *MP,* while Western media present the election through the ethnic aspect *Kikuyu, Luo, ethnic,* and use even ideologically marked words *tribe* or *tribalism*. The ideological aspects of these stereotyping interpretation frameworks can be found also in Ray (2008).

In the last experiment we analyzed the differences in sub-concepts suggested by OntoGen for the whole data set through differences in topics covered by local and Western media. In Figure 5 we can find four different automatically extracted topics, number of all articles for each topic, number of articles by class, and contrasting keywords. The number of articles informs us which topics were more frequently addressed in the Western and in the local articles. The focus on "Kikuyu, police and town" (Topic 1) and on the two main political candidates (Topic 2) is typical for WE. On the other hand, when the articles talk about ODM, member of the parliament or police the articles are nearly exclusively from LO. Also talking about the Annan's mediation team is the local angle. The SVM keywords can be understood as different views on the same topic e.g. for Topic 1, covering the violence and crisis, the Western media adopt ethnic frame, while Kenyan press is concerned about the consequences of the violence for the people (*child, victim, camp*).

Topic	**Contrasting SVM keywords**	
Topic 1: (135 art.) kikuyu, police, town	WE (101 art.): kikuyu, opposition, luo, ethnic, tribe, protest, valley, rift_valley, kalenjin, burn	LO (34 art.): youth, child, victim, town, food, estate, resident, kisumu, camp, road
Topic 2: (154 art.) mr kibaki, mr odinga, mr	WE (110 art.): mr_kibaki, mr_odinga, opposition, kikuyu, tribalism, odinga, voting, power, opposition_leader, kibaki	LO (44 art.): media, risk, appeal, statement, concern, editor, ranneberger, crisis, provide, report
Topic 3: (87 art.) odm, mp, police	WE (3 art.): mungiki, embakasus, protect, ballot, secret, left, bloodshed, apply, duty, clerk	LO (84 art.): odm, mp, police, mr, party, court, house, law, office, pnu
Topic 4: (88 art.) annan, talk, odm	WE (18 art.): opposition, annan, deal, mr_annan, talk, prime_minister, agree, opposition_leader, prime, mr_kibaki	LO (70 art.): odm, team, pnu, solution, mediation, proposal, discuss, meeting, mr, president_kibaki

Figure 5: Analysis of differences between local and Western coverage of same (automatically generated) topics.

5. Conclusions

The topic of this study is the analysis of the articles covering the 2007 Kenyan elections and post-election crisis, aimed at capturing the differences between local (Kenyan) and Western (British and US) newspaper articles. The main motivation was to use quantitative methods as a starting point for finding interesting observations for further qualitative discourse analysis (we are currently preparing joint article with R. Coesemans from Antwerp Center of Pragmatics).

For finding contrasting patterns, differentiating between the local and Western media, we adopted several methods. Firstly, a simple tag cloud representation offered a basic view of the corpus. Since we were analyzing the articles covering the same topic, the majority of the words were the same. However, words with different weight and specially those present only in one of the two tag clouds were of special interest. The two main methods for finding contrasting patterns, differentiating between the local and Western media were contrast set mining and semi-automated topic ontology construction. We applied also SVM contrasting keywords detection. The study shows that meaningful contrasting patterns can be detected by the proposed methodology and show that the ethnic and tribal aspects are the frameworks of interpretation of crisis in Western media, while political angle is more typical in local media.

6. References

Bay, S. D. and Pazzani, M. J.. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3): 213 – 246, 2001.

Fairclough, N. *Media Discourse*. Arnold, London, 1995.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Communication of the ACM*, 39 (11), p. 27–34, 1996.

Feldman, R. and Sanger, J. *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data.* CUP, New York, 2007.

Fortuna, B., Galleguillos, C. and Cristianini, N. Detecting the bias in media with statistical learning methods. In Ashok N. et al. (ed.), *Text Mining: Theory and Applications.* Taylor and Francis Publisher, 2008.

Fortuna, B., Grobelnik, M. and Mladenić, D. OntoGen: Semi-automatic Ontology Editor. In *Proc. of Human Interfaces*, LNCS 4558: 309–318, Springer, 2007.

Gramberger, D., Lavrač, N. Expert-Guided Subgroup Discovery: Methodology and Application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

Kralj Novak, P., Lavrač, N., Gamberger, D. and A. Krstačić. CSM-SD: Methodology for contract set mining through subgroup discovery. *Journal of biomedical informatics* 42 (1): 113-122, 2009.

Podpečan, V., Žakova, M. and Lavrač, N. Workflow Construction for Service-Oriented Knowledge Discovery. Isola 2010 (to be published).

Pollak, S. Text classification of articles on Kenyan elections. In *Proc. of the 4th Language & Technology Conference*, Poznań, 2009.

Ray, C. How the word 'tribe' stereotypes Africa. *New African* 471, 2008.

Van Dijk, T. A. *Power and the news media.* In D. Paletz (Ed.), Political Communication and Action: 9-36. Cresskill, Hampton Press, 1995.

Verschueren, J. *Context and structure in a theory of pragmatics*. Studies of Pragmatics 10: 13-23, 2008.

Witten, I.H. and Frank, E. *Data Mining Practical Machine Learning Tools and Techniques.* Elsevier, San Francisco, 2005.

Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proc. of the 1st European Conference on Principles of Data Mining and Knowledge Discovery:* 78–87, 1997

Towards a Lexicon of XIXth Century Slovene

Tomaž Erjavec[1], Christoph Ringlstetter[2], Maja Žorga[3], Annette Gotscharek[2]

[1] Department for Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si
[2] Centre for Language and Information Processing, University of Münich
Schellingstrasse 10, 80799 Münich
kristof@cis.uni-muenchen.de, annette@cis.uni-muenchen.de
[3]maja.zorga@gmail.com

Abstract

Historical Slovene texts are being increasingly digitized and made available on the internet in the scope of digital libraries, but so far no language-technology support is offered for processing, searching and reading such materials. Appropriate lexical resources for historical Slovene language could significantly increase such support, by enabling better automatic OCR correction, full-text searching and by modernizing archaic language. This paper describes the first steps in creating a historical lexicon of Slovene, which will map archaic word-forms into modern word-forms and lemmas. The process of lexicon acquisition relies on a proof-read corpus of Slovene books from the XIXth century, a large lexicon of contemporary Slovene language, and LeXtractor, a tool to map historical forms to their contemporary equivalents via a set of rewrite rules, and to provide an editing environment for lexicon construction. The envisioned lexicon should not only help in making digital libraries more accessible but also provide a quantitative basis for linguistic explorations of historical Slovene texts.

Prvi koraki v izdelavi leksikona slovenščine devetnajstega stoletja

Čedalje več slovenskih historičnih besedil je digitaliziranih in dostopnih na spletu v okviru digitalnih knjižnic, vendar zaenkrat še ni na voljo jezikovnotehnološke podpore za obdelavo, iskanje in branje takšnih gradiv. Ustrezni leksikalni viri za historično slovenščino bi lahko z omogočanjem popravkov avtomatsko prepoznanega besedila, iskanja po celotnem besedilu in modernizacijo arhaičnega jezika občutno izboljšali tako podporo. Članek opiše prve korake v razvoju historičnega leksikona slovenščine, ki bo pripisal arhaičnim besednim oblikam sodobne besedne oblike in leme. Proces gradnje slovarja se naslanja na korigirani korpus slovenskih knjig 19. stoletja, obsežen leksikon sodobnega slovenskega jezika in orodje, ki omogoča tako preslikavo historičnih oblik v njihove sodobne ustreznice s pomočjo prepisovalnih pravil kot urejevalno okolje za gradnjo slovarja. Tako zastavljeni leksikon ne bo le omogočil večjo dostopnost digitalnih knjižnic, temveč bo predstavljal tudi kvantitativno osnovo za jezikoslovne raziskave historičnih slovenskih besedil.

1. Introduction

In the context of digital libraries human language technology support can bring increased functionality esp. for full-text search and information retrieval. The most obvious task is automatic lemmatisation of text, which abstracts away from the morphological variation encountered in heavily inflecting languages, such as Slovene. The user can thus query for e.g. *mati* (*mother*) and receives portions of text containing this word in any of its inflected forms (*matere, materi, materjo*, etc.). Support for lemmatisation, as well as morphosyntactic tagging is well-advanced for modern-day Slovene (Erjavec & Džeroski, 2004). However, the situation is very different for historical Slovene, where no such detailed research has yet been carried out for the language.

Historical Slovene language[1] brings with it a number of problems related to automatic processing:

- due to the low print quality, optical character recognition (OCR) produces much worse results than for modern day texts; currently, such texts must be hand-corrected to arrive at acceptable quality levels;
- full-text search is difficult, as the texts are not lemmatised and use different orthographic conventions with different archaic spellings, typically not familiar to the user;
- comprehension of the texts for most users can also be problematic, esp. with texts older than 1850 which use the Bohoričica alphabet.[2]

The above problems would be alleviated by using a large lexicon of historical Slovene language giving the mapping of historical word-forms into their modern-day equivalents with associated lemmas. OCR engines could make use of such a lexicon to guide the recognition process; texts could be lemmatised enabling better search; and the texts could be transcribed using modern day equivalents of the word-forms to facilitate reading.

Developing a lexicon of historical Slovene is a very timely undertaking, as a large number of books and periodicals from the XIXth century are being made available on the internet, e.g. in the context of the dLib.si digital library[3] (Krstulović and Šetinc, 2005) and the Slovene literary classics in WikiSource[4] – Hladnik (2009) gives an overview of digitisation efforts and availability of Slovene texts on the internet.

The lexicon we are developing has a simple structure, where each entry contains the following fields:

[1] In this paper we concentrate on the Slovene language from the XIXth century; the problems are, of course, worse going further back in time, but even here, due to the late development of the written Slovene word and its spelling standardisation, there are substantial differences to contemporary Slovene.

[2] The Bohoričica alphabet had different conventions in writing various Slovene sounds, e.g. »*shaloft*« is the modern-day »*žalost*«, which makes it confusing for today's readers.

[3] http://www.dlib.si/

[4] http://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika

- a word-form that has been witnessed in a proof-read historical text
- the equivalent word-form from contemporary Slovene
- the contemporary lemma of the word-form
- the lexical morphosyntactic properties of the lemma

Compiling such a lexicon with sufficient coverage is a non-trivial process: a representative corpus of proof-read historical texts must be compiled, a comprehensive modern-day lexicon must be obtained, and the word-stock of the former must be matched against the latter. Problems arise in methodological issues (not all historical word-forms even have a modern day equivalent) as well as technological ones (having a good software environment for lexicon construction).

The rest of this paper is structured as follows: in Section 2 we present the language resources we currently use for lexicon construction, in particular the AHLib historical corpus and FidaPLUS contemporary language lexicon; in Section 3 we introduce the LeXtractor software environment used for lexicon construction; Section 4 discusses the main issues so far discovered in building the lexicon; and Section 5 gives some conclusions and directions for further work.

2. The corpus and lexicon

For the historical lexicon of Slovene to be built using the envisaged methodology, three language resources are needed: a proof-read reference corpus of historical texts, a large lexicon of modern-day Slovene, and the patterns of historical spelling variation. In this section we detail the first two resources, and leave the third for the discussion in the next section.

2.1. The AHLib corpus

The corpus we currently use was compiled in the scope of the project *Deutsch-slowenische / kroatische Übersetzung* 1848–1918 (Prunč, 2007). The project addressed the linguistic study of Slovene books translated from German in the period 1848–1918, where a large portion of the effort went towards building a digital library (compiling a corpus) of these translations. To this end, the books were first scanned and OCRed, and then, for a portion of the corpus, the transcription was hand-corrected, marked-up with structural information, and, for a few books, lemmatised; this process was supported by a web interface (Erjavec, 2007).

The subcorpus chosen for building the historical lexicon includes all the AHLib proof-read books written before the year 1900, where the oldest one was published in 1847. There are all together 71 such books, of which the majority (56) are fiction (mostly novels) while 15 are non-fiction (from self-help books for farmers, to text-books on astronomy, chemistry, etc.). All together the corpus contains approximately 2.2 million running words. While certainly small compared to most corpora of contemporary language, it is large and varied enough to enable us to start building the lexicon.

2.2. The FidaPLUS lexicon

The lexicon of contemporary Slovene used was extracted from the FidaPLUS corpus[5] (Arhar and Gorjanc, 2007), a large corpus of contemporary Slovene, where each word was automatically annotated with its morphosyntactic description (MSD) and lemma. The MSDs are compact strings that give the morphosyntactic features of the word form, and can be decomposed into features, e.g. the MSD `Ncmsn` is equivalent to the feature set `Noun, Type = common, Gender = masculine, Number = singular, Case = nominative`.

The lexicon was gathered from the corpus by extracting all the triplets consisting of the word-form, lemma and MSD. The word-forms were lowercased, and the word-boundary symbol added to the start and end of the word (c.f. next section). Using regular expressions, entries with anomalous "words" were removed, and only those lexical items with a frequency greater than 4 were retained. The MSDs were also reduced to the lexical features, e.g. from `Ncmsan` to `Ncm`, which simplifies the task of the lexicographer when adding new words to the lexicon. With this we arrived at a lexicon, which is large enough to serve as a reference lexicon of modern word-forms. The lexicon contains about 600,000 word-forms and 200,000 lemmas.

3. LeXtractor and approximate string matching

This section first explains the general ideas and principles guiding our corpus-based construction of lexica for historical language and then describes a web tool for collaborative construction of historical lexica, initially conceptualized for German (Gotscharek et al., 2009), and its adaptation for the Slovenian lexicon project.

3.1 Corpus based lexicon construction

Given a sufficiently large historical corpus, we ignore all words found in a contemporary lexicon of the language processed, as well as special contemporary vocabulary such as names, geographic expressions, etc. The remaining words are analyzed by their frequency of occurrence in the historical corpus. This frequency-based construction ensures that the lexicon soon enables a reasonable recall over the word tokens that represent historical variants of contemporary words.

In order to minimize the cognitive load of the lexicographers, we employ a number of advanced NLP techniques. Our intention is the following ideal division of work: it is the role of the machine to produce meaningful suggestions of what to include into the lexicon; the lexicographers are enabled to concentrate on the linguistic decision according to the corpus material; they can just confirm or reject the suggestions. In the real production process (cf. Section 4), difficult cases where more complex actions and unsupported input of the lexicographers is needed also occur. In what follows we describe the resources that are used to come close to the idealistic goal.

Word lists for contemporary vocabulary. To separate between contemporary words and historical spellings, a collection of word lists of contemporary vocabulary is used. We use special lists for names and geographic expressions, as well as a large list D^{mod} that covers the contemporary standard vocabulary of the processed language.

List of patterns. For Slovene as well as for other languages, many historical spelling variants can be traced back to a set of rewrite rules or "patterns" that locally

[5] http://www.fidaplus.net/

explain the difference between contemporary and historical spelling. The most prominent pattern for Slovene is *r→er* as exemplified by the pair *brž→beržˇ*. Based on corpus inspection and as a side result of lexicon construction, we currently collected a list P of 57 patterns for Slovene; of these, 26 are for transliteration (e.g., *e→ê* or *š→ſh*), and the rest for "proper" changes in spelling. It should be noted that our patters can also be sensitive to the word boundary, as some spelling changes occur only at the start or the end of the word, e.g. *žganjem→žganjam*, where the inflectional ending *-am* has changed into modern-day *-em*. To enable this functionality, the words in D^{mod} are embedded in a special character (@), e.g. *@žganjem@*, and the appropriate patterns make use of this symbol, e.g. *em@→am@*.

Matching modulo patterns. We employ a tool for matching modulo patterns. The tool uses the word list D^{mod} and the list of patterns *P* as background resources. Given an input token w' occurring in the historical corpus, all entries *w* in D^{mod} are computed where w' can be obtained from *w* by applying one or several patterns. The output list is ranked, preferring candidates *w* where a small number of pattern applications are needed to rewrite *w* into w'. With each suggestion *w* the tool also outputs the set of patterns that are used to rewrite *w* into w'. The tool is implemented as a finite-state device. The lexicon D^{mod} is represented as a deterministic finite-state automaton. For traversal of the automaton, a special procedure has been implemented that takes pattern variation into account, using the list of patterns *P*.

Lemmatizing contemporary word-forms. The output of the above process are one or several contemporary word-forms *w* which correspond to a given historical token *w'*. It remains to assign the correct lemma(s) and part-of-speech (lexical category) to the word-form(s) *w*. A lemmatiser for the processed language is used to map a contemporary inflected word-form *w* to all possible corresponding lemmas. In the case of Slovene, "lemmatizing" of *w* is implemented by the modern Slovene lexicon. The lexicon holds full forms with the lemma and morphological information attached. This enables us to add linguistic features like part-of-speech and morpho-syntactic information to the entry of *w'*.

3.2 A web-tool for collaborative construction of lexica for historical language

A web-based tool was designed to implement the workflow of NLP supported collaborative lexicon construction. The main modules of the web-tool are a managing module, which guarantees that no conflicts arise when several lexicographers simultaneously work on the vocabulary of the corpus, an analyzer module, and the graphical user interface; the latter two are described in more detail below.

Given a historical string *w'* observed in the corpus, the analyzer module first suggests corresponding contemporary word-forms *w* from the contemporary lexicon D^{mod} based on matching. Each interpretation *w* comes with the set of patterns that were applied. Second, for a given contemporary word-form *w*, the analyzer computes all the lemma(s) – including part-of-speech information – which may underlie the word-form *w*.

The confirmed entries for the historical lexicon are stored in a special database. Standard entries of the database consist of the historical string as found in the corpus, the corresponding contemporary word-form and lemma, the part-of-speech category, pointers to concordances[6] in the historical corpus which serve as attestations for the given interpretation, and the name of the person who created the entry. Note that a historical string can be associated with several entries of the database. The database also contains "non-standard" entries such as named entities, abbreviations, and historical words that do not have a corresponding contemporary lemma.

The graphical user interface visualizes the different workflows to create lexicon entries of the words in the corpus. Figure 1 shows the frequency list mode with the pattern based strings on the left and the non-derivable strings on the right hand side. If the user selects a token *w* of the left list, she is taken to a new screen that visualizes the possible interpretations of *w*. By an *interpretation* we mean a pattern based derivation of *w* from a valid contemporary word-form (cf. Figure 2). The user now confirms or rejects the proposed interpretations. For each confirmed interpretation, the linguistic readings in terms of the corresponding lemma(s) have to be determined.

Fig. 1 GUI for collaborative lexicon building, corpus mode. Unchecked word-forms derivable by patterns from contemporary words are presented in the left column, ordered by frequency; the non-derivable word-forms are in the right-hand column.

[6] A *concordance* is a text window containing an occurrence of the word form.

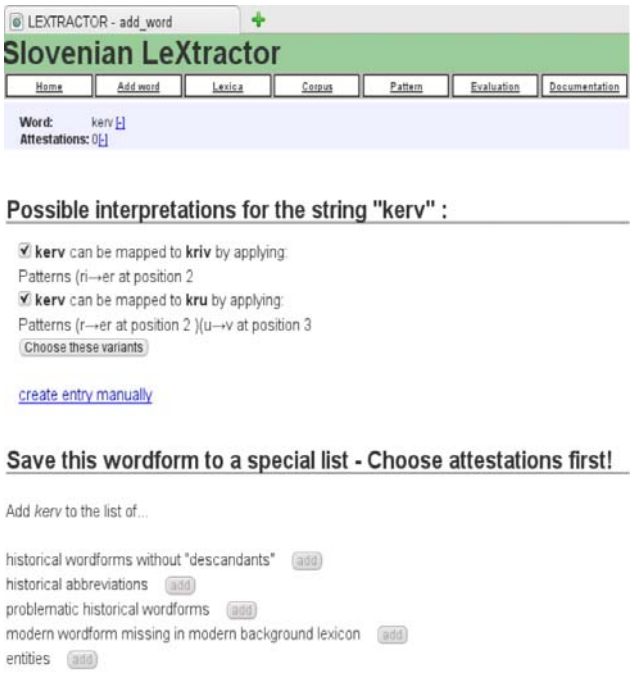


Fig. 2 Selecting possible interpretations for "*kerv*". The system suggests "*kirv*" and "*kru*". Alternatively, "*kerv*" can be added to special lists visible in the lower part.

In our case, readings based on the contemporary lexicon are suggested by the system. Each reading is confirmed or rejected. Before the lexicographer may confirm a reading she has to select at least one attestation, i.e. a concordance where the reading in question is the correct one. To this end, all concordances are shown graphically (cf. Figure 3).

For every confirmed reading, a separate lexicon entry is created that includes the associated attestations. If a processed string has other than pattern based mappings to a contemporary word form or lacks a contemporary explanation, it is included into one of the following *special sublexica*: historic words without a contemporary equivalent; historic abbreviations; historic word-forms which lack a simple transition pattern; named entities; missing words of the contemporary lexicon (cf. Figure 2, lower part).

Entries of the right frequency list are more complicated because they are not rule-based variants in terms of patterns. The system can't suggest the mapping of a historical string *w'* to its contemporary equivalent *w* automatically, so *w* has to be specified manually. If necessary, the lexicographer can also assign these entries to the special lexica mentioned above. If the lexicographer sees a derivation from a contemporary word using a *new* pattern *p* she can suggest to add *p* to the list of patterns. In the current version, there are no automated update mechanisms for the list of patterns and the matching procedure.

Document mode. The lexicographer may also decide to work on a specific text. On the basis of the current lexicon and the matching rules, the text is visualized with all words marked according to their lexical explanation (cf. Figure 3). Additional information is provided through mouse events. We distinguish: contemporary words, checked entries of the lexicon for historical word-forms, entries of the left (right) frequency list shown in the corpus mode, and non-explained strings. If a string is activated in the document mode, the sequential processing is the same as for the corpus mode.

The system is web based and collaborative. Both issues are of great importance for the project. As the involved lexicographers do not work at the same location, flexibility concerning their individual workplaces is needed. Since the professional abilities of the contributors and the complexity of certain lexical entries differ significantly, a workflow was created that leaves more challenging entries in the frequency list to the trained historical linguists, whereas the other lexicographers deal with the simple cases.

From our present perspective, corpus, matching rules, and lexicon should be considered as a joint knowledge base. Given a set of patterns we may use historical word-forms and corresponding contemporary word-forms stored in the lexicon and in addition the corpus for deriving meaningful probabilities or edit weights for the patterns. As a matter of fact, frequency based lexicon construction also helps to find new relevant patterns. In this sense, lexicon and corpus provide empirical evidence for patterns (rules) and help to fine-tune approximate matching. Conversely, we have seen above how refined matching procedures help to speed up lexicon construction. Summing up, this shows that refinement of matching procedures and lexicon construction can be directly interleaved in a kind of bootstrapping procedure.

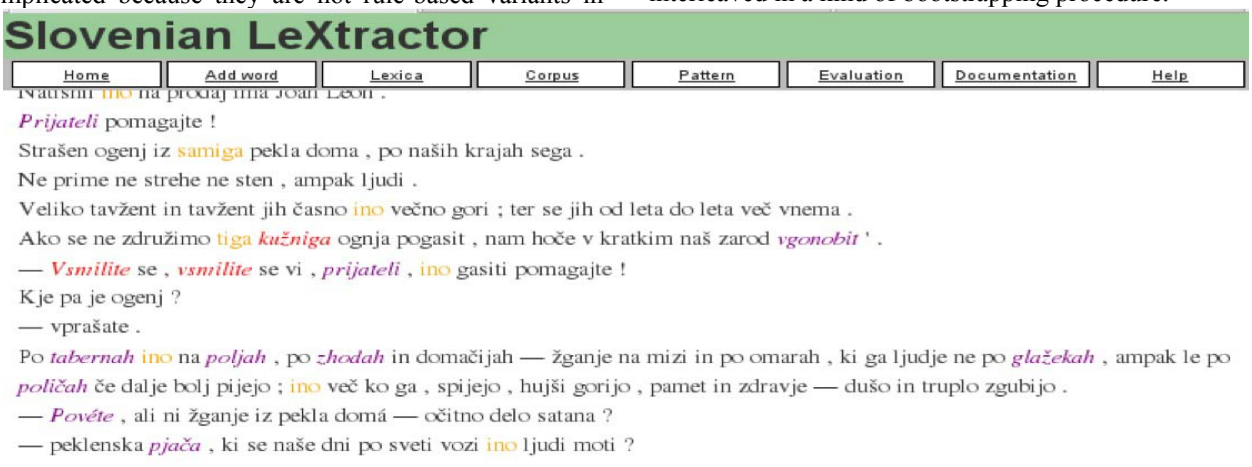


Fig. 3 GUI for collaborative lexicon building: document mode with highlighting, in which different types of words are presented in different colours.

3.3 Adapting the system to Slovene

In addition to integrate language specific background resources, as for example, the rewrite pattern set and the modern lexicon into the system, internal software adjustments were also necessary. To start lexicon building, the tokenizer had to be enabled to cope with special characters, such as *ſ* and *Ş* introduced by the alphabet of historical Slovene. Furthermore, the software had to be adapted to the specific format of the Slovene modern lexicon that differs from the format used for the German variant of the lexicon building tool. Since the Slovene rewrite patterns include word boundaries, the contextual treatment of patterns had to be extended accordingly. During the first round of lexicon production a list of further issues turned up which is reported on in the next section.

4. Discussion

Currently only a few hundred word-forms have been added to the lexicon of historical Slovene; rather than going for quantity, we first concentrated on the various types of issues that arise in lexicon construction. In this section we discuss the main problems – and solutions – that we have encountered in our work so far. Below we provide a typology of cases which arise in the construction of the lexicon. The first five are already a part of LeXtractor itself:

1. *Historical word-forms without descendants*: such is a case with pairs of similar words, of which only one survived in modern Slovene; the other is thus included under this category. For example, both *pervle* and *pervič* (*firstly*$_{[adv.]}$) existed in 19th century, but only *prvič* is used in modern Slovene.
2. *Problematic historical word-forms*: although there is such a proposed category in LeXtractor, we are currently avoiding including words in it, as we are still determining the general methodology of how to deal with such cases.
3. *Named Entities*: when entities are pattern-based, they pose a specific problem. Such was the case with the given name *Ménart*. Since modern Slovene (usually) does not include diacritical marks in writing, patterns for diacritic removal were added, so that the word-form can be found in the modern background lexicon. The LeXtractor tool does not allow for adding the word-forms to the list of entities once a pattern has been chosen. This leads to making a list of attestations where a word-form is used as a name, adding the attestations under sub-lexicon Entities, and manually adding the modern string without the diacritical mark, in our case, *Menart*.
4. *Modern word-form missing in modern background lexicon*: some word-forms are still alive in modern (though not necessary standard) Slovene, but are missing from our contemporary background lexicon. Such is the case of the word *tavžent*, a deformed German word *tausend* (*thousand*$_{[num.]}$), that is missing from the background lexicon, even though the word is very much alive in spoken Slovene. Another example is word-form *Ogerska* (*Hungary*$_{[sg.,loc.]}$), which only exists as an adjective in the background lexicon, and not as a geographical name, as is the case in historical corpus. The proposed solution is either to add these words in modern background lexicon, or to have an edit option in LeXtractor, which would allow the lexicographer to add MSD information to the word-form.
5. *Identical word-forms*: in the otherwise trivial case when a historical word-form exactly corresponds to a modern word-form it can happen that the entry in the lexicon is a false friend; for example, *serca* is an archaic form for *srca* (*heart*$_{[sg,gen]}$), but at the same time is identical to a form of the contemporary lemma *serec* (*horse of a gray colour*). The problem is solved when we ascribe a pattern to the word-form, *r→er*, which transcribes *serca* into a univocal *srca*.
6. *Missing readings*: sometimes the correct MSD is missing in the background lexicon, and is consequently also missing in LeXtractor. Such was the case with word-forms *dobé* and *mrtve* that represent two different approaches of dealing with such problems. The word-form *dobé* was transcribed into *dobe* (both *they get* and *era*$_{[sg.\ gen.\ \text{or}\ pl.\ nom.]}$), but the background lexicon only offered the latter reading, i.e. *Ncf: noun, common, feminine*. We decided to change the pattern (*ijo→e*) and modernize the otherwise possible, but in contemporary Slovene archaic verb form *dobe* into a more common modern form *dobijo*. This extracted the proper reading for *dobé* from the background lexicon. The second possible scenario of a missing MSD is much more complicated. The possible solution for it is the same solution we propose for modern word-forms missing in modern background lexicon. Sometimes a historical word-form was used differently than modern word-form, which means that the modern background lexicon cannot offer the missing MSD. For example, the word-form *mrtve*, it is not only an adjective of a feminine plural for *dead*, it is also an accusative plural form for a masculine noun *mrtvi* (*the dead*), as well as a nominative plural form for a feminine noun *the dead*. This last use, however, is foreign to contemporary Slovene, which only uses masculine plural noun *mrtvi*. An edit option, with which the lexicographer could manually add the missing MSD information, would again be needed.
7. *Historical word-form corresponds to more than one modern word-form of the same lemma*: such is a case with the word-form *veči*, that doesn't only transcribe into *večji* (*bigger*), but also into *večja*, *večje* and *večjo* (*veči žejo*/*večjo žejo*, *veči groze*/*večje groze*, *veči del*/*večji del*, *veči nesreča*/*večja nesreča* etc.). The solution to this is to first add the suggested readings and attestations and then create additional entries manually. The lexicographer needs to pay special attention with such cases, because the entry procedure must be performed in a single step: once we stop working on the word-form *veči* and work on another word-form, the only way to return *veči* is to destroy all the entries for it, recalculate and start again. Once again an edit option would be needed.
8. *Change of inflection*: a form of the missing reading or MSD is an inaccurate reading. Sometimes the word form for a specific declination has changed during

time: for example, the word-form *serci* (from *srce, heart*), was used both as the nominative and accusative dual form, *najne serci* (*our two hearts*), as well as the historical singular dative and locative form, *k serci*, *pri serci*. Contemporary Slovene uses the form *srci* only for the first case (and as instrumental plural form, though no such attestations were found in historical corpus), whereas the modern singular dative form is *srcu*. The dilemma that arises is this: should a new pattern, and hence a new reading, be added transcribing *u→i*, so that the full and correct MSD can later be extracted not only for the dual but also for the singular dative form? If this is not done and the lexicographer just ascribes the information, that *srci* is a common neutral noun, later, more specific MSD extraction could not recognize that *srci* is also a historical singular dative form.

9. *The background lexicon offers too many possible readings*: another frequent problem that needed systematical solving is a case when a word-form, transcribed into a modern word-form, offers more grammatical readings (MSDs) than the historical corpus shows are needed. For example, the word form *ravna* only appeared in the historical corpus as a form of a verb *ravnati* (*to straighten*), even though it could also be a feminine adjective form of *raven* (*straight*). There are two possible solutions for such a case. One is to exclude the reading for adjective because the word-form historically did not exist as an adjective, the other is to mark that there were no attestations found, even though the word-form *ravna* already existed as an adjective. We have opted to exclude the reading when the word as such doesn't appear in Pleteršnik's dictionary, published in 1894—1895, and on the other hand mark that there were no attestations found for the proposed reading, if the word does appear in his dictionary, as was the case for *ravan.* In the future, when texts before 1847 are added, it would also be wise to include cited older dictionaries as a reference.
10. *Historical word-form is written separately, modern word-form is not*: sometimes, historical word-forms are written separately, even though in modern Slovene they are written as one word. Such is the case of compound words *najprej* (*firstly*), historically *nar pervo*, *zase* (*for him/her/itself*), historically *za-se*, and *čezenj* (*over him*), historically *čeznj*, *čez-nj*, and in one instance also *čes-nj*. Words with the prefix *nar* were sometimes written separately, like *nar pervo*, and sometimes together, *narbolj* (*mostly*). Since some compound words were written with a hyphen and others were not, a possible solution would be to merge the prefix with the following word, so that LeXtractor recognizes the compound word in the background lexicon, with or without applying patterns. A solution more complex to implement would be for the lexicographer having the option of marking that two words actually form one and ascribe modern compound word to the unit.

5. Conclusions

The paper presented the first steps in building a lexicon of XIXth century Slovene, using a historical corpus, a contemporary lexicon of Slovene, spelling variation patterns, and the LeXtractor software. So far we have mostly concentrated on setting up the resource and program environment and methodological issues, which have been discussed in the present paper.

In further work we plan to intensively start adding entries to the lexicon, extend the corpus, esp. with newspapers and older books, as well as address the remaining methodological issues, such as tokenisation, which, as discussed, can be different in historical words from their contemporary equivalents.

Current work has also been exclusively empirically driven, i.e. we addressed only issues that directly arise out of the lexical items found in the corpus. In the future we plan to take into account the linguistic research that has been done so far on historical Slovene language, as discussed e.g. in Orožen (1996). Maybe our computational approach might also reveal new quantitative and qualitative linguistic insights into the language as used in XIXth century Slovenia.

Acknowledgements

The work presented in this paper was supported by the EU FP7 ICT project IMPACT, "Improving Access to Text".

References

Špela Arhar and Vojko Gorjanc. 2007. Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. [Corpus FidaPLUS: a new generation of the Slovene reference corpus] *Jezik in slovstvo*, 52(2).

Tomaž Erjavec and Sašo Džeroski. 2004. Machine Learning of Language Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1):17–41.

Tomaž Erjavec. 2007. Architecture for Editing Complex Digital Documents. Proceedings of the Conference on Digital Information and Heritage. Zagreb. pp. 105-114.

Annette Gotscharek, Andreas Neumann, Ulrich Reffle, Christoph Ringlstetter and Klaus U. Schulz. 2009. Enabling Information Retrieval on Historical Document Collections - the Role of Matching Procedures and Special Lexica. *Proceedings of the ACM SIGIR 2009 Workshop on Analytics for Noisy Unstructured Text Data (AND09)*, Barcelona.

Miran Hladnik. 2009. Infrastruktura slovenistične literarne vede [Infrastructure of Slovene Literary Studies]. In *Obdobja 28 – Infrastruktura slovenščine in slovenistike*. pp. 161–69.

Zoran Krstulović and Lenart Šetinc. 2005. Digitalna knjižnica Slovenije – dLib.si. [The digital library of Slovenia – dLib.si] *Informatika kot temelj povezovanja: zbornik posvetovanja*, pp. 683-689.

Martina Orožen. 1996. *Oblikovanje enotnega slovenskega knjižnega jezika v 19. stoletju.* [The formation of a unified Slovene literary language in the XIXth Century.] Ljubljana, Filozofska fakulteta.

Erich Prunč. 2007. Deutsch-slowenische/kroatische Übersetzung 1848-1918 [German-Slovene/Croatian translation, 1848-1918]. *Ein Werkstättenbericht. Wiener Slavistisches Jahrbuch 53/2007*. Austrian Academy of Sciences Press, Vienna. pp. 163-176.

Recognition of odonyms in Serbian language

Staša Vujičić*, Duško Vitas*, Miloš Utvić †

* Faculty of Mathematics
Studentski trg 16, 11 000 Belgrade, Serbia

† Faculty of Philology
Studentski trg 3, 11 000 Belgrade, Serbia
{stasa, vitas, misko}@matf.bg.ac.rs

Abstract

In this paper we present the problem of recognizing street names and other odonyms in ads in Serbian language in order to develop a system which automatically serves the users who make an offer or a demand. The experiment was limited to ads that are related to supply and demand of residential objects. The complexity of the model of representation of street names is, on one hand, due to them being multiply nested named entities, and on the other, to differently motivated variations in the structure of the same name. This paper presents the models of street names created through specially developed lexical resources and finite transducers in the Unitex system, as well as results of experiments of their application on the corpus of contemporary Serbian language and series of ads.

1. Introduction

Recognition of names of the streets, boulevards, avenues, roads, etc., is an integral part of the problem of recognition and categorization of named entities (Chinchor et al., 1999). In lexical analysis, they appear as multiword units or contingent sequences of more simple words. Their recognition as street names, apart from that as named entities, reduces the number of ambiguities in the text. For example, in the name of the street *Knez Mihailova ulica*, Knez Mihailo is not the owner of the street, but a significant figure in the Serbian history after whom the street got its name. The paper discusses the problem of recognition of street names in ad text and the corpus of contemporary Serbian language[1] in order to construct transducers that will annotate the recognized names with corresponding XML tags, in accordance with (Chinchor et al., 1999). In addition to recognizing the very street names, special transducers have been developed in the Unitex system that also recognize adjectives for a place that more closely describe the locations referred to in the ads. Identified names and complements of place are then used as input data for a query in a geographic information system in order to try to find the matching pairs of ads, if they exist, and thus meet the needs of users.

The class of street names is limited in the experimental phase, primarily to Serbian names. The corpus contains foreign names only in the exceptional cases. Odonyms of other languages are rarely found in the corpus of contemporary Serbian language, and the ones that are mostly well known odonyms such as *Menhetn* (*Manhattan*) or *Jelisejska polja* (*Elysian Fields*).

In the construction of transducers, lexical resources developed for Serbian language were used, especially dictionaries of proper names (Krstev et al., 2005a) and local grammars that have been developed by the bootstrap method proposed in (Gross, 1999), based on the analysis of the corpus of contemporary Serbian.

[1] http://www.korpus.matf.bg.ac.rs

2. Description of the problem

The names of streets in the Serbian language are mostly constructed from the names of famous people, geographic concepts or important dates, so that identification of street names means identification of one or more nested named entities. Only exceptionally are those names neutral in relation to named entities (such as, for example *Slavujev venac* (*Nightingale's wreath*), *Cvetni trg* (*Flower square*) or *Cvetna ulica* (*Flower street*).

The complexity of the problem of recognizing street names in the Serbian language stems from several sources rooted in the system of Serbian language.

According to the orthography of Serbian language, if the street name begins with the word *ulica* (*street*), the initial letter needs to be capitalized, thus ***U**lica*; but if the word ***u**lica* is not in the first place, it is then written in lower case: ***U**lica Simina* and *Simina* ***u**lica*. It is often deviated from this spelling rule in written texts, so the form ***u**lica Simina* can also be found in the corpus of modern Serbian language. Also, spelling of some personal names is subject to fluctuations. The proper name *Mihailo* is often written as *Mihajlo* (e.g. *Bulevar Miha (i + j) la Pupina*).

A street name can be written in Cyrillic or Latin alphabet, which in some cases leads to ambiguity in interpretation of the Latin name, as in the examples *Ulica Viljema Šekspira* (*William Shakespeare's street*) where *Viljem* has two Cyrillic representations (*Вилјем* and *Вильем*) or *Ulica Kralja Petra I Oslobodioca* where symbol *I* can be interpreted as a Roman numeral or conjunction *i* (*and*).

Serbian derivative system in many cases allows the replacement of the structure *ulica N* with *A+Pos ulica*, where *N* is a noun or noun phrase, and *A+Pos* the appropriate possessive or relational adjective. For example, *Ulica Kneza Mihaila* can be replaced by *Knez-Mihailova ulica*. Moreover, in such cases, it is often possible to omit the word *ulica*, so that only the adjective refers to the odonym (*Knez-Mihailova*), which introduces additional ambiguity. This possibility creates further complexity. If the street name is followed by a noun phrase, one of its parts is usually selected to move to the front of the name of the street. Thus, for example, *Ulica*

Ilije Garašanina becomes *Garašaninova ulica* (possessive adjective of the last name), but *Ulica Vase Čarapića - Vasina ulica* (possessive adjective of the first name).

The street names that include toponyms consist a special problem. As a rule, the structure of these names is *A+Rel ulica*, where *A + Rel* is an appropriate relational adjective. For example, *Francuska ulica* (*French street*). In case the toponym is a multimember word, a complex transformation occurs within the string that refers to the toponym. For example, in the name of a street *Novosadska ulica* (*Novi Sad street*), *novosadska* is a relational adjective derived from a complex toponym Novi Sad (Utvić, 2008). In Serbian language, the possessive adjectives (for example, *novosadska*) are written with a small initial letter, except in this very case because of the orthography rule that the street names are written with an initial letter capitalized (*Novosadska ulica*). Considering the rich inflexion system of Serbian language, street names are subject to transformation rules of multimember words, where it is necessary to take into account the complex conditions of matching.

In addition to the problems listed, street names are subject to frequent change. Only 30 of over 5,000 streets in Belgrade have borne the same name for more than a century, while during the last four years more than 500 changed their name. Therefore, a custom model of description of the semantic relationships, based on the one developed within the project Prolex (Maurel et al., 2006), can be applied to identification of the street names indicating the same local toponym. Name change may apply to the change of the odonym type (e.g. *Ulica 29. Novembra (The street of the 29th of November)*, became *Bulevar despota Stefana (Boulevard of Despot Stefan)*), but we should also keep in mind the cases of the exact same name occuring in two types of odonyms, indicating different locations (e.g. *Bulevar Nikole Tesle (Boulevard of Nikola Tesla)* and *Ulica Nikole Tesle (Nikola Tesla's)*).

3. Models

The first step in constructing a model for recognition of street names was creation of the dictionaries of possible street names.

Lexical resources developed for the Serbian language that are used for recognition are dictionaries of proper names (Krstev, Vitas and Gucul, 2005a), dictionaries of celebrities and a general dictionary of Serbian language. Below is an overview of the dictionary of proper names:

```
Vasa,Vasa,N1741+Hum+NProp+First+Nick+SR
Cyarapicx,Cyarapicx,N28+NProp+Hum+Last+SR
Cetkin,Cetkin,N1002+NProp+Hum+Cel+Hist
Bonaparta,Bonaparta,N1685+NProp+Hum+Cel+Hist
Klara,Klara,N1637+NProp+Hum+First+EN+Val=Clara+
```

The entries in the dictionary consist of a word form, lemma and part of speech and inflexion information. N stands for a noun, NProp stands for a proper name, Hum stands for human, First for the first name, Last for the surname, Nick for the nickname, Cel for celebrity and so on.

Here is an excerpt from the corpus of modern Serbian language with odonyms:

```
Preko Autokomande, pa u Ulicu Maksima Gorkog.

Potrcya niz ulicu. Nasta prava pometnxa.

Prodje zaobilazno, sporednim ulicama.

Iz jedne beogradske ulice, krajem osamdesetih

Preko Knez Mihailove ulice do Malog Kalemegdana.

Klub iz ulice Zdravka Cyelara.

Simbol Knez-Mihailove ulice sxezdesetih godina.
```

The first thing to notice in the structure of street names in the corpus of contemporary Serbian language is that odonyms may appear after the words street (*ulica*), road (*put*), loop (*petlja*), alley (*sokače*), boulevard (*bulevar*), pier (*kej*), coast (*obala*), square (*trg*). In exceptional cases, when odonyms are of foreign origin odonyms words like avenue (*avenija*) or canal (*kanal*) can also be found.

The first structure that we can distinguish is the one with personal nouns. They appear, for example, in the following names: *Terazije*, *Slavija*, *Zeleni venac*.

The next structure that we see in the analysis of text and speech is in the form of the name of odonym, followed by the name of a famous person in genitive, or very similar forms of the name of odonym followed by names of several celebrities in genitive. So we have, for example, *ulica Lole Ribara*, *bulevar Milutina Milankovića*, *trg Nikole Pašića*. Particular attention should be paid to constructing a genitive form of names of male and female celebrities, which is discussed in detail in the paper (Krstev, Vitas and Gucul, 2005b). As synonymous to this structure, there is also a structure built from a possessive adjective resulting from the name of celebrities, optionally followed by the name of an odonym. For example, *Dositejeva*, *Simina*, *Dobračina*.

Another distinguished structure is the form of the name of odonym followed by the title and the name of the famous person in genitive, where more celebrity names can also appear. For example, *ulica Kneza Mihaila*, *ulica majke Jevrosime*, *ulica braće Jugovića*, *ulica poručnika Spasića i Mašare*.

There are two ways to present the structures of the form: name of an odonym, date, and name of an odonym, number and noun. Namely, it is allowed to write numbers which are part of the structure as words and signs for numbers, where one should also take into account whether the number is an ordinal, in which case it should be followed by a period. For example, *ulica 27. marta* (*The street of the 27th of March*) or *ulica Dvadeset i sedmog marta* (*The street of the twenty-seventh of March*), *ulica 1300 kaplara* (*The street of 1,300 corporals*).

The structures constructed from toponyms are built with relational adjectives. These structures are in the form: relational adjective, name of an odonym. Thus, for example, the street name *Makedonska ulica* is constructed from a toponym *Makedonija*, and similarly, *Mostarska petlja* is constructed from *Mostar* and *Balkanska ulica* from *Balkan*.

4. Presenting the models with finite automata

Odonym names used in the construction of other structures were placed in a separate subgraph **odonimi.grf,** given in the Figure 1.

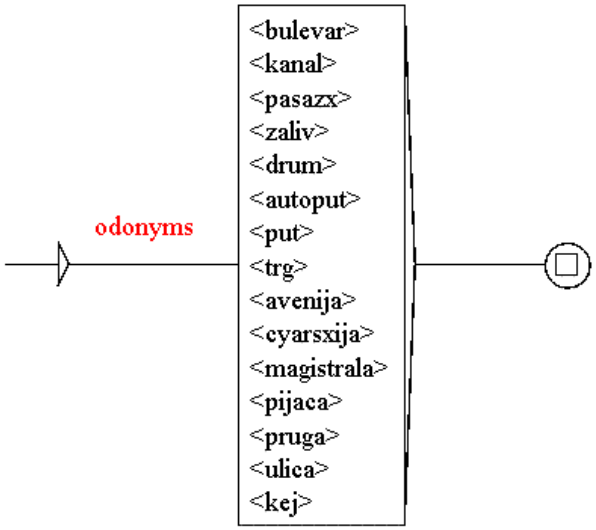


Figure 1. odonimi.grf

In recognition of street names, a structure that can optionally follow the structures described in Chapter 3 can also be used, and it is separated into a special subgraph **broj.grf**. This graph recognizes the structures of type *u broju 3 (in number 3), na broju 3 (on number 3), u br 5 (in No 5), u br. 8 (in No. 8), broj 56 (number 56), br 9 (No 9), br. 80 (No. 80), 57* .

Also, when constructing the above described structures, we use subgraphs **datum.grf** and **imena.grf** that describe dates or names of famous people in genitive. Dates are constructed from ordinal numbers (the number followed by a period) and names of months, or ordinal numbers written in letters followed by the name of the month. In the graph **imena.grf** we use names of famous people, which are tagged with semantic categories *+NProp* (proper name) and *+Hum* (human) in the dictionary.

Therefore, the graph **ulicaPoznataLicnost.grf** is built from subgraphs **odonimi.grf** and **imena.grf**, and graph **ulicaTitulaPoznataLicnost.grf** from subgraph **odonimi.grf**, titles *<N+PropHum>* and subgraph **imena.grf** where conjunctions can also appear.

The graph **RelacioniPridevUlica.grf** is built from the relational adjective of a toponym *<A+Rel+Nprop+Top>* followed by a subgraph **odonimi.grf**. and the graph **PrisvojniPridevUlica.grf** is built from the possessive adjective of a famous person *<N+NProp:2>*, followed by a subgraph **odonimi.grf**.

The last step in this phase of the experiment is to expand the developed finite state automata that recognizes street names so it can tag the appearance of street names in the text with XML labels, in accordance with (Chinchor et al., 1999).

When such expanded graphs are applied to the corresponding resources, while prepositions deciding whether something is a street name or not are also included in the analysis, the result is the output with tagged street names, as shown in an excerpt below:

```
iz ulice Zdravka Cyelara<PREP+p2>iz</PREP+p2>
<ENAMEX>ulice Zdravka Cyelara</ENAMEX> prvo se
 pilxarica zxena  iz Ulice 27.
marta<PREP+p2>iz</PREP+p2> <ENAMEX>Ulice 27.
marta</ENAMEX> broj 25 a.
i 27 marta 1941 na ulice Beograd <ENAMEX>ulice
Beograda</ENAMEX>  i drugih gradova
Na uglu ulice Milena <ENAMEX>ulice
Milena</ENAMEX> se drzxala za ruku
na raskrsnici ove ulice i Ulice Slobodana
Penezicxa Krcuna <ENAMEX>Ulice Slobodana
Penezicxa Krcuna</ENAMEX>
kroz tunel koji vodi do Ulice Teodora
Drajzera<PREP+p2>do</PREP+p2> <ENAMEX>Ulice
Teodora
```

5. Application to the searching of ads

This paper is limited to ads in which housing is demanded or offered, regardless of whether it concerns apartments, houses, rooms, offices, garages, etc. and the possibility to answer three questions: "*Who-whom?*" , "*what?*" and "*where?*". In order for the latter stage of the experiment to offer full functionality, i.e. to synchronize the customer needs and offering: the very type of users "*who-whom*?" (male, female, family ...), in the form in which the ads are entered it is specified: whether the user offers or demands, the type of the object in question "*what?*" (room, apartment, house ...) and the address where the facility is located, including the details that further determine the user's wishes, in response to "*where?*": "*blizu*" (*near*), "*u blizini*" (*close to*), "*u krugu*" (*within*), "*kod*" (*at*), "*preko puta*" (*opposite*). This gives the following format of the ad:

```
<oglas id="1">
     <namena> ponuda/potražnja </namena>
     <objekat>    soba/    stan/    kuća…
     </objekat >
     <tekst>…</tekst>
 </oglas>
```

As it is presented at the beginning of this paper, the street name recognition, but also further clues on the location, are provided using the system Unitex. In addition to Unitex, which was used to identify street names in the text, GIS (Geographic Information System) will be used in the later stages of the experiment, in which maps of the city center with marked streets will be included. The references such as "*u blizini*" (*close to)*, "*oko*" (*around the*), "*u krugu*" (*within*), ... allow the GIS system to call the appropriate functions and get the list of street names that match the required criteria, which can later be used in the re-searching of ads in order to find those ads that match the user's enquiry fully or with a certain threshold of tolerance.

By analyzing the ads with housing offers and demands, we concluded that users often do not provide completely accurate street names that would satisfy their needs, but rather use descriptions of locations in relation to the major objects in the certain area. Such objects are often described by colloquial names rather than formal, official names. For example, "*kod bloka 70*" (*at block 70*) or "*kod Piramide*" (*near the Pyramid*), define the same area of the city, which could be described more formally by the street name "*u ulici Jurija Gagarina*" (*in the Jurija Gagarina's street*).

The next step in the experiment would be to incorporate the detailed map of Belgrade into the GIS system, which, in addition to the marked streets, would contain information about facilities and points of interest, whether they are sights (libraries, monuments, fountains,

parks, forests ...), or shopping centers, markets and points that are referred to by their non-official names in everyday language, and to provide name recognition of such objects in texts of different types of ads.

6. Conclusion

In this paper we presented the basic structure of models for recognition of street names in the Serbian language and their practical application. It should be noted that, although currently achieved results are satisfying, work on this issue is at an early stage and has yet to be developed in future. In the later stages of the experiment, models for identifying street names should be enriched, more detailed maps of the city should be included and the possibilities of its search should be extended.

7. References

Chinchor, Nancy; Brown, Erica; Ferro, Lisa; Robinson, Patty. 1999. 1999 Named Entity Recognition Task Definition (version 1.4). Technical Report, SAIC, http://www.nist.gov/speech/tests/ie-er/er_99/doc/ne99_taskdef_v1_4.pdf.

Friburger, Nathalie; Maurel, Denis. 2004. Finite-state transducer cascades to extract named entities in texts. Theor. Comput. Sci. 313(1): 93-104

Gross, Maurice. 1999. "A bootstrap method for constructing local grammars". In Contemporary Mathematics. Proceedings of the Symposium, 18-20 December 1998, Belgrade, Serbia, N. Bokan (ed.), University of Belgrade, pp. 229-250.

Krstev, Cvetana; Vitas, Duško; Maurel, Denis; Tran, Mickael. 2005a. Multilingual Ontology of Proper Names. In Proc. of Second Language & Technology Conference, Poznań, Poland, April 21-23, Wydawnictwo Poznańskie Sp. z o.o, Poznań

Krstev, Cvetana; Vitas, Duško; Gucul, Sandra. 2005b. Recognition of Personal Names in Serbian Texts. In Proc. of the International Conference Recent Advances in NLP RANLP, 21-23 September 2005, Borovets, Bulgaria, eds. G. Angelova et al.

Maurel D., Tran M., Vitas D., Grass T., Savary A. 2006. Prolex: Implantation d'une ontologie multilingue des noms propres, Rapport interne du Laboratoire d'Informatique de l'Université François-Rabelais de Tours, n°286, 47 p.

Utvić, Miloš. 2008. Konačni automati u regularnoj imenskoj derivaciji. Matematički fakultet. Beograd

Automatic Construction of Wordnets by Using Machine Translation and Language Modeling

Martin Saveski*, Igor Trajkovski†

* Faculty of Computing, Engineering and Technology,
Staffordshire University,
College Road, Stoke-on-Trent, Staffordshire, UK
saveski.martin@gmail.com

† Faculty of Electrical Engineering and Information Technologies,
Ss. Cyril and Methodius University,
Rugjer Boshkovik bb, PO Box 574, Skopje, Macedonia
itrajkovski@feit.ukim.edu.mk

Abstract

WordNet is one of the most valuable lexical resources in the Natural Language Processing community. Unfortunately, the benefits of building a WordNet for the Macedonian language have never been recognized. Due to the time and labor intensive process of manual building of such a lexical resource, we were inspired to develop a method for its automated construction. In this paper, we present a new method for construction of non-English WordNets by using the Princeton implementation of WordNet as a backbone for their construction along with Google's translation tool and search engine. We applied the new method for construction of the Macedonian WordNet and managed to develop a WordNet containing 17,553 words grouped into 33,276 synsets. However, the method in consideration is general and can also be applied for other languages. Finally, we report the results of an experiment using the Macedonian WordNet as a means to improve the performance of the text classification algorithms.

Avtomatska izdelava wordneta z uporabo strojnega prevajanja in jezikovnega modeliranja

Wordnet velja za enega najbolj uporabnih leksikalnih virov na področju računalniške obdelave naravnega jezika, vendar za makedonščino še ne obstaja. Ker je ročna izdelava tovrstnega vira izjemno dolgotrajna in draga, smo se odločili za gradnjo z avtomatskimi pristopi. V prispevku predstavljamo metodo za izdelavo wordneta v izbranem ciljnem jeziku, pri čemer izhajamo iz angleškega Prinecton WordNeta, za generiranje sinsetov pa uporabimo dvojezični slovar, Googlov spletni strojni prevajalnik in iskalnik. Čeprav je na ta način mogoče izdelati wordnet za kateri koli jezik, smo v pričujoči raziskavi generirali makedonski wordnet, ki vsebuje 17.553 besed oz. 33.265 sinsetov. Izdelan wordnet tudi preizkusimo na sistemu za avtomatsko klasifikacijo besedil in s tem preverimo njegovo uporabnost v praksi.

1. Introduction

WordNet (Fellbaum, 1998) is a lexical database for the English language. It groups the English words into sets of cognitive synonyms (synsets) which represent different concepts. Each synset contains a gloss (explanation of the concept captured by the synset) and links to other synsets, which define the place of the synset in the conceptual space.

The public release of the Princeton WordNet (PWN), encoding the English language inspired many researchers around the world to develop similar lexical resources for other languages. As of today, there have been more than sixty WordNets built worldwide for more than fifty languages[1]. Moreover, WordNet had become an ideal tool and source of motivation for researchers from various fields. A plethora of applications which use WordNet have been developed including: word sense disambiguation, text categorization, text clustering, query expansion, machine translation, and many others.

Unfortunately, this potential has never been utilized for the Macedonian language and other than traditional lexical resources, such as dictionaries and lexicons, we are not aware of any current large lexical resources such as WordNet ontology for the Macedonian language.

Although, the manual construction of such lexical resource is most accurate, as far as linguistic soundness is concerned, it requires a lot of time and resources. Therefore, we have developed a method for automated construction of WordNets by using the PWN as a backbone for the construction and Google's translation tool and search engine.

The method is based on the assumption that the conceptual space modeled by PWN is not depended on the language in which it is expressed. Furthermore, we assume that the majority of the concepts exist in both languages, the source and target language, but only have different notations. Given that the conceptual space is already represented in English by the PWN, our goal is to find the corresponding concept notations in the target language by finding the proper translations of the synset members. However, we are aware of the fact that the WordNet produced by our method will be strongly influenced by the effectiveness in which PWN conceptualizes the world. Moreover, we are aware that PWN is not a perfect lexical resource and that all of its mistakes and drawbacks will also be inherited in the WordNet that is produced. Even if a lot of the parts of the produced WordNet remain in English, we believe that it will still be valuable for many WordNet applications in the target language, as a result of the WordNet structure.

The reminder of this paper is organized as follows: in the next section we provide a short overview of a related work after which we will describe our approach and methodology for the construction of the Macedonian WordNet. In sections 3 and 4, we present the results and

[1] http://www.globalwordnet.org/gwa/wordnet_table.htm

explain the usage of the WordNet in practical applications. Lastly in section 5, we discuss the pros and cons of our approach and ideas for future work.

2. Related Work

Due to the time consuming and labour intensive process of manual construction of WordNet, many automated and semi-automated construction methods have been proposed. This section provides a short overview of the methods for automated construction found in literature and considered most interesting.

An attempt to build a Macedonian WordNet was previously made by Aleksandar Pechkov in a scope of coursework. However, none of the deliverables from this study are publicly available.

Fišer D. and Sagot B. (2008) used a multilingual parallel corpus to construct Slovene (SloWNet) and French (WOLF) WordNets. They have PoS tagged, lemmatized, sentence, and word aligned the corpus in order to produce five multilingual lexicons which included French and four multilingual lexicons which included Slovene. Apart from Slovene and French, WordNets for the other languages (Romanian, Czech, and Bulgarian) have already been built and linked to PWN as part of the BalkaNet project (Tufis, 2000). Next, each of the lexicon entries produced is assigned a synset id from the WordNet of the corresponding language. Finally, the intersection of the synset ids of the entries is computed and assigned as a synset id to the Slovene and French words in the lexicon entry.

Changki L. and JungYun S. (2000), for the purpose of construction of Korean WordNet, define the problem of WordNet construction quite differently than the other methods discussed in this section. Namely, each Korean word is mapped to a list of English translations, each of which is expanded with the PWN synsets in which it belongs. Thus, the problem of WordNet construction is defined as finding the adequate English synset for a given Korean word. The authors propose six heuristics: maximum similarity, prior probability, sense ordering, IS-A relation, word match, and co-occurrence. Most interesting was found the word match heuristic which assigns a score to a given candidate synset according to the portion of overlapping words in the English dictionary definition of the Korean word and the English synset gloss and usage examples. Finally, in order to make a final decision, the heuristics are combined by using decision tree learning, where manually mapped senses are used as training data.

Barbu E. and Mititelu B. (2007) developed four other heuristics for automated construction of the Romanian WordNet. Namely, the intersection, WordNet Domains, IS-A relation, and dictionary definitions heuristics were proposed. The last two were found very similar to the IS-A relation and word match heuristics mentioned in the previous paragraph. More attention was paid to the intersection and WordNet Domains heuristics. The second makes use of the WordNet Domains project, which linked the PWN synsets with a set of 200 domain labels from the Dewey Decimal library classification. By using a collection of domain classified documents, all Romanian words in the EN-RO dictionary are labeled with the same domain labels as in WordNet Domains. Thus, when translating a source synset only, the translation candidates which match the synset domain are considered. These experiments proved to be very interesting since they were evaluated against the manually constructed Romanian WordNet and a formal measure of their performance was given.

3. Methodology

3.1. The Approach

Given the assumptions mentioned in the introductory section, the problem of automated construction of the Macedonian WordNet can be formulated as follows: Given a synset from PWN, the method should find a set of Macedonian words which lexicalize the concept captured by the synset.

The first step is by using an English – Macedonian (EN-MK) machine readable dictionary (MRD) to find the translations of all words contained in the synset. These translations are called *candidate words*. Since not all English words have Macedonian translations or are not contained in the MRD, for quality assurance it is assumed that if more than 65% of the words contained in the synset can be translated, then the concept captured by the synset can be expressed with a subset of the candidate words. Thus, the performance of the method is strongly influenced by the size and quality of the MRD used. For this reason, we have spent a lot of time and effort building a large and accurate in-house-developed MRD (Saveski, 2010). The MRD contains 181,987 entries i.e. 61,118 English and 79,956 Macedonian unique terms, where each English word is mapped into a set of Macedonian translations grouped by part of speech. The synsets which did not contain enough known words were skipped and retained in English.

However, not all of the candidate words reflect the concept represented by the synset. Therefore, a subset of words must be selected.

Let that the original synset contain ***n*** English words:

$w_{1,} \ldots, w_{i,} \ldots, w_n$,

and the word w_i has ***m*** translations,

$cw_{1,} \ldots, cw_m$, in the MRD.

Since the MRD has no means of differentiating between word senses, the set of translations of w_i ($cw_1 \ldots cw_m$) will contain the translations of all senses of the word w_i. It is a task of the method to determine which of these words, if any, correspond to the concept captured by the synset. Stated in this way, the problem of translating the WordNet synsets is essentially a *word sense disambiguation* (WSD) problem.

This is not very encouraging because WSD is still an open problem, but nevertheless gives us some pointers which may help in determining the best candidate words. Throughout the history of Artificial Intelligence, many approaches and algorithms have been proposed to solve the problem of WSD. Dagan I. and Itai A. (1994) stated that by using the word sense dictionary definition and a large textual corpus, the sense in which the word occurs can be determined. In other words, the words in the dictionary definition of the word sense tend to occur in the corpus more often, closely to the word in question, when the word is actually in the sense defined, and less often when the word represents other senses.

In terms of the problem of WordNet construction, this means that if the synset gloss can be translated, it will give us a good approximation of which of the candidate words are most relevant for the synset in question. Since manual translation of the glosses is not possible (translating the glosses is equivalent to translating the PWN), the English-to-Macedonian machine translation tool available through Google on the Web was chosen to be used. Although the Google EN-MK translation tool was not extremely accurate at the time of conducting this study, its performance was good enough to capture the meaning of the gloss. From the observations, it was concluded that the most common mistakes made by the translation tool were inappropriate selection of the genre and case of the words. However, this does not affect the use of the gloss translation as an approximation of the correlation between the candidate words and the synset.

The next crucial element for applying the statistical WSD technique is a large Macedonian textual corpus. Although, we are aware of some small textual corpora, mostly newspaper archives available on the Web, any attempt of collecting a large, domain independent corpus is not known to exist. Using a small and domain dependent corpus may significantly affect the performance of the method. On the other hand, collecting a large textual corpus from scratch requires a lot of time and resources, which were not available for this study. Therefore, an alternative method for measuring the correlation between the translated gloss and the candidate words was considered.

Namely, the *Google Similarity Distance* (GSD) proposed in (Cilibrasi & Vitanyi, 2007), calculates the correlation between two words/phrases based on the Google result counts returned when using the word, phrase, and both as a query. Most importantly, the result of applying the GSD is a similarity score between 0 and 1 representing the semantic relatedness of the candidate word and the translated synset gloss. The GSD is calculated for each candidate word and then the words are sorted according to their similarity.

Next, the candidate words are selected based on the following two criteria:

1. the words must have GSD score greater than: 0.2,
2. the words must have GSD score greater than: 0.8 * the maximum GSD score among the candidates.

The first criterion ensures that the words exceed minimum correlation with the gloss translation while the second makes discrimination between the words which lexicalize the concept captured by the synset and those that do not. The coefficients in both criteria were determined experimentally.

Finally, the words selected are included in the resulting Macedonian synset while the other candidate words are considered as not lexicalizing the concept captured by the synset. Figure 1 depicts the method explained in this section.

3.2. Google Similarity Distance

Google Similarity Distance (GSD) is a *word/phrase semantic similarity distance* metric developed by Rudi Cilibrasi and Paul Vitanyi proposed in (Cilibrasi & Vitanyi, 2007). The measure is based on the fact that words and phrases acquire meaning from the way they are used in the society and from their relative semantics to other words and phrases. The World Wide Web is the largest database of human knowledge and contains context information entered by millions of independent users.

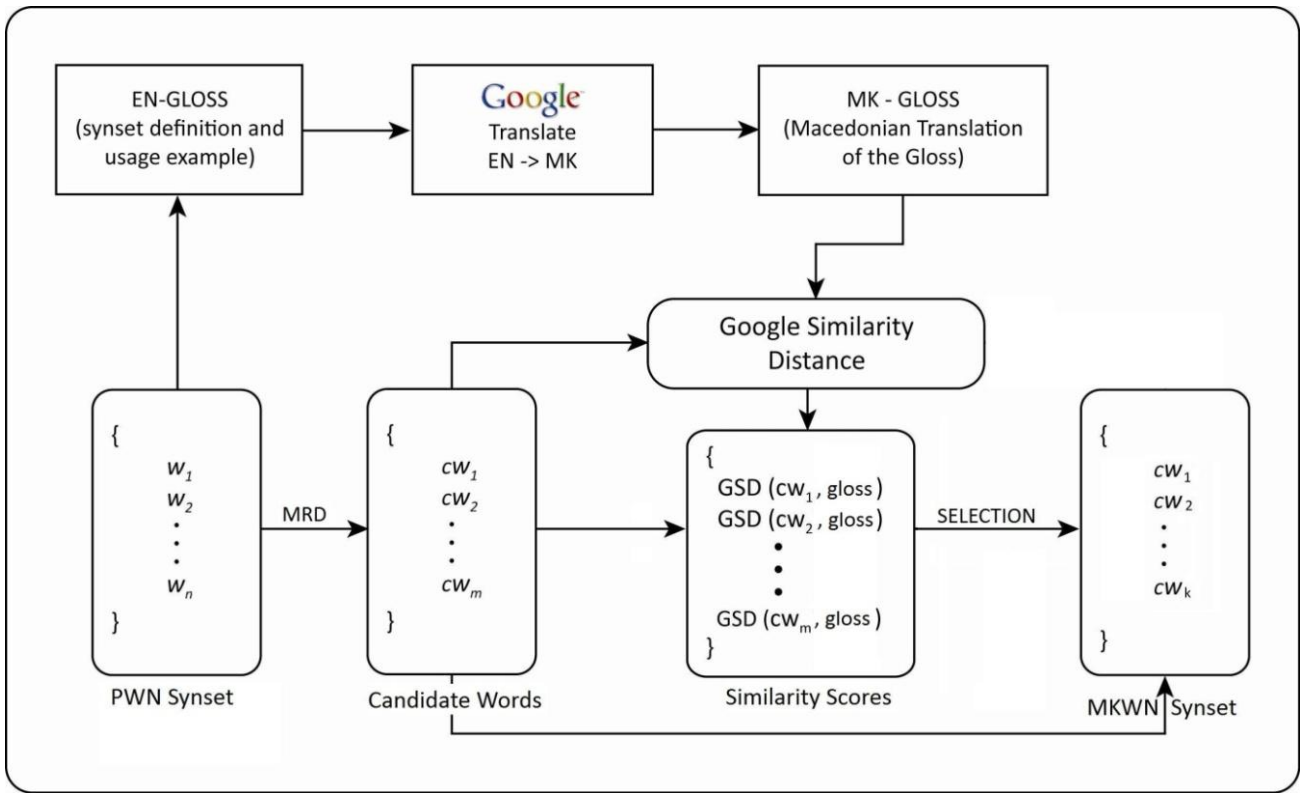


Figure 1. The Google Similarity Distance Method,
(*n*: dimension of the PWN synset, *m*: the number of candidate words, *k*: dimension of the resulting synset)

The authors claim that by using a search engine, such as Google, to search this knowledge, the semantic similarity of words and phrases can be automatically extracted. Moreover, they claim that the result counts of the words in question estimate the current use of the words in the society. As defined in (Cilibrasi & Vitanyi, 2007), the normalized Google Similarity Distance between words/phrases *x* and *y* is calculated as:

$$GSD(x,y) = \frac{\max\{\log f(x),\ \log f(y)\} - \log f(x,y)}{\log N - \min\{\log f(x),\ \log f(y)\}}$$

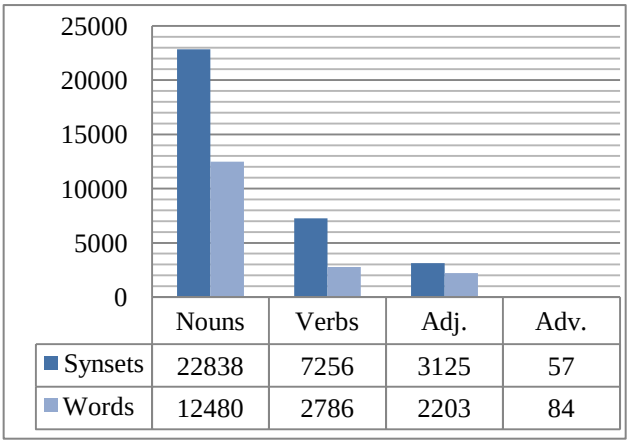
where *f(x)* and *f(y)* denote the result counts returned for *x* and *y*, respectively, and *f(x, y)* denotes the result count when both *x* and *y* are included in the query. The normalization factor *N*, can be chosen but has to be greater than the maximum result count returned. In our case, *f(x)* is the result count returned when the candidate word is included in the query, *f(y)* is the result count of the gloss translation, and *f(x, y)* is the result count when both are included in the query.

Here, the similarity distance is defined by using Google as a search engine, but is applicable with any search engine which returns aggregated result counts. The authors observed that the distance between words and phrases measured in different periods of time is almost the same. This shows that the measure is not influenced by the growth of the index of the search engine and therefore it is stable and scale invariant.

One possible drawback of the method is that it relies on the accuracy of the result counts returned. The Google index changes rapidly over time and the result counts returned are only estimated. However, linguists judge that the accuracy of the Google result counts is trustworthy enough. In (Keller & Lapata, 2003) it is shown that web searches for rare two-word phrases correlated well with the frequency found in the traditional corpora, as well as with human judgment of whether those phrases were natural.

3.3. Comparison with the Intersection Heuristic

In order to evaluate the results of our method, we also applied the Intersection heuristic proposed in (Barbu & Mititelu, 2007), and we have compared the results produced by both methods. The results of applying this heuristic, when compared with the manually produced WordNet, showed most successful results during the experiments for automated construction of the Romanian WordNet. As reported by the authors, on a selected subset of synsets an error rate of only 2% has been achieved.



	Nouns	Verbs	Adj.	Adv.
Synsets	22838	7256	3125	57
Words	12480	2786	2203	84

Figure 2. The size of the produced WordNet

After applying this heuristic for the construction of the Macedonian WordNet, we found out that **45%** of the synsets produced by the GSD method contained exactly the same words. However, because the two methods rely on different rules for translating the synsets, each succeeds to translate different subsets of PWN.

The last step of the construction of the MWN was to combine the synsets produced by both methods in order to produce a single WordNet. Namely, the synsets which could be translated by using both methods but did not result with the same words were produced by using the following rules. If the synset could be translated by using only the monosemous-word rule of the Intersection heuristic (Barbu & Mititelu, 2007), then the synset is produced by applying the GSD method. On the other hand, if the intersection rule of the Intersection heuristic is applicable, then the synset is produced by applying that rule. The rules are based on the fact that the GSD method and the intersection rule of the Intersection heuristic are more restrictive than the monosemous word translation rule.

Figure 2 shows the number of words and synsets produced by combining both methods, grouped by part of speech. It is important to note that all words included in the WordNet are lemmas.

4. Results and Evaluation

4.1. Using the MWN for Text Classification

The most common practices for evaluation of the quality of the automatically built WordNets are manual verification of the synsets produced (e.g. (Changki & JungYun, 2000)) or their comparison with the synsets of the manually developed WordNets, if such exist (e.g. (Barbu & Mititelu, 2007)). Although the manual verification of the synsets developed during the automatic construction of the Macedonian WordNet would be the most accurate and objective evaluation, it would require a lot of time and resources and thus is not an option. Also, as previously mentioned, a manually developed WordNet for the Macedonian language is not available and thus there is no golden standard against which we can evaluate the WordNet produced.

However, it is important to note that the initial objective of this study was not to develop a WordNet which will be a perfect lexical resource, but rather to develop a resource which will give us the opportunity to include semantics in the already developed techniques for Machine Learning (ML) and Natural Language Processing (NLP). Therefore, it was considered that it is much more suitable to evaluate the WordNet developed by its performance in a particular NLP/ML application and by the possible improvements that its usage may allow.

Namely, we were interested in how the use of the Macedonian WordNet will influence the performance of the text classification algorithms. This is only one of the plethora of applications of WordNet. However, it was considered mainly because the performance of the classification algorithms can be measured unambiguously and compared easily.

4.2. The Experiment

The first step towards defining a method for measuring the semantic similarity between two text documents using WordNet is to define how the distance between two WordNet synsets can be measured. We have adopted the *Leacock-Chodorow (LCH)* (1998) and *Wu and Palmer (WUP)* (1994) conceptual distance measures. The LCH measure defines the distance of the concepts (synsets) in terms of the number of nodes between the two synsets in the hierarchy while the WUP measure is based on the number of arcs between the synsets. For more information and comparison of the measures the interested reader can consult (Budanitsky, 1999) and (Budanitsky & Hirst, 2001). Next, since one word can be found in many synsets, we have extended the synset distance measures to word-to-word level. Namely, the distance between two words is defined as the minimum distance (maximum similarity) between the synsets where the first word was found and the synsets where the second word was found. Finally, by using the method defined in (Mihalcea et al., 2006), we extended the semantic word-to-word similarity measure to text-to-text semantic similarity. This measure combines the metrics of word-to-word similarity and *word specificity* (inverse document frequency - *idf*) into a single measure which can be used as an indicator of the semantic similarity of two texts.

During the experiment we compared the performance when using the following three similarity measures:

1. Semantic text similarity based on LCH synset similarity,
2. Semantic text similarity based on WUP synset similarity,
3. Cosine Similarity.

The Cosine Similarity is a classical approach for comparing text documents where the similarity between two documents is defined as the cosine of the angle between the two document vectors. This measure is used as a base line for comparison of the performance of the other two metrics.

In addition, we made use of the KNN - K Nearest Neighbors classification algorithm as a method which is easy to implement and allows the similarity measures to be compared unambiguously. To speed up the classification and improve the performance of the algorithm, during the training phase, we structured the data samples in an *inverted index* (Manning et al., 2008).

For the purpose of the experiment a corpus of Macedonian news articles was used. The articles are taken from the archive of the A1 Television Website published between January 2005 and May 2008. As table 1 shows, the corpus contains 9,637 articles i.e. 1,289,196 tokens classified in 6 categories.

Category	Articles	Tokens
Balkan	1,264	159,956
Economy	1,053	160,579
Macedonia	3,323	585,368
Sci/Tech	920	17,775
World	1,845	222,560
Sport	1,232	142,958
TOTAL	**9,637**	**1,289,196**

Table 1. A1 Corpus, size and categories

4.3. Results

Figure 3 compares the performance of the three similarity metrics by their F-Measure score. As seen in the figure, the LCH-semantic similarity fails to improve the performance of the Cosine Similarity metric. The main reason for the low performance of this measure is due to its inability to calculate the similarity between words with different part of speech. The WUP-semantic similarity metric, on the other hand, has improved classification performance and outperforms both the Cosine Similarity and LCH-semantic similarity metrics by ***6.7%*** and ***20.6%***, respectively. When compared to the Cosine Similarity, as a baseline, this metric manages to find more patterns in the text documents. This is especially evident in the documents from the Sci/Tech and Economy categories.

Although by doing this experiment we cannot argue about the validity of the WordNet produced, we can conclude that the information encoded by the WordNet is meaningful and accurately models the real world. Moreover, we have practically shown that the Macedonian WordNet can be used to include semantics in the existing ML and NLP algorithms and to improve their performance.

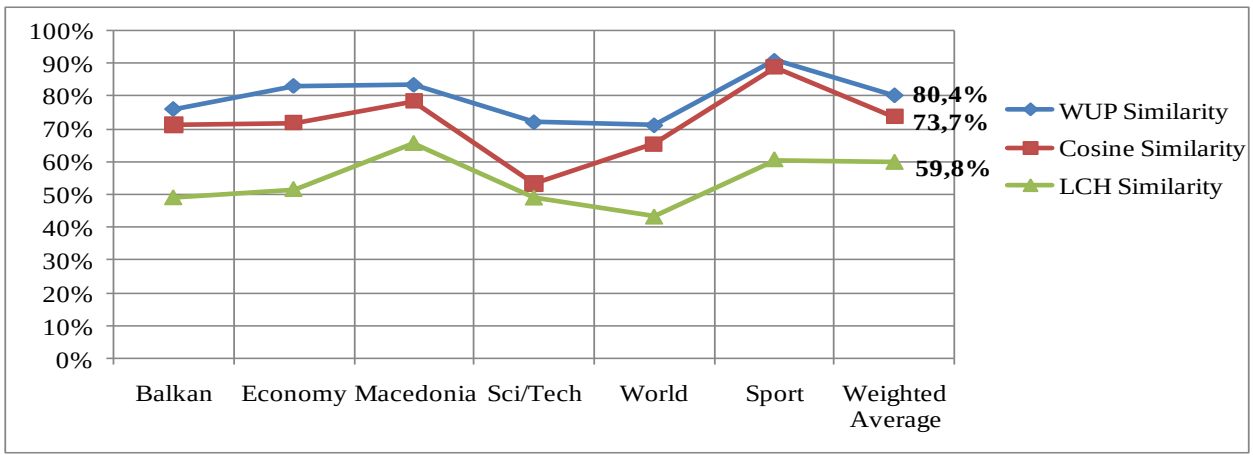


Figure 3. Comparison of all text similarity measures (F-measure)

5. Conclusion and Future Work

In this paper we have proposed a new method for automated construction of WordNets. The method relies on a bilingual dictionary and PWN, as a backbone for the construction and uses Google's machine translation and result counts to make a selection between candidate words. The method presented has been successfully applied for the construction of the Macedonian WordNet but can also be applied to other languages if machine readable dictionary and translation system are available. We have experimentally evaluated the accuracy of the produced WordNet. By using it as a mean to include semantics in the text classification algorithms, we have managed to improve the performance achieved by the standard techniques.

However, our method currently considers each candidate word independently, not taking into account the semantic relatedness which exists between some of the candidate words. In the future, we plan to investigate how the candidate words can be clustered (grouped) prior to assigning them to the synset. We want to consider how, based on the individual similarity between the words and the similarity of each word and the gloss, it can be determined which group is most suitable to express the concept captured by the synset. In this way, we can compensate for some of the possible mistakes made during both the translation of the gloss and the measuring the semantic similarity between the candidate word and the gloss. Moreover, the probability of incorrectly assigning a group of words to a synset is much lower than the probability of incorrectly assigning an individual word.

Next, we would like to repeat the text classification experiment by using larger corpus of text documents and to investigate whether this improvement in the performance will also be evident. Moreover, we are interested in how the use of more complex word-in-context-to-word-in-context similarity measure will influence the performance.

Finally, we plan to conduct similar experiments for other WordNet applications, such as text clustering and word sense disambiguation, and to apply this method for construction of other non-English WordNet.

6. References

Barbu, E. & Mititelu, B. V. (2007). Automatic Building of Wordnets. In Proceedings of *Recent Advances in Natural Language Processing IV*, John Benjamins, pp. 217--226, Amsterdam, 2007.

Budanitsky, A. & Hirst, G. (2001). Semantic distance in WordNet: An experimental, application oriented evaluation of five measures. In Proceedings of the *NAACL Workshop on WordNet and Other Lexical Resources*.

Budanitsky, A. (1999). *Lexical Semantic Relatedness and its Application in Natural Language Processing* [Online]. Accessed from: http://www.cs.toronto.edu/

Changki, L. & JungYun, S. (2000). Automatic WordNet mapping using word sense disambiguation. In Proceedings of the *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cilibrasi, R. & Vitanyi, M. B. (2007). The Google Similarity Distance. In the proceedings of *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, pp. 370--383.

Dagan, I. & Itai, A. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus. In the proceedings of *Computational Linguistics 1994*, vol. 20, pp. 563--596.

Fellbaum, C. Et al. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT press.

Fišer, D. & Sagot, B. (2008). Combining Multiple Resources to Build Reliable Wordnets. In Proceedings of *Text, Speech and Dialogue (LNCS 2546)*, Springer 2008, pp. 61--68, Berlin: Heidelberg.

Keller, F. & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. In the proceedings of *Computational Linguistics 2003*, vol. 29:3, pp. 459--484.

Leacock, C. & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In Fellbaum C., *WordNet: An electronic lexical database*, pp. 265--283.

Manning, C. D., Raghavan, P. & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Mihalcea, R., Courtney, C. & Strapparava, C. (2006). Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In Proceedings of *American Association for Artificial Intelligence*.

Saveski, M. (2010). *Development of a WordNet Prototype for the Macedonian Language*. Bachelor (Hons) Thesis, Staffordshire University, UK.

Tufis, D. (2000). BalkaNet: design and development of multilingual Balkan wordnet. In Proceedings of the *Romanian Journal of Information Science and Technology*, vol. 7(1-2).

Wu, Z. & Palmer, M. (1994). Verb semantics and lexical selection. In Proceedings of the *Annual Meeting of the Association for Computational Linguistics*.

Obtaining Information Beyond Speech Technologies for a User-Adaptive Multimodal Dialogue System

Gonzalo Espejo *, **Nieves Ábalos** *, **Gregor Podrekar** † **and Ramón López-Cózar** *

* Dep. of Languages and Computer Systems,
CITIC-UGR. University of Granada
{gonzaep, nayade}@correo.ugr.es; rlopezc@ugr.es

† Faculty of Electrical Engineering,
University of Ljubljana
gregapo@gmail.com

Abstract

In this article we consider our plans to improve an already implemented multimodal dialogue system for control of appliances in an ambient intelligent environment, called *Mayordomo*. We think that by means of the planned improvements, the dialogue system will become more user-adaptive and thus the interaction will be more user-friendly. The adaptation will be implemented considering information not taken into account in the original version of the system, specifically user localization and identification. The paper discusses our recent research on methods to achieve this sort of information, as well as our plans to make use of user profiles.

Pridobivanje neverbalnih informacij v večmodalnem sistemu za dialog, sposobnem prilagoditve uporabniku

Prispevek obravnava naše načrte za nadaljnje delo pri izboljšavi delujočega sistema za večmodalni dialog Mayordomo, ki je namenjen krmiljenju naprav v okolju ambientalne inteligence. Predvidevamo, da bodo opisane izboljšave sistem še bolj približale uporabniku. Pri implementaciji izboljšane verzije bomo dodatno upoštevali predvsem informacijo o identiteti ter lokaciji uporabnika, kar v prvotni različici sistema ni bilo upoštevano. V prispevku opisujemo postopke, s katerimi pridobivamo opisane nove podatke, ter načrte za izdelavo uporabniških profilov.

1. Introduction

This multimodal dialogue system employed for the elaboration is *Mayordomo* (Ábalos et al., 2010). The aim of Mayordomo is to centralize control of appliances in a home. The interaction with appliance can be through spontaneous speech or through a traditional GUI interface, based on keyboard and mouse.

Besides handling appliance, the system support different types of users, depending on the administrator level and the experience with *Mayordomo*. The system administrator has privileges to perform special actions, for example, installing and uninstalling appliances Also, restrictions are allowed to some users. For instance, parents can forbid that children watch TV after 10 p.m. The system creates a log of all actions carried out within the environment by any user.

In this article we purpose an upgrade of *Mayordomo* by making it user-adaptive. So, the user can be located and identified by the system in an implicit way, being not necessary saying where the user is, ridding him off about providing unnecessary information. This information about the user can be used to optimize the dialogue with the user.

The paper is organized as follows. First, some methods are presented for automatically locating and recognizing the user such as RFIDs, cameras, and microphone arrays. In section 2 we explain the user-adaptive dialogue system. Finally, section 3 presents the conclusions and outlines possibilities for future work.

2. Methods to locate and recognize the user with *Mayordomo*

Using Mayordomo, the localization and identity of the user can be deduced using different ways to extract information, or even using all these ways together. In this section we describe three of the most common ways: radio frequency identification (RFID), cameras and microphone arrays. Using these devices, the user is not required to say where and who he is as the system can get this information automatically.

2.1. Radio Frequency IDentification (RFID)

RFIDs are considered one of the main ways for physical browsing, which is an interaction paradigm that associates digital information with physical objects (Aghajan et al., 2010). The interaction takes place using mobile terminals, such a mobile phones or cards that have one tag associated with a reader device. Each tag contains a universal resource identifier (URI, or Web address). The readers are distributed throughout the environment. Each RFID reader has a serial or identification number. When one of the tagged objects is scanned by the reader, the tag and the identification number of the reader are sent to the dialogue system. The distance between the physical object and the RFID reader can range from a centimetre to four meters. Usually, in closed and small environments this distance is not very large.

To be useful for our purposes, the dialogue system must know who is the user carrying each tag. Taking into account this information, the system can find out which was the last RFID reader used by the user, and thus deduce his localization.

RFID devices have been used in several research projects concerned with user localization. For example, López-Cózar et al. (2006) used RFID devices to locate students in an educational environment, whereas Haya et al. (2004) employed these devices in a home environment.

2.2. Cameras

In addition to provide information about user localization, cameras can help in detecting whether the user is talking or not, for example, using lip-motion information (Hazen et al., 2003).

Using cameras, the user can be identified by recognizing his face. This process can be carried out in two stages: face detection and face recognition. In the first stage the face is localized in the sequence of images provided by the camera. Many approaches can be used for this task, and many solutions already exist. One of them, which is used by Mayordomo, is based on the open source computer vision library openCV. This face detector implements a version of the face detection technique, first developed by Paul Viola and Michael Jones, which is commonly known as the Viola-Jones detector (Bradski & Kaehler, 2008). This technique was later extended by Rainer Lienhart and Jochen Maydt. The detector uses Haar-like wavelets created by adding and subtracting rectangular image regions, and then thresholding the result.

After the face of the user has been detected and extracted from the sequence of images, we must carry out face recognition. A commonly used method for doing this is by using Principal Component Analysis (PCA) (Albiol et al., 2005). Using this method, the feature vector that represents the face, is calculated from an individual face image and then a decision is made regarding the candidate face, for example, using a number of classifiers. A weakness of this camera approach is that the user's face might not be visible for the camera as it can be hidden by other objects.

2.3. Microphone arrays

A microphone array consists of multiple microphones placed at different spatial locations. Microphone arrays have been often used to solve, or at least reduce, the effects of the so called *cocktail party* problem. These arrays are very apparent for hands-free speech recognition applications, such as *Mayordomo*, in which it is not possible to use a close-talking microphone (Cherry, 1953). For example, Brandstein et al. (1997) used a linear intersection and employed sensor array time-delay estimate information. Also, Sturim et al. (1997) implemented the tracking of multiple talkers using this type of array.

As the input to the microphone array is the user's voice, which is specific for each person, these arrays can be useful as well to identify the speaker. The process of doing this is called *automatic speaker recognition* (Reynolds & Rose, 1995). As with any other automatic recognition process, speaker recognition consists of two stages: feature extraction and classification. The aim of feature extraction is to remove unnecessary information from the sensor data, and convert the properties of the signal which are important for pattern recognition to a format that simplifies the distinction of the classes. Mel Frequency Cepstral Coefficients (MFCCs) are normally used as features, whereas Gaussian Mixture Models (GMMs) is the most common method for classification.

Speaker recognition can be *text dependent* or *text independent*. The difference is that in the first type the meaning of the text is employed for the recognition task, whereas it is not used in the second type. The second type is more convenient for our system and thus the one that we will use.

High level features can also be used for text independent speaker recognition, such as word idiolect, pronunciation, phone usage and prosody. However, as we have observed that with *Mayordomo* the dialogues are very short, we think these features are not very appropriate.

2.4. Combining different approaches

Face recognition and automatic speaker recognition share a potential drawback, namely, they can be unreliable due to distortions. In the case of speaker recognition, distortion is normally caused by background noise, whereas in the case of face recognition it is caused by changes in lighting and in the user's appearance. In order to make the process of user identification more reliable, combining both methods is possible and has already been investigated (Albiol et al., 2005). In order to get the best performance possible, we plan to make *Mayordomo* combine the methods discussed above, and decide about the user's localization and identity on the basis of the combined data.

3. Information management for user-adaptive dialogue

In this section we describe how we plan to use the information about user identification in *Mayordomo* in order to enable a more user-friendly interaction.

3.1. Input information about user identification

Once we know who is the current user, we can adapt the system's performance. To do this, the first step will be designing a user model. Taking into account (Webb et al., 2001), the four main issues to describe are:

(1) the cognitive processes that underlie the user's actions (*behaviour*);

(2) the differences between the user's skills and expert's skills (degree of *expertise*);

(3) the user's behavioural patterns or *preferences*; and

(4) the user's *characteristics*.

In our case, the user model will be based on a number of parameters used to adapt the dialogue system, for example age, previous experience using the system, user preferences, interaction language, etc. These parameters will be stored in a *user profile*.

The user profile for a new user can be built either *manually*, in which case a system developer just fills in all the information that he knows about each user, or *automatically*, in which case the system prompts the user for his/her personal data and for general information (e.g.,

name, age, language, gender, etc.). This second approach was used by Lucas et al. (2009) to develop user profiles for a dialogue system that controls a Hi-Fi device.

The management of all the information stored in the user profiles will be carried out by a *Profile Manager*, which will interact with some modules in the dialogue system, for example, the Dialogue Manager.

Table 1 shows an excerpt of a user profile to be created for our system. It represents a profile filled in by a bilingual woman called Mary, who is twenty-five years old. She prefers speech-based system's interaction in English, although she wants to read the text in the graphical interface (user interface) in Spanish. She is a system administrator and thus is allowed to use all the appliances in the house (in this example, lights and TV). Besides, when Mary is at home, she normally watches channel number six on TV and likes a lightning ambient created by all the lights in the living room at medium brightness. Finally, she prefers speech for the input interaction to the system and the graphical interface for the output (the system does not interact with her using speech).

General user information	
Name	Mary
Genre	Female
Age	25
Language preferences (used by the system)	
ASR	English
UI	Spanish
TTS	English
Characteristics	
Level	Administrator (3)
Privileges	Light – Allowed
	TV – Allowed
User preferences	
Light	Medium brightness
TV	Channel 6
Interaction user-system	
Input	Speech
Output	Graphical interface

Table 1**.** *Example of user profile*

Our system's implementation follows the definition of *sessions* and *dialogue turns* presented in Pargellis et al. (2004). *Mayordomo* will control the interaction with the user by means of a set of dialogues which start with an activation command and finish with a close command. A dialogue is a sequence of turns, i.e. interactions in which either the user or the system say something.

In the following sections, we explain some parameters of our user profiles.

3.2. Age

Generally, the dialogue systems which adapt their performance to older or younger users create ranges of age for the users. For example, in Georgila et al. (2008) older participants were aged between 50 and 85, whereas younger participants were aged 20 to 30.

Wolters et al.' (2009) group studied what would happen if the relevant user groups were not delineated by age but by characteristic patterns of behaviour. They analysed whether users can be grouped according to the way in which they interact with the system. The statistical analysis methodology that they used allows identifying "extreme" users and "typical" users, without using age in the initial analysis. They quantified the *interaction style* (the linguistic choices that users make) of each user based on a linguistic analysis of their dialogues. Examples of relevant linguistic choices are choosing between different expressions of agreement (e.g., "yes" against "that's fine"), or using politeness expressions like "please". Finally, they have shown that being old does not necessarily mean acting old. Even though older users were more likely to have a "social" interaction style than younger users, a sizeable proportion of older users preferred the same "factual" interaction style as younger users.

For the implementation of the *Mayordomo* system, we have chosen the common approach of creating ranges of age for users, as done by Georgila et al. (2008). Our analysis is made in two levels: lexical and concerned with dialogue acts. Similarly as Wolters et al. (2009), Georgila et al. (2008) argued that older users produce longer dialogues than younger users. They also have a richer vocabulary and use a larger variety of speech acts. Besides, younger users tend to restrict themselves to speech acts that are of immediate relevance to the task. A third of all words uttered by younger users are "yes" or "no", despite 13% for older users. In order to express approval or disapproval, older users are more likely to use expressions other than "yes", such as "fine". Also, they are more likely to use expressions that are more appropriate in human/human interactions, such as forms of "goodbye" or "thank you". Moreover, Georgila et al. (2008) compared their study with the word-level analyses of the MeMo corpus of Gödde et al. (2008), and concluded that the social interaction words that distinguish between older and younger users appear to be task-specific.

Table 2 shows an example of interaction between two users and *Mayordomo*. As said above, Mary is a 25 years old woman, whereas Damien is a 65 years old man. The example shows that *Mayordomo* adapts the dialogue depending on the age of the user, creating more complex and polite dialogues in the second case.

Younger user (Mary)	*Older user (Damien)*
S: Hi Mary!	S: Welcome Damien.
U: Hi Mayordomo. Please switch on the lights.	U: Good morning. Please switch on the lights.
S: Where?	S: Excuse me, where do you want me to switch on the lights?
U: In the living room.	U: In the living room, please.
S: Done.	S: The lights in the living room are now turned on.
	U: Thank you.

Table 2. *Example of dialogue adaptation by age*

3.3. Preferences

Lucas et al. (2009) proposed to implement systems' adaptation taking into account a number of *usage statistics*, which are divided in different groups according to domain-knowledge. This enables the system to infer the preferences for each user among the different statistic groups, and among the different items belonging to the same group. To sum up, the statistic counts are used to build a model for user preferences, which allows the system to suggest hypotheses in incomplete information situations. When in a system-user interaction the dialogue manager finds that there is lack of information, instead of directly asking the user, the system must look first into the user profile in order to decide whether the user has a clear preference for a certain action.

Table 3 shows two sample expected interactions with *Mayordomo*. The first one will consider stored usage statistics, whereas the second will not. *Mayordomo* will adapt the dialogue depending on the stored usage information, asking for more information in the second case.

Using stored information	Not using stored information
S: Hi Mary.	S: Hi Mary.
U: Switch on the TV	U: Switch on the TV
S: Done.	S: Done.
Note: the system statistics about the user show that Mary prefers channel six and loud TV volume.	U: Please select channel six.
	S: Done.
	U: Turn up the volume.
	S: Done.

Table 3**.** *Example of dialogue adaptation by preferences*

3.4. Characteristics (privileges)

Another important issue in our system is in terms of usage permissions (*user privileges*). *Mayordomo* stores information related to user-appliance interaction with a tag regarding user permission for each appliance. If the interaction is not allowed, *Mayordomo* informs the user, as can be observed in **Table 4.**

3.5. Management (degree of expertise)

Mayordomo enables different kinds of user depending on the number of times they have used the system. For example, **Table 4** shows that Mary is a novice user because it is the first time that she interacts with the system. Instead, Peter is an expert user because he has used it many times and knows perfectly how to use it. By modifying the behaviour of the Dialogue Manager and the output of the Natural Response Generator module, the system avoids giving basic information to expert users in order to avoid boring them with already known information.

4. Conclusions and future work

In this paper we have described some methods to obtain extra information from the user of *Mayordomo*, a multimodal dialogue system used in home environments.

This information goes beyond the strict orders or queries that can be done through dialogue. The extra-information can be divided into two groups: location and identity.

The location refers the exactly point or room where the user is. This location can be deduced using RFIDs, cameras or microphone arrays; or even, by combining of all.

The second group, identity information, deals with the different characteristics of users. The goal is to provide better adapted information depending on some characteristics, such as age, preferences, privileges and degree of expertise.

As future work we plan one evaluation of the different methods to obtaining information, as well as a comparative between all of them, with particular attention in how the information is received and adapted. Also, users will answer some questionnaires with the aim of compare the adaptation results with the true preferences and likes.

Novice / basic user	*Expert user*
S: Hi Rose.	S: Hi Peter
U: Which appliances are there in the hall?	U: Switch on the light.
S: There are two lights (ceiling light and lamp) and piped music.	S: Which light?
U: Please switch on the light.	U: The ceiling light.
S: Sorry, which light? The ceiling light or the lamp?	S: You are not allowed to do this.
U: The ceiling light.	
S: You are not allowed to use the ceiling light. Please contact the system administrator to get assistance.	

Table 4. *Examples of dialogue adaptation by degree of expertise*

5. Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology, under project TIN2007-64718 Adaptive Hypermedia for Attention to Different User Types in Ambient Intelligence Environments.

6. References

Ábalos, N., Espejo, G., López-Cózar, R., Callejas, Z. (2010). Sistema de Diálogo Multimodal para una Aplicación de Inteligencia Ambiental en una Vivienda. Procesamiento del Lenguaje Natural, 44 (pp: 51—58)

Aghajan, H., López-Cózar, R, Augusto, J. C. (2010). Human-centric Interfaces for Ambient Intelligence. Academic Press.

Albiol, A., Torres, L., Delp, E. J. (2005). A fully automatic face recognition system using a combined audio-visual approach. Vision, Image and Signal Processing, IEE Proceedings, vol. 152, (pp: 318—326).

Bradski, G., Kaehler, A. (2008). Learning OpenCV. O'Reilly Media.

Brandstein, M., Adcock, J., Silverman, H. (1997). A closed-form location estimator for use with room environment microphone arrays. Speech and Audio Processing IEEE Transactions, 5 (pp. 45–50)

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and two ears. Journal of the Acoustic Society of America, 25 (pp. 975—979)

Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., Moore, J., Watson, M. (2008). A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems. In Proceedings of the 6th International Conference on Language Resources and Evaluation.

Gödde, F., Möller, S, Engelbrecht, K.-P., Kühnel, C., Schleicher, R., Naumann, A., Wolters, M. (2008). Study of a speech-based smart home system with older users. In International Workshop on Intelligent User Interfaces for Ambient Assisted Living (pp: 17–22).

Haya, P. A., Montoro, G., Alamán, X. (2004). A prototype of a context-based architecture for intelligent home environments. In Proceedings of the International Conference on Cooperative Information Systems (pp. 477—491)

Hazen, T., Weinstein, E., Kabir, R., Park, A., Heisele, B. (2003). Multi-modal face and speaker identification on a handheld device. In Proceedings of the Workshop Multimodal User Authentication (pp. 120—132)

López-Cózar, R., Callejas, Z., Montoro, G., Haya, P. (2006). DS-UCAT: Sistema de Diálogo Multimodal y Multilingüe Para un Entorno Educativo. In Proceedings of the IV Jornadas en Tecnología del Habla (pp. 135—140)

Lucas, J.M., Fernández, F., Salazar, J., Ferreiros, J., San Segundo, R. (2009). Managing Speaker Identity and User Profiles in a Spoken Dialogue System. Procesamiento del Lenguaje Natural, 43 (pp: 77—84)

Pargellis, A.N., Kuo, H.K.J., Lee, C.H. (2004). An Automatic Dialogue Generation Platform for Personalized Dialogue Applications. Speech Communication, 42: (pp: 329—351).

Reynolds, D. A., Rose, R. C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transaction on Speech and Audio Processing, vol. 3, No. 1, (pp. 72)

Sturim, D., Brandstein, M., Silverman, M. (1997). Tracking multiple talkers using microphone-array measurements. In Acoustics, Speech, and Signal Processing, IEEE International Conference, volume 1, (pp. 371–374)

Webb, G., Pazzani, M. and Billsus, D. (2001). Machine Learning for User Modeling. User Modeling and User-Adapted Interaction, 11: (pp: 19--29)

Wolters, M., Georgila, K., Moore, J., MacPherson, S.(2009). Being Old Doesn't Mean Acting Old: How Older Users Interact with Spoken Dialog Systems. ACM Transactions on Accessible Computing, Vol. 2, No. 1, Article 2.

Formant Frequencies In Children With Normal Hearing And Profound Or Severe Hearing Impairments

Martina Ozbič*, Damjana Kogovšek[#], Daniil Umanski[†]

*Faculty of education, University of Ljubljana,
Kardeljeva ploščad 16, 1000 Ljubljana, Slovenija
Martina.ozbic@pef.uni-lj.si
[#] Faculty of education, University of Ljubljana,
Kardeljeva ploščad 16, 1000 Ljubljana, Slovenija
Damjana.kogovsek@gmail.com
[†] Leiden University Centre for Linguistics, Leiden Institute for Brain and Cognition (LIBC), Leiden University
Albinusdreef 2, locatiecode C3Q-46, 2333 ZA, Leiden
daniil.umanski@gmail.com

Abstract

The purpose of the present study was to discover the differences in vowel formant production (F1 and F2) in 33 children, aged 5-9 years, with a different hearing status (11 children with normal hearing (NH), 9 children with prelingual severe (SHI) (mean of hearing loss in dBHL, better ear=68,53, SD=18.90) and by 13 children with prelingual profound hearing impairment (PHI) (mean of hearing loss in dBHL, better ear=106.70, SD=4.82). Formant frequencies associated with 7 Slovenian vowels (/i/, closed /e/, open /e/, /a/, open /o/, closed /o/, /u/), produced during naming pictures or reading words from the Slovenian articulation test were obtained. All first formant and second formant frequencies of high front and back vowels of the speakers with hearing impairment were significantly different from those of the normal-hearing children. The findings suggest the role of the auditory feedback in vowel production in speakers with hearing impairment. The knowledge may be used in speech therapy in visual monitoring of vowel production in HI speakers and in acoustical engineering, to give more stress on high frequencies during acoustic processing.

Povzetek

Namen raziskave je analiza razlik v formantni produkciji (F1 in F2) 33 otrok, starih od 5 do 9 let, ki so polnočutni (11 otrok), naglušni (9 otrok, boljše uho=68,53, SD=18.90) ali gluhi (13 otrok, boljše uho=106.70, SD=4.82). Sedem samoglasnikov slovenskega jezika (polglasnik smo opustili) smo analizirali iz zvočnih posnetkov imenovanja ali branja besed artikulacijskega testa. Vsi prvi in drugi formanti sprednjih zgornjih ter zadnjih samoglasnikov otrok z izgubo sluha se statistično pomembno razlikujejo od formantov polnočutnih otrok. Rezultati opozarjajo na vlogo slušne povratne zanke za samoglasniško produkcijo. Podatke lahko uporabimo pri logopedski terapiji ob vidnem spremljanju samoglasniškega izgovora ter v zvočnih aplikacijah, ter opozarjajo na večje upoštevanje visokih frekvenc pri akustičnem procesiranju.

1. Introduction

Several authors claim that the speech production of individuals with severe prelingual hearing impairment is different from the speech of those with profound hearing impairment and from that of normal-hearing individuals, due to an inefficient auditory feedback (Murphy and Dodds , 2007, 248-250; Waldstein, 1990; Markides, 1983). The articulation manoeuvres for vowels are controlled through audition and through kinaesthesia (Nasir & Ostry, 2008; Nasin & Ostry; 2006, Purcell & Munhall, 2006). The indirect evidence of this statement can be drawn from results of several studies on formant production of deaf speakers (i.e. Waldstein, 1990; Nikolaidis & Sfakkianaki, 2007).

Speakers with hearing impairment show less differentiated vowels and a more centralised vowel space. F1 and F2 formant frequencies show reduced ranges during the production of the different vowel qualities, and there can be an extensive overlap of vowel areas and a tendency toward the neutral schwa (Angelocci, Kopp & Holbrook, 1964; Ryalls, Larouche & Giroux, 1983; Fletcher, 1995). This reduced differentiation of vowels has been attributed to limited auditory feedback and the relative invisibility of articulatory gestures needed for vowel production (Monsen, 1976). Higher frequencies tend to be more affected, as hearing sensitivity is greatly reduced above 1000 Hz for individuals with hearing impairment. As a result, more errors have generally been reported for the high and the middle vowels compared to the low ones and for the front than the back vowels. The high frequency, low intensity F2 formants of the high vowels are more likely to be affected than the lower-frequency, more intense F2 formants of the back vowels (Nicolaidis & Sfakiannaki, 2007). Taking into account the fact that residual hearing most often covers low frequencies better than high ones, this is an indirect prove that vowels are controlled by the auditory feedback most. In Subtelny, Whitehead and Samar's report on the speech production of four deaf women, the formant structure disclosed consistent neutralisation of vowels, with F2 values clustering in the 1500–2100 Hz frequency range, which is attributed to the observed restricted horizontal movements of the tongue within the oral and pharyngeal cavities. If these restrictions affect the production of all vowels, a lower F2 might be assumed for the front vowels, which normally have a high F2. A higher F2 frequency would be anticipated for back vowels, which normally have a low F2 (Subtelny, Whitehead, Samar 1992, 574-579). Schenk, Baumgartner, and Hamzavi (2003) analyzed the frequencies of the first and second formants and the vowel spaces of selected vowels in word-in-context

condition of 23 postlingually deafened and 18 normal-hearing speakers were compared. All first formant frequencies (F1) of the postlingually deafened speakers were significantly different from those of the normal-hearing people. The values of F1 were higher for the vowels /e/ (418±61 Hz compared with 359±52 Hz, *P*=0.006) and /o/ (459±58 compared with 390±45 Hz, *P*=0.0003) and lower for /a/ (765±115 Hz compared with 851±146 Hz, *P*=0.038). The second formant frequency (F2) only showed a significant increase for the vowel /e/ (2016±347 Hz compared with 2279±250 Hz, *P*=0.012). The results of Waldstein (1990) exploring selected properties of consonants, vowels, and suprasegmentals in the speech of seven totally postlingually deafened individuals demonstrates that all deafened subjects showed a reduction in formant frequency ranges and in the acoustic vowel space. The reduced range values suggest a smaller degree of tongue movement in articulation. Waldstein reports an increased variability of formant frequencies for both F1 and F2.

2. Goal of the paper

The purpose of this study was to find out the differences in vowel formant production between children with normal hearing and those children with prelingual profound and severe hearing impairment. The hypotheses were:

H1: The F2 values of anterior vowels are lower and F2 formant values of the posterior vowels are higher in the hearing-impaired groups, according to the degree of hearing impairment, compared with the values of normal-hearing individuals.

H2: The ranges of F1 from high to low vowels and of F2 from anterior to posterior vowels are smaller in the hearing impaired groups, according to the degree of hearing impairment, compared with the values of normal-hearing children.

H3: The formant space in the F2-F1 plane is the smallest in children with profound hearing impairment, followed by children with severe hearing impairment and the greatest formant space in the F2-F1 plane is in normal-hearing children.

3. Materials and methods

Participants: The experimental group[1]: twenty-two Slovenian children from 5 to 9 years old[2] with profound (13) or severe (9) hearing impairment were included in the study (severe hearing impairment: 5 males, 4 females, mean of age=7.6 years SD=1,51; profound hearing impairment: 8 males, 5 females, mean of age=7.7 years SD=1.32). All presented severe and profound sensorineural prelingual deafness in unaided condition (Table 1). All were children from non-inclusive kindergartens and schools for the deaf and hard of hearing in centres for speech and language impairments in Ljubljana, Maribor and Portorož in Slovenia. This means that they were sign language users in the deaf community and were orally trained for communication with those who don't use sign language and for bilingual education (Slovenian sign language and Slovenian oral language). All of them were detected and diagnosed in early childhood during neonatology screening tests or up to the 3 years of age. All of the children were fitted with analogue behind-the-ear hearing aids and did not have any developmental disorders. Three had deaf parents. None were cochlear implant users. All the speakers used the hearing aid continuously (often or always during the day), mainly during the educational process, which was bilingual and bimodal, i.e. Slovenian and/or sign language. The school curriculum requires to use both oral Slovenian and sign language (production and perception).

The control group: Eleven Slovenian normal-hearing children (7 males, 4 females) aged from 5 to 9 (M=7.0 years, SD=1.18) were included in the study as the control group. None of the normal-hearing children had any developmental diseases or disorders.

According to the Welch robust test of equality of means, the experimental and control groups were not statistically different in age (p=0.401); the Chi square test shows that also the differences in the distribution of gender were not statistically relevant (p=0.931).

	N	Mean	**SD**	**SE**	Min	Max
	Mean of hearing loss in dBHL, right ear					
SHI	9	77.91	22.10	7.37	48.27	103.73
PHI	13	109.72	4.57	1.27	104.36	119.09
	Mean of hearing loss in dBHL, left ear					
SHI	9	71.83	17.27	5.76	45.91	91.82
PHI	13	109.32	7.49	2.08	98.64	124.55
	Mean of hearing loss in dBHL, better ear					
SHI	9	68.53	18.90	6.30	45.91	90.00
PHI	13	106.70	4.82	1.33	98.64	114.09
	Mean of hearing loss in dBHL, worse ear					
SHI	9	81.21	18.97	6.32	49.27	103.73
PHI	13	112.29	6.11	1.70	105.14	124.55

Table 1: Descriptive statistics of the mean values of hearing loss for better and worse ear, and for right and left ear (NH=normal hearing, SHI=severe hearing impairment, PHI=profound hearing impairment)

Variables: age, degree of hearing loss (mean of hearing loss in dBHL for right and left ear and for better and worse ear), F1 and F2 formant frequency values for seven vowels (/i/, closed /e/, open /e/, /a/, open /o/, closed /o/, open /o/ and /u/) of Slovenian language (Cronbach's alpha: 0.972). The schwa is omitted due to frequent omission or substitution with closed /e/ or open /e/ in the speech of deaf subjects.

Data acquisition: The Three-position test of articulation for Slovenian (Globačnik, 1999) and an additional list of seven words were used. The test battery of articulation, used by all speech and language therapists in Slovenia, is a set of well-known, frequent words with simple and complex phoneme structures. In the set of the elicited words the most frequent stressed

[1] The research was made in accordance with the Declaration of Helsinki (1983).

[2] The beginning of the mutation can be fixed at the age of 10–11 years, according to Hacki, Heitmüller (1999).

syllables were the initial syllables. The most frequent stressed syllable structure was CV.

Analysis tools: The speech of all participants was recorded on a Sony TCD-D8 DAT recorder with a Sennheiser MD 441 U microphone, which has an even frequency response from 0 to 20 kHz. The recordings were monitored on-line by the investigator by visual inspection of the VU meter on the tape recorder. The recordings were sampled at 16 KHz and digitized using a HP notebook. The data were stored with CoolEdit2000 software for sound data processing and analysed with tools for speech analysis (Praat version 5.1.40 and SpeechAnalyzer SIL version 3.0.1). The fundamental and first and second formants were taken from the most stable portion of the vowel in stressed syllables. Typically this was the midpoint of the vowel. If the centre portion of the vowel did not yield the most stable spectra, measurements were taken slightly earlier or later than the midpoint. Formants were selected by way of convergence among values derived from spectrographic displays and FFT first, assisted by LPC analyses in both software packages. First approximations for formant frequencies were provided by spectrographic display and FFT analysis, with supportive measurements obtained from LPC analysis. We analysed from min 2 (open /e/) to 33 instances per vowel for each speaker. The statistical analysis was performed with WASP 18.0 for Windows.

Statistical analysis: frequency analysis was used to describe the frequencies of the variables, the distribution and to test the hypothesis H2 and H3, Kolmogorov-Smirnov test was used to test the normal distribution (all variables are normal distributed at $p<0.05$); ANOVA, Welch analysis and post-hoc Bonferroni analysis were used to test the hypothesis H1 and H2 and to analyse the statistically relevant differences between formant values (means) in children with normal hearing, and those with severe and profound hearing impairment.

3. Results

The results show some differences in vowel production between the three groups (Table 2, Figure 1). The mean F1 values show a shifted vertical space, with frequency mean range of 620/590 – 991 Hz for the children with severe hearing impairment and a vowel space with frequency range of 639/662 – 1007 Hz in children with profound hearing impairment, compared to that of the children with normal hearing and a frequency range of 532/551 – 879 Hz.

Comparisons of the second formants of the vowels produced by the children with PHI and SHI revealed neutralisation of vowels, with F2 values clustering in the 1104 – 2509 Hz frequency range for severe impairment and in the 1222 – 2494 Hz frequency range for PHI (in subjects with NH, F2 values cluster in the 975 – 2910 Hz frequency range), which is attributed to the restricted horizontal movements of the tongue within the oral and pharyngeal cavities, according to Subtenly, Whitehead, Samar (1992, 574-579) and Engwall (1999).

		N	Mean	SD	Min	Max.
/i/ f1	NH	11	532	40.55	467	583
	SHI	9	620	73.22	483	719
	PHI	13	662	90.32	503	883
/i/ f2	NH	11	2910	334.97	2410	3493
	SHI	9	2509	267.10	2127	2853
	PHI	13	2494	306.58	1981	2932
Closed /e/ f1	NH	11	536	56.18	455	655
	SHI	9	627	100.23	507	798
	PHI	13	652	142.50	431	1005
Closed /e/ f2	NH	11	2686	267.25	2398	3352
	SHI	9	2359	205.17	2047	2632
	PHI	13	2120	464.70	815	2612
Open /e/ f1	NH	11	738	158.07	519	1065
	SHI	9	692	133.68	464	867
	PHI	13	736	204.04	497	1153
Open /e/ f2	NH	31	2400	201.57	2174	2812
	SHI	14	2267	219.41	1968	2553
	PHI	25	2152	345.83	1338	2732
/a/ f1	NH	32	879	126.12	719	1051
	SHI	14	991	76.20	882	1095
	PHI	25	1007	148.69	784	1293
/a/ f2	NH	11	1572	209.49	1143	1828
	SHI	9	1727	152.92	1506	2001
	PHI	13	1754	198.81	1457	2006
Open /o/ f1	NH	11	608	118.49	494	824
	SHI	9	799	130.99	671	1078
	PHI	13	793	134.19	628	1052
Open /o/ f2	NH	11	1206	121.20	1040	1425
	SHI	9	1448	113.29	1298	1686
	PHI	13	1535	207.99	1253	1783
Closed /o/ f1	NH	11	554	52.55	503	684
	SHI	9	670	81.52	570	841
	PHI	13	691	123.36	459	941
Closed /o/ f2	NH	11	1110	88.40	976	1255
	SHI	9	1301	112.47	1163	1527
	PHI	13	1348	222.14	982	1813
/u/ f1	NH	32	551	38.90	492	614
	SHI	14	590	65.58	491	675
	PHI	25	639	115.99	434	808
/u/ f2	NH	32	975	115.87	744	1095
	SHI	14	1104	136.24	881	1243
	PHI	25	1222	141.81	1060	1556

Table 2: Descriptive statistics of the formant values (in Hertz) for the Slovenian vowels in children with normal hearing (NH), severe (SHI) and profound hearing loss (PHI)

		N	Mean	SD
d_i_f1	NH	11	.06	40.55
	SHI	9	-87.75	73.22
	PHI	13	-130.49	90.32
d_i_f2	NH	11	-.45	334.97
	SHI	9	400.96	267.10
	PHI	13	416.39	306.58
d_closed_e_f1	NH	11	.39	56.18
	SHI	9	-90.85	100.23
	PHI	13	-115.80	142.50
d_closed_e_f2	NH	11	.47	267.25
	SHI	9	326.82	205.17
	PHI	13	566.07	464.70
D_open_e_f1	NH	10	.17	158.07
	SHI	9	45.54	133.68
	PHI	13	1.63	204.04
d_open_e_f2	NH	10	-.17	201.57
	SHI	9	132.94	219.41
	PHI	13	248.17	345.83
d_a_f1	NH	11	.1250	126.12
	SHI	9	-111.94	76.20
	PHI	13	-127.76	148.69
d_a_f2	NH	11	.13	209.49
	SHI	9	-155.37	152.92
	PHI	13	-181.58	198.81
d_open_o_f1	NH	11	.40	118.49
	SHI	9	-191.00	130.99
	PHI	13	-185.00	134.19
d_open_o_f2	NH	11	.16	121.20
	SHI	9	-242.18	113.29
	PHI	13	-329.26	207.99
D_closed_o_f1	NH	11	-.02	52.55
	SHI	9	-116.23	81.52
	PHI	13	-136.74	123.36
d_closed_o_f2	NH	10	-.41	88.40
	SHI	9	-190.73	112.47
	PHI	13	-238.15	222.14
d_u_f1	NH	10	.25	38.90
	SHI	9	-38.52	65.58
	PHI	13	-88.39	115.99
d_u_f2	NH	11	.3977	115.87
	SHI	9	-129.05	136.24
	PHI	13	-246.54	141.81

Table 3: Descriptives of the deviation among SHI and PHI speakers in formant frequency (Hertz) from the standard formant value as derived from the control group data set (NH)

Generally, the means of F1 are usually slightly higher in the anterior and posterior vowels in the PHI compared with formant values of those with SHI; with the SHI group the means of F2 are usually higher in the anterior vowels and lower in the posterior vowels compared to the PHI. Compared with formant values of the children with NH, a lower F2 might be assumed in children with PHI and SHI for the front vowels, which normally have a high F2, and a higher F2 frequency would be produced for the back vowels, which normally have a low F2. The F1 values are normally higher for the front and the back vowels with children with both PHI and SHI (except F1 of open /e/ where the speakers with SHI have lower F1 than PHI and NH speakers).

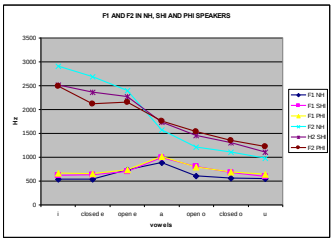


Figure 1: Vowel production: F1 and F2 in of NH, SHI and PHI children

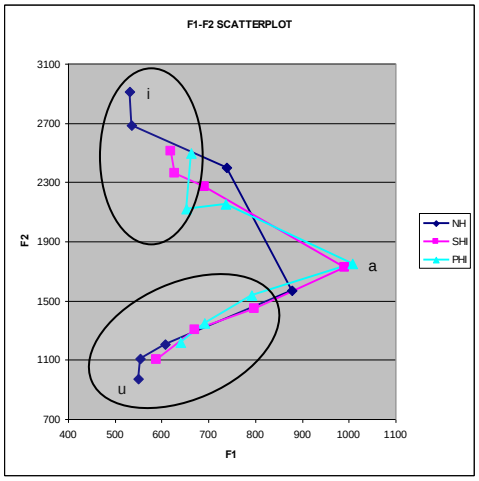


Figure 2: Vowel production in F2 – F1 plane of NH, SHI and PHI children

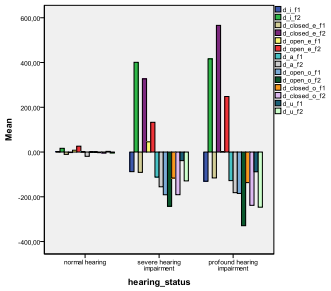


Figure 3: The deviation among SHI and PHI speakers in formant frequency from the standard formant value as derived from the control group data set (NH)

Test of homogeneity of variances				
Vowel / formant	Levene Statistic	df1	df2	Sig.
Open /o/ f2	6.370	2	28	.005*
/u/ f1	5.168	2	30	,012*
Welch Robust Tests of Equality of Means				
/i/ f1	13.196	2	17.092	.000*
/i/ f2	5.804	2	19.291	.011*
Closed /e/ f1	5.411	2	16.967	.015*
Closed /e/ f2	8.150	2	19.925	.003*
Open /o/ f1	7.207	2	17.605	.005*

Open /o/ f2	13.995	2	18.607	.000*
Closed /o/ f1	10.656	2	17.666	.001*
Closed /o/ f2	11.678	2	18.499	.001*
/u/ f1	3,868	2	17,016	,041*
/u/ f2	10,715	2	18,850	,001*

Table 4: Test of homogeneity of variances and Welch Robust Tests of Equality of Means

(I) hearing status	(J) hearing Status	Mean Difference (I-J)	Std. Error	Sig.
Dependent Variable /i/ f1				
NH	SHI	-87.8063*	32.53751	.034
	PHI	-130.5462*	29.65682	.000
Dependent Variable /i/ f2				
NH	SHI	401.4079*	137.81961	.020
	PHI	416.8406*	125.61781	.007
Dependent Variable closed /e/ f1				
NH	PHI	-116.1948*	44.60315	.042
Dependent Variable closed /e/ f2				
NH	PHI	565.6009*	142.74826	.001
Dependent Variable open /o/ f1				
NH	SHI	-191.3882*	60.79372	.012
	PHI	-185.3900*	55.92209	.008
Dependent Variable open /o/ f2				
NH	SHI	-242.3359*	76.60001	.011
	PHI	-329.4118*	70.46176	.000
Dependent Variable closed /o/ f1				
NH	SHI	-116.2069*	42.11623	.029
	PHI	-136.7210*	38.38748	.004
Dependent Variable closed /o/ f2				
NH	SHI	-190.3168*	72.07830	.039
	PHI	-237.7365*	65.69688	.003
Dependent Variable /u/ f1				
NH	PHI	-88.6045*	34.35591	.045
Dependent Variable /u/ f2				
NH	PHI	-246.9376*	54.14627	.000

Table 5: Multiple Comparisons - Bonferroni post hoc analysis of vowel formant variables between children with NH, with SHI and PHI (* The mean difference is significant at the .05 level)

According to the hypothesis H2, we can state that NH speakers show a larger range in F2 formant production from anterior to posterior vowels (2910 Hz-975 Hz=1935 Hz) in comparison with speech production in children with SHI (2509 Hz-1104 Hz=1405 Hz) and PHI (2494 Hz-1222 Hz=1272 Hz). In the F1 formant production of high and low vowels, NH speakers show higher F1 values (532/551 Hz-879 Hz=347/328 Hz) in comparison with the vowel production in children with PHI (662/639 Hz–1007 Hz=368/445 Hz) and those with SHI (620/590 Hz–991 Hz=401/371 Hz). As a result, ranges in NH speakers are larger than in SHI and PHI speakers.

Comparing the children with profound and severe hearing impairment, generally, there are smaller standard deviations in the severe hearing impairment group and greater standard deviations in the profound hearing impairment group. In the children with profound and severe hearing impairment the standard deviations in some variables (i.e. F1 of /i/, closed and open /e/, open and closed /o/, /u/ and F2 of open /e/ and /o/, closed /o/ and /u/) are much greater than in those with normal hearing; the highest standard deviation is found in the group with profound hearing impairment.

From Figures 1 and 2, it is clear that the formant space (F1-F2 scatterplot) with profound hearing impairment is smaller than that with severe impairment and that the space in SHI children is smaller than that in the NH children. The greatest differences are in the anterior vowel production: normal-hearing children differentiate the three anterior vowels much better than those with severe and profound hearing impairment, especially the closed and open /e/. The greatest differences are between children with normal hearing and children with profound hearing impairment, especially in the second formant values (Table 2).

To clarify the differences we computed the deviations among speakers with SHI and PHI in formant frequency from the standard formant values as derived from the control group data set (NH speakers). Table 3 shows the greatest deviation in PHI speakers for all formant values, except in F1 of open /e/ and open /o/. These two vowels are influenced by the dialect spoken by speakers, and they are very variable. Considering that open /e/ and open /o/ are more visible than front and back vowels and more open, the speakers with PHI may use more visible feedback than the speakers with SHI. The Figure 3 is very informative: all the formant values in speakers with SHI and PHI deviate from the control group data set. Greater deviations are visible in speakers with PHI, comparing with those of SHI speakers; more detailed, the anterior F2 values are lower in frequency and the difference is positive, whereas the F2 of back vowels and the F1 values of all vowels are higher and the difference is negative.

As the degree of hearing loss increased the deviation in frequency from the standard second formant values of the front values, as derived from the control group data set, increased (positive values) and the values of the F2 were lower than those in the control group data set; the deviation in the back values decreased, the first formants increased in the extreme back and front vowels and decreased in the middle-low vowel.

Simultaneously comparing the means of the three groups (Table 4) yielded the results that at $p<.05$ all the variables except F1 of /u/ and F2 of open /o/ are homogeneous in variance. As seen in Table 4, all the variables of high front and back rounded vowels, except F1 and F2 of /a/ and open /e/ did not pass the robust test of equality of means (Welch method, $p<.05$). The greatest differences in the second formant are for the anterior high vowel /i/ and for the back high rounded vowel open /o/, and the greatest differences in first formant values are for the back high rounded vowels open /o/ and closed /o/.

The results of analysing the differences between the three pairs of groups (NH – SHI, NH – PHI, SHI – PHI) in the Bonferroni post-hoc analysis, the values in Table 5 show that the only formant values that are consistent between the three groups, are the first and second

formants of the vowels /a/ and open /e/, the most central vowels, where the auditory control is minimal: the only movement required is a jaw vertical movement with minimal tongue movement.

According to the results, we can accept the first, second and third hypotheses in their entirety. The F2 values are more at risk for neutralisation / clustering / overlapping than F1 values. In addition to the generally reduced perception of F2 formants, tongue placement along the front-back axis in the oral cavity is difficult to perceive visually. On the other hand, better residual hearing in the region of F1 frequencies and relatively better visibility of tongue height associated with jaw displacement, which can be accessible in speech reading, makes variation in F1 more prominent (Nicolaidis & Sfakiannaki, 2007). Comparing our speakers and formant values, and the conclusions of the cited research, we can say that there is a general belief that the formant space in speakers with hearing impairment is reduced, the F2 line inverted, and F2 values clustered. Comparing to the results of Subtenly et al. (1992), with F2 frequency range 1500-2100 Hz, in our analysis F2 values clustered from 2509 Hz to 1727 HZ. Shizuo and Ryuzaemon (1957) reported that [i] and [o] in speakers with hearing impairment , aged 6-11 years, deviated most from the production in NH speakers. In our research great deviations were in /i/ and (closed and open) /o/ and in closed /e/ and /u/. Schenk, Baumgartner and Hamzavi (2003) showed differences in all F1 values and differences in F2 in [e] only. Our study – comparing NH and HI groups - found differences in all high front and back vowels (in F1 and F2). Only open /e/ and /a/ didn't differ in the three groups. It is important to underline that Slovenian language has 13 stressed vowels (7 long and 6 short) and 6 unstressed short ones; a direct comparison with other languages with other vowel systems is consequently inappropriate. We may compare the rules of vowel neutralisation, clustering or overlapping, the tendencies of range reduction. We can say that our study and the cited ones are similar.

The presented findings might have an impact to the language teaching process for the children with hearing disabilities: they offer to speech and language therapists the norms of formant production and encourage the use of audio-video feedback tools for speech production monitoring to widen the formant production space.

Future studies should examine the F2/F1 ratios in vowel production in NH, SHI and PHI group in order to understand the similarities between speakers and groups and to compare these ratios with ratios of older speakers (adult males and females).

4. References

Angelocci, A.A., Kopp, G.A: Holbrook A. (1964). The Vowel Formants of Deaf and Normal - Hearing Eleven - to Fourteen-Year-Old Boys. Journal of Speech and Hearing Disorders, 29, 2, 156 - 170.

Engwall, O. (1999). Vocal tract modelling in 3D. TMH-QPSR, 40, 1-2: 031-038.

Fletcher, S.G. (1975). Visual articulatory modelling and shaping: a new approach to developing speech of the deaf. Proceedings of the 18th International Congress on Education of the Deaf. 1995; Volume II: 757-758.

Globačnik, B. (1999). Ocena artikulacije govora. Ljubljana, CenterKontura.

Hacki, T., Heitmüller, S. (1999). Development of the child's voice: premutation, mutation. International Journal of Pediatric Otorhinolaryngology, 49, Issue null, 141-144.

Markides, A. (1983). The Speech of Hearing Impaired Children. Manchester: Manchester University Press.

Monsen, R. (1976). Normal and reduced phonological space: the production of English vowels by deaf adolescents. Journal of Phonetics 4, 189-198.

Murphy, E., Dodds, B. (2007). Hearing impairment. In: Dodds, B. (eds): Differential diagnosis and treatment of childen with speech disorder (pp. 244 – 257). London and Philadelphia, Whurr publishers.

Nasir, S.M., Ostry, D.J. (2008). Speech Motor Learning in Profoundly Deaf Adults. Nature Neuroscience, 11, 1217 - 1222 .

Nasir, S.M. Ostry, D.J. (2006). Report: Somatosensory Precision in Speech Production. Current Biology 16, 1918–1923, October 10..

Nicolaidis, K., Sfakiannaki, A. (2007). An acoustic analysis of vowels produced by Greek speakers with hearing impairment. ICPhS, Saarbrücken. Accessed November 11, 2008. Available: from http://www.icphs2007.de/conference/Papers/1358/1358.pdf.

Purcell, D.W., Munhall, K.G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. J. Acoust. Soc. Am. 120, 966–977.

Ryalls, J., Larouche, A., Giroux, F. (2003). Acoustic comparison of CV syllables in French–speaking children with normal hearing, moderate-to-severe and profound hearing impairment. Journal of Multilingual Communication Disorders, 1, 99-114.

Schenk, B.S., Baumgartner, W.D., Hamzavi, J.S. (2003). Effect of the loss of auditory feedback on segmental parameters of vowels of postlingually deafened speakers. Auris Nasus Larynx , 30, 333-339.

Shizuo, H., Ryuzaemon, K. (1975). Some properties of formant frequencies of vowels by deaf and hard of hearing children. The Journal of the Acoustical Society of Japan, 31, 3, 163-169.

Subtelny, J.D., Whitehead, R.L., Samar, V.J. (1992). Spectral study of deviant resonance in the speech of women who are deaf. Journal of Speech and Hearing Research, 35, 574-579.

Waldstein, R.S.(1990). Effects of postlingual deafness on speech production: implications for the role of auditory feedback. J. Acoust. Soc. Am. 88, 2099–2114.

Indeks avtorjev / Author index

Zbornik 13. mednarodne multikonference **INFORMACIJSKA DRUŽBA - IS 2010**
Proceedings of the 13th International Multiconference **INFORMATION SOCIETY - IS 2010**
Zvezek C / Volume C

JEZIKOVNE TEHNOLOGIJE
LANGUAGE TECHNOLOGIES

14. do 15. oktober 2010, Ljubljana / 14th - 15th October 2010, Ljubljana, Slovenia
Uredila / edited by: Tomaž Erjavec / Jerneja Žganec Gros