Uredniki:

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

dr. Jerneja Žganec Gros
Alpineon d.o.o, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Tisk: Birografika BORI d.o.o.
Priprava zbornika: Mitja Lasič
Oblikovanje naslovnice: dr. Damjan Demšar
Tiskano iz predloga avtorjev
Naklada: 50

Ljubljana, oktober 2008

Konferenco IS 2008 sofinancirata
Ministrstvo za visoko šolstvo, znanost in tehnologijo
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)
004.8(063)

KONFERENCA Jezikovne tehnologije (6 ; 2008 ; Ljubljana)
Zbornik Šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008, Ljubljana, Slovenia : zbornik 11. mednarodne multikonference Informacijska družba - IS 2008, zvezek C = Proceedings of the Sixth Language Technologies Conference, October 16th-17th, 2008 : proceedings of the 11th International Multiconference Information Society - IS 2008, volume C / uredila, edited by Tomaž Erjavec, Jerneja Žganec Gros. - Ljubljana : Institut "Jožef Stefan", 2008. - (Informacijska družba, ISSN

1581-9973)

ISBN 978-961-264-006-4
1. Jezikovne tehnologije 2. Language Technologies 3. Informacijska družba 4. Information society 5. Erjavec, Tomaž, 1960- 6. Mednarodna multi-konferenca Informacijska družba (11 ; 2008 ; Ljubljana)
241520896

Zbornik 11. mednarodne multikonference
INFORMACIJSKA DRUŽBA - IS 2008
Zvezek C

Proceedings of the 11th International Multiconference
INFORMATION SOCIETY - IS 2008
Volume C

Zbornik
Šeste konference JEZIKOVNE TEHNOLOGIJE

Proceedings of the
Sixth Language Technologies Conference

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

16. do 17. oktober 2008 / October 16th - 17th, 2008
Ljubljana, Slovenia

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2008

V svojem enajstem letu ostaja multikonferenca Informacijska družba (http://is.ijs.si) ena vodilnih srednjeevropskih konferenc, ki združuje znanstvenike z različnih raziskovalnih področij, povezanih z informacijsko družbo. V letu 2008 smo v multikonferenco povezali osem neodvisnih konferenc. Informacijska družba postaja vedno bolj zapleten socialni, ekonomski in tehnološki sistem, ki je pritegnil pozornost vrste specializiranih konferenc v Sloveniji in Evropi. Naša multikonferenca izstopa po širini in obsegu tem, ki jih obravnava.

Rdeča nit multikonference ostaja sinergija interdisciplinarnih pristopov, ki obravnavajo različne vidike informacijske družbe ter poglabljajo razumevanje informacijskih in komunikacijskih 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 z 31-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2008 sestavljajo naslednje samostojne konference:

- BIOMA 2008 – Optimizacijske metode po vzorih iz narave in njihova uporaba
- Inteligentni sistemi
- Jezikovne tehnologije
- Kognitivne znanosti
- Rudarjenje podatkov in podatkovna skladišča (SiKDD 2008)
- Slovenija pred demografskimi izzivi 21. stoletja
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Vzgoja in izobraževanje v informacijski družbi

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija. Zahvaljujemo se tudi Ministrstvu za visoko šolstvo, znanost in tehnologijo za njihovo 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 2008 sta se programski in organizacijski odbor odločila, da bosta podelila posebno priznanje Slovencu ali Slovenki za izjemen prispevek k razvoju in promociji informacijske družbe v našem okolju. Z večino glasov je letošnje priznanje pripadlo prof. dr. Ivanu Rozmanu. Čestitamo!

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

FOREWORD - INFORMATION SOCIETY 2008

In its 11th year, the Information Society Multiconference (http://is.ijs.si) continues as one of the leading conferences in Central Europe gathering scientific community with a wide range of research interests in information society. In 2008, we organized eight independent conferences forming the Multiconference. Information society displays a complex interplay of social, economic, and technological issues that attract attention of many scientific events around Europe. The broad range of topics makes our event unique among similar conferences. The motto of the Multiconference is 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 31 years of tradition in excellent research publications.

The Information Society 2008 Multiconference consists of the following conferences:

- BIOMA 2008 – Bioinspired Optimization Methods and their Applications
- Cognitive Sciences
- Collaboration, Software and Services in Information Society
- Data Mining and Data Warehouses (SiKDD 2008)
- Education in Information Society
- Intelligent Systems
- Language Technologies
- Slovenian Demographic Challenges in the 21st Century

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.

In 2008, the Programme and Organizing Committees decided to award one Slovenian for his/her outstanding contribution to development and promotion of information society in our country. With the majority of votes, this honor went to Prof. Dr. Ivan Rozman. Congratulations!

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 revieewers 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
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

Organizational 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
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

KAZALO / TABLE OF CONTENTS

Language Technologies ***1***

PREDGOVOR / PREFACE 3

RECENZENTI / REVIEWERS 4

One corpus-based semantic model, many semantic tasks / Marco Baroni 7

Rapid Deployment of Speech Processing Systems to New Languages and Domains / Tanja Schultz 9

Postopek za izbiro govornih segmentov pri vgrajeni polifonski združevalni sintezi govora / Jerneja Žganec Gros, Aleš Mihelič, Mario Žganec, Uliana Dorofeeva, Nikola Pavešić 10

Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast News / Andrej Žgank, Marko Kos, Bojan Kotnik, Mirjam Sepesy Maučec, Tomaž Rotovnik, Zdravko Kačič 16

Vpliv predhodne segmentacije govor/negovor na segmentacijo govorcev / Matej Grašič, Marko Kos, Zdravko Kačič 20

Označevanje vrste diskurznih označevalcev / Darinka Verdonik 25

Validacija označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint / Darinka Verdonik, Andrej Žgank, Agnes Pisanski Peterlin 29

Črpanje primerov za japonsko-slovenski slovar iz vzporednega korpusa / Kristina Hmeljak Sangawa, Tomaž Erjavec 33

Predstavitev in analiza slovenskega wordneta / Darja Fišer, Tomaž Erjavec 37

Samodejno luščenje slovarja iz vzporednega korpusa s pomocjo vmesnega jezika in pomenskega razdvoumljanja / Peter Holozan 43

Oblikoskladenjske specifikacije in označeni korpusi JOS / Tomaž Erjavec, Simon Krek 49

Oblikoskladenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko oblikoskladenjsko označevanje slovenščine / Špela Arhar, Nina Ledinek 54

Vpliv namembnosti korpusa na označevanje besedilnega gradiva za »Večjezični korpus turističnih besedil« / Vesna Mikolič, Ana Beguš, Davorin Dukič, Miha Koderman 60

iKorpus in luščenje izrazja za Islovar / Špela Vintar, Tomaž Erjavec 65

AvID: Audio–Video Emotional Database / Rok Gajšek, Anja Podlesek, Luka Komidar, Gregor Sočan, Boštjan Bajec, Vitomir Štruc, Valentin Bucik, France Mihelič 70

Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language / Vesna Satev, Nicolas Nikolov 75

Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata / Damir Čavar, Ivo-Pavao Jazbec, Siniša Runjaić 80

Productivity of concepts in Serbian Wordnet / Jelena Tomašević, Gordana Pavlović-Lažetić 86

A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators / Tim vor der Br¨uck, Sven Hartrumpf, Hermann Helbig 92

Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages / Jernej Vičič 98

Part-of-Speech Tagging of Slovenian, 12 years after / Primož Jakopin, Aleksandra Bizjak Končar 104

Improving morphosyntactic tagging of Slovene by tagger combination / Jan Rupnik, Miha Grčar, Tomaž Erjavec 110

Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian / Željko Agić, Marko Tadić, Zdravko Dovedan 116

Indeks avtorjev / Author index ***123***

Zbornik 11. mednarodne multikonference
INFORMACIJSKA DRUŽBA - IS 2008

Proceedings of the 11th International Multiconference
INFORMATION SOCIETY - IS 2008

Zbornik
Šeste konference JEZIKOVNE TEHNOLOGIJE

Proceedings of the
Sixth Language Technologies Conference

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

http://is.ijs.si

16. do 17. oktober 2008 / October 16th - 17th, 2008
Ljubljana, Slovenia

PREDGOVOR K ZBORNIKU ŠESTE KONFERENCE »JEZIKOVNE TEHNOLOGIJE«

V pričujočem zborniku so objavljeni prispevki s šeste konference "Jezikovne tehnologije", ki je potekala 16. in 17. oktobra 2008 v Ljubljani, v okviru multikonference "Informacijska družba" IS'2008. Konferenca zaznamuje tudi deset let, kolikor je minilo od prve slovenske konference o jezikovnih tehnologijah, ki je potekala leta 1998 v Cankarjevem domu v Ljubljani. 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 mednarodne konference o jezikovnih tehnologijah, organizirane v okviru IS. Zbornik vsebuje 23 prispevkov, ki obravnavajo široko paleto raziskav; posebej izstopa veliko število prispevkov o izdelavi korpusov in drugih jezikovnih virov, dobro zastopani pa so tudi prispevki 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'2008.

Preface to the Proceedings of the Sixth Language Technologies Conference

These proceedings contain the contributions for the Sixth Language Technologies Conference, which took place on October 16th and 17th 2008 in Ljubljana, in the scope of the Information Society Multiconference, IS'2008. This conference also marks ten years that have passed since the first Slovene conference on language technologies, which took place in 1998 in Ljubljana. This year's 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 international conference on Language Technologies organised in the scope of IS. The proceedings contain 23 contributions, which present a wide variety of research topics; especially numerous are contributions dealing with creation and usage of corpora and other language resources, while papers about speech technologies are also well represented. 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 and demo presentations, the programme committee of the conference and the organising committee of IS 2008.

Tomaž Erjavec, Jerneja Žganec Gros
Ljubljana, October 2008.

RECENZENTI

- prof. dr. Steven Bird, Dept. of Comp. Sci. and Software Engineering, University of Melbourne (Avstralija)
- prof. dr. Nick Campbell, ATR (Japonska)
- doc. dr. Jan Cernocký, Faculty of Information Technology, Brno Technical University (Češka)
- dr. Simon Dobrišek, Fakulteta za elektrotehniko, Univerza v Ljubljani
- prof. dr. Christoph Draxler, Institute of Phonetics and Speech Communication, Ludwig-Maximilians Uni. Munich (Nemčija)
- doc. dr. Tomaž Erjavec (predsednik), Odsek za tehnologije znanja, Institut "Jožef Stefan"
- dr. Anna Esposito, Institute for Advanced Scientific Studies (Italija)
- dr. Anna Feldman, Dept. of Linguistics and Dept. of Computer Science, Montclair State University (ZDA)
- prof. dr. Sadaoki Furui, Graduate School of Information Science and Engineering, Tokyo Institute of Technology (Japonska)
- prof. dr. Carmen Garcia-Mateo, ETSI Telecomunicacion, Vigo University (Španija)
- doc. dr. Vojko Gorjanc, Filozofska fakulteta, Univerza v Ljubljani
- prof. dr. Nancy Ide, Department of Computer Science, Vassar College (ZDA)
- doc. dr. Bojan Imperl, Iskratel d.o.o.
- prof. dr. Ivo Ipsić, Faculty of Engineering, University of Rijeka (Hrvaška)
- prof. dr. Zdravko Kačič, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
- dr. Adam Kilgarriff, Lexical Computing Ltd (Velika Britanija)
- prof. dr. Ewan Klein, HCRC/ICCS, University of Edinburgh (Velika Britanija)
- prof. dr. Steven Krauwer, Utrecht University (Nizozemska)
- doc. dr. Cvetana Krstev, Arts Faculty, University of Belgrade (Srbija in Črna gora)
- dr. Siegfried Kunzmann, European Media Laboratory GmbH (Nemčija)
- dr. Birte Loenneker-Rodman, International Computer Science Institute Berkley (ZDA)
- prof. dr. Markéta Lopatková, Institute of Formal and Applied Linguistics, Charles University (Češka)
- prof. dr. Bente Maegaard, Centre for Language Technology, University of Copenhagen (Danska)
- prof. dr. Jean-Pierre Martens, Dept. of Electronics and Information Systems, University of Gent (Belgija)
- prof. dr. France Mihelič, Fakulteta za elektrotehniko, Univerza v Ljubljani
- prof. dr. Bernd Moebius, University of Stuttgart (Nemčija)
- doc. dr. João Paulo Neto, Spoken Language Laboratory, INESC-ID (Portugalska)
- dr. Elmar Nöth, Technical Faculty, Friedrich-Alexander University Erlangen-Nuremberg (Nemčija)
- prof. dr. Karel Pala, Faculty of Informatics, Masaryk University (Češka)
- prof. dr. Serge Sharoff, Centre for Translation Studies, University of Leeds (Velika Britanija)
- prof. dr. Marko Stabej, Filozofska fakulteta, Univerza v Ljubljani
- prof. dr. Yannis Stylianou, University of Crete (Grčija)
- prof. dr. Rastislav Šuštaršič, Filozofska fakulteta, Univerza v Ljubljani
- prof. dr. Marko Tadić, Department of linguistics, University of Zagreb (Hrvaška)
- dr. Jörg Tiedemann, Alfa-Informatica, Rijksuniversiteit Groningen (Nizozemska)
- prof. dr. Tamás Váradi, Linguistics Institute, Hungarian Academy of Sciences (Madžarska)
- dr. Darinka Verdonik, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
- doc. dr. Špela Vintar, Filozofska fakulteta, Univerza v Ljubljani
- doc. dr. Andreja Žele, Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU
- dr. Jerneja Žganec Gros (predsednica), Alpineon, d.o.o.
- dr. Janez Žibert, Univerza na Primorskem

REVIEWERS

- Steven Bird, Dept. of Computer Science and Software Engineering, University of Melbourne (Australia)
- Nick Campbell, ATR (Japan)
- Jan Cernocký, Faculty of Information Technology, Brno Technical University (Czech Republic)
- Simon Dobrišek, Faculty of Electrical Engineering, University of Ljubljana (Slovenia)
- Christoph Draxler, Inst. of Phonetics and Speech Comm., Ludwig-Maximilians Uni. Munich (Germany)
- Tomaž Erjavec (chair), Dept. of Knowledge Technologies, Jožef Stefan Institute (Slovenia)
- Anna Esposito, Institute for Advanced Scientific Studies (Italija)
- Anna Feldman, Dept. of Linguistics and Dept. of Computer Science, Montclair State University (USA)
- Sadaoki Furui, Graduate School of Inf. Science and Engineering, Tokyo Institute of Technology (Japan)
- Carmen Garcia-Mateo, ETSI Telecomunicacion, Vigo University (Spain)
- Vojko Gorjanc, Faculty of Arts, University of Ljubljana (Slovenia)
- Nancy Ide, Department of Computer Science, Vassar College (USA)
- Bojan Imperl, Iskratel Ltd (Slovenia)
- Ivo Ipsić, Faculty of Engineering, University of Rijeka (Croatia)
- Zdravko Kačič, Faculty of Electrical Engineering and Computer Science, University of Maribor (Slovenia)
- Adam Kilgarriff, Lexical Computing Ltd (UK)
- Ewan Klein, HCRC/ICCS, University of Edinburgh (UK)
- Steven Krauwer, Utrecht University (The Netherlands)
- Cvetana Krstev, Arts Faculty, University of Belgrade (Serbia and Montenegro)
- Siegfried Kunzmann, European Media Laboratory GmbH (Germany)
- Birte Loenneker-Rodman, International Computer Science Institute Berkley (USA)
- Markéta Lopatková, Institute of Formal and Applied Linguistics, Charles University (Czech Republic)
- Bente Maegaard, Centre for Language Technology, University of Copenhagen (Denmark)
- Jean-Pierre Martens, Department of Electronics and Information Systems, University of Gent (Belgium)
- France Mihelič, Faculty of Electrical Engineering, University of Ljubljana (Slovenia)
- Bernd Moebius, University of Stuttgart (Germany)
- João Paulo Neto, Spoken Language Laboratory, INESC-ID (Portugal)
- Elmar Nöth, Technical Faculty, Friedrich-Alexander University Erlangen-Nuremberg (Germany)
- Karel Pala, Faculty of Informatics, Masaryk University (Czech Republic)
- Serge Sharoff, Centre for Translation Studies, University of Leeds (UK)
- Marko Stabej, Faculty of Arts, University of Ljubljana (Slovenia)
- Yannis Stylianou, University of Crete (Greece)
- Rastislav Šuštaršič, Faculty of Arts, University of Ljubljana (Slovenia)
- Marko Tadić, Department of linguistics, University of Zagreb (Croatia)
- Jörg Tiedemann, Alfa-Informatica, Rijksuniversiteit Groningen (Netherlands)
- Tamás Váradi, Linguistics Institute, Hungarian Academy of Sciences (Hungary)
- Darinka Verdonik, Faculty of Electrical Engineering and Computer Science, Uni of Maribor (Slovenia)
- Špela Vintar, Faculty of Arts, University of Ljubljana (Slovenia)
- Andreja Žele, Fran Ramovš Institute of the Slovenian Language, ZRC-SAZU (Slovenia)
- Jerneja Žganec Gros (chair), Alpineon Ltd (Slovenia)
- Janez Žibert, University of Primorska (Slovenia)

One corpus-based semantic model, many semantic tasks

Marco Baroni

Center for Mind/Brain Sciences (CIMeC)
University of Trento
Palazzo Fedrigotti, C.so Bettini 31
38060 Rovereto (TN), Italy
marco.baroni@unitn.it

1. Introduction

The past 20 years have seen the birth of many systems that learn aspects of semantics from patterns of co-occurrence in naturally occurring data, mostly in the form of large scale linguistic corpora. *Relation extraction* algorithms (Hearst, 1992; Banko et al., 2007) look for pairs that instantiate a certain relation type, for example the *is-a* relation (cars are vehicles, wolves are mammals, ...). *Word space models* (Sahlgren, 2006) measure taxonomic similarity among concepts by the degree to which they tend to share similar contexts. These models achieve very good results in tasks such as synonym identification and categorization. Finally, corpus-based semantics recently began to tackle aspects of *semantic compositionality*, the ability to (recursively) create composite meanings from simpler ones (Kintsch, 2001; Mitchell and Lapata, 2008). Compositionality is a complex phenomenon, but one particularly important facet of it pertains to compatibility constraints on composition (Resnik, 1996; Erk, 2007): Even if you have not heard either sentence before, *I killed a kangaroo* makes more sense than *I killed a book*.

I argue that one of the most important goals for corpus-based semantics in the next few years is to develop a common corpus-derived model that can be tuned to perform all these tasks and more. This is a necessary step both in theoretical terms (human semantic knowledge is flexible enough to allow us to perform all these tasks) and for practical purposes (embedding semantic knowledge in applications should be a matter of tuning an existing resource to a new task, and it should not require going back to a corpus and training a new ad-hoc system each time).

2. Semantic models as target+type+feature tuples

All the tasks sketched above can indeed be performed by a model that stores corpus-based information as a weighted list of *target+type+feature* tuples. The target list will typically be extracted from the corpus according to some criterion (e.g., the top N words, the top N words filtered by part-of-speech, etc.). Types are strings that cue the relation that exists between the target and the feature. For example, they can be dependency paths extracted from a parse of the corpus, or (generalizations of) lexico-syntactic patterns. The features are words (or multi-word elements) that are syntagmatically linked to the targets by the relevant type (e.g., the verb *to eat* might be a feature of the target *apple*, with relation type *obj*). Finally, weighting is based on standard statistical measures such as Mutual Information or Log-Likelihood Ratio (Evert, 2004), that mark the degree of syntagmatic association between the elements of a tuple.

3. Examples

While any model equipped with the relevant information would do, I present here examples from StruDEL (Baroni et al., submitted), a fully unsupervised model trained on a large corpus of English Web pages (about 2 billion tokens). Table 1 reports the top 10 tuples extracted by StruDEL for the targets *book* and *tiger* (using Log-Likelihood Ratio weighting).

Given these tuples, the task of relation extraction is rather straightforward. We can identify, either by hand or using standard weakly supervised methods, types that instantiate the relation of interest, and look for target-feature pairs that occur in highly weighted tuples with the extracted types. For example, it is pretty clear from the examples in Table 1 that we could use the type *T in F* (among others) to harvest features that instantiate the location relation. In Baroni et al. (submitted) we show that this approach compares favorably against a state-of-the-art ad-hoc relation extraction algorithm, when searching for locations and functions.

Taxonomic similarity, such as has been traditionally extracted by word space models, can be computed by building a matrix whose rows are the targets, the columns (dimensions) are type+feature pairs, and the tuple weights fill the corresponding cells (e.g., the *book* vector will have a *T for F: reader-n* dimension with value 5024.4). Cosines or other standard measures of similarity can be computed from this matrix. Table 2 shows the nearest neighbours of *book* and *tiger* collected in this way. As these examples suggest, and as we confirmed quantitatively in Baroni et al. (submitted), vectors constructed from type-feature dimensions can do very well in tasks requiring information about taxonomic similarity.

For the final task, modeling compatibility restrictions on compositionality, I focus on the classic problem of checking whether a noun is appropriate to fill a verbal argument slot. First, I extract the targets (nouns) occurring in the top N (in my experiments: 30) tuples that contain the verb of interest as feature and a type marking the argument relation of interest. I call these nouns the "model set" for the slot. I then compute the similarity of other nouns to the centroid

book			*tiger*		
type	*feature*	*weight*	*type*	*feature*	*weight*
T for F	reader-n	5024.4	T in F	jungle-n	183.5
T by F	author-n	4406.7	T in F	zoo-n	164.4
obj	read-v	4314.3	F on T	lion-n	103.3
T in F	library-n	3925.1	F by T	maul-v	101.4
F in T	chapter-n	3303.4	subj	kill-v	93.6
obj	write-v	2919.0	F on T	stripe-n	90.1
subj	publish-v	2454.5	T in F	cage-n	83.3
T from F	publisher-n	2417.2	F such as T	specie-n	77.3
F of T	page-n	1836.4	F for T	habitat-n	73.3
T on F	subject-n	1422.4	T from F	extinction-n	73.7

Table 1: Top ranked tuples for targets *book* and *tiger*

book		*tiger*	
neighbour	*cosine*	*neighbour*	*cosine*
story	0.27	gorilla	0.33
magazine	0.25	elephant	0.29
photograph	0.08	bear	0.28
wrapper	0.07	lion	0.28
tape	0.06	fox	0.24
disc	0.05	monkey	0.21
painting	0.03	leopard	0.20
table	0.03	wolf	0.17
tool	0.03	rhino	0.17
library	0.02	penguin	0.16

Table 2: Nearest neighbours of books and tigers

object		*instrument*	
noun	*cosine*	*noun*	*cosine*
kangaroo	0.52	heroin	0.33
snake	0.29	stone	0.31
person	0.27	antibiotic	0.30
fun	0.22	brick	0.25
robot	0.15	hammer	0.23
flower	0.12	sympathy	0.14
stone	0.10	flower	0.14
lake	0.09	smile	0.13
book	0.08	person	0.12
graduation	0.08	graduation	0.12

Table 3: Potential object and instrumental arguments of *to kill*

of the model set, using the same vector representation described above. The prediction is that only nouns compatible with the slot will have high similarity to the model set (this is similar to the method of Erk, 2007, but I start from the tuples already stored in the model, rather than harvesting ad-hoc data from the corpus).

Results for the object and "instrumental" (*with*) slots of the verb *to kill* and 10 test items (none of them in the respective model set) are reported in table 3. With the exception of the relatively high position of *fun* as an object, the ranking reflects my semantic compatibility intuitions almost perfectly.

4. Conclusion

Given the impressive growth of corpus-based semantics in recent years, it is time to ask to what extent different tasks that have been tackled with ad-hoc methods can be integrated into a single system. There are good reasons to think that this integration can be straightforward, and that the core format in which corpus-derived information should be stored is as weighted target+type+feature tuples.

5. References

M. Banko, M. Cafarella, S. Soderland, M. Broadhead and O. Etzioni. 2007. Open information extraction from the Web. *Proceedings of IJCAI 2007*: 2670-2676.

M. Baroni, E. Barbu, B. Murphy and M. Poesio. Submitted. StruDEL: A corpus-based semantic model based on properties and types.

K. Erk. 2007. A simple, similarity-based model for selectional preferences. *Proceedings of ACL 2007*: 216-223.

S. Evert 2004. *The statistics of word cooccurrences: Word pairs and collocations*. Ph.D. thesis. University of Stuttgart.

M. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. *Proceedings of COLING 1992*: 539-545.

W. Kintsch. 2001 Predication *Cognitive Science* 25: 173-202.

J. Mitchell and M. Lapata. 2008. Vector-based models of semantic composition. *Proceedings of ACL 2008*: 236-244.

P. Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61: 127-159.

M. Sahlgren. 2006. *The Word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis. Stockholm University.

Rapid Deployment of Speech Processing Systems to New Languages and Domains

Tanja Schultz

University of Karlsruhe and Carnegie Mellon University
tanja@cs.cmu.edu

The performance of speech and language processing technologies has improved dramatically over the past decade, with an increasing number of systems being deployed in a large variety of applications, such as spoken dialog systems, speech summarization and information retrieval systems, and speech translation systems. Most efforts to date were focused on a very small number of languages with large number of speakers, economic potential, and information technology needs of the population. However, speech technology has a lot to contribute even to those languages that do not fall into this category. Languages with a small number of speakers and few linguistic resources may suddenly become of interest for humanitarian and military reasons. Furthermore, a large number of languages are in danger of becoming extinct, and ongoing projects for preserving them could benefit from speech technology.

With more than 6900 languages in the world and the need to support multiple input and output languages, the most important challenge today is to port speech processing systems to new languages rapidly and at reasonable costs. Major bottlenecks are the lack of data and language conventions, and the gap between technology and language expertise. The lack of data results from the fact that today's speech technologies heavily rely on statistically based modeling schemes, such as Hidden Markov Models and n-gram language modeling. Although statistical modeling algorithms are mostly language independent and proved to work well for a variety of languages, the parameter estimation requires vast amounts of training data. Large-scale data resources are currently available for less than 50 languages and the costs for these collections are prohibitive to all but the most widely spoken and economically viable languages. In addition, a surprisingly large number of languages or dialects lack a standardized writing system which hinders web harvesting of large text corpora or the construction of dictionaries and lexicons. Last but not least, despite the well-defined process of system building it is very cost- and time consuming to handle language-specific peculiarities, and it requires substantial language expertise. Unfortunately, it is extremely difficult to find system developers who simultaneously have the necessary technical background and significant insight into the language in question. Consequently, one of the central issues in developing systems in many input and output languages is the challenge of bridging the gap between language and technology expertise. In my talk I will introduce state-of-the-art techniques for rapid language adaptation and present existing solutions to overcome the ever-existing problem of data sparseness and the gap between language and technology expertise. I will describe the building process for speech recognition and speech synthesis components for new unsupported languages and introduce tools to do this rapidly and at lost costs.

The talk describes the SPICE Toolkit (Speech Processing - Interactive Creation and Evaluation), a web based toolkit for rapid language adaptation to new languages. The methods and tools implemented in SPICE enables user to develop speech processing components, to collect appropriate data for building these models, and to evaluate the results allowing for iterative improvements. Building on existing projects like GlobalPhone and FestVox, knowledge and data are shared between recognition and synthesis; this includes phone sets, pronunciation dictionaries, acoustic models, and text resources. SPICE is an online service (http://cmuspice.org). By archiving the data gathered on-the-fly from many cooperative users we hope to significantly increases the repository of languages and resources and make the data and components for new languages available at large to the community. By keeping the users in the developmental loop, SPICE tools can learn from their expertise to constantly adapt and improve. This will hopefully revolutionize the system development process for new languages.

Postopek za izbiro govornih segmentov pri vgrajeni polifonski združevalni sintezi govora

Jerneja Žganec Gros[1], Aleš Mihelič[1], Mario Žganec[1], Uliana Dorofeeva[1], Nikola Pavešić[2]

[1]Alpineon R&D, Iga Grudna 15, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia

Povzetek

V prispevku predstavljamo postopek za izbiro govornih segmentov pri polifonski združevalni sintezi govora, pri katerem smo s poenostavitvami postopkov iskanja poti po grafu vplivali na hitrost postopka za izbiro govornih segmentov tako, da se to čim manj odraža na kvaliteti govora. Izbrani segmenti so še vedno optimalni, le cene lepljenja segmentov, na katerih temelji izbira, so manj natančne. Postopek je primeren za uporabo v vgrajenih sintetizatorjih govora.

An Efficient Unit-Selection Method for Embedded Concatenative Speech Synthesis

Memory and processing power requirements are important factors when designing TTS systems for embedded devices. We describe an accelerated unit-selection methods, which we designed for an embedded implementation of a polyphone concatenative TTS system. The results of objective measurements of computational speed, along with results of subjective listening tests, which have been conceived according to ITU-T recommendations, are provided at the end of the paper.

1. Uvod

Omejitve pri procesnih zmogljivostih in količini pomnilnika, ki jih srečamo pri vgrajenih sistemih, vplivajo tudi na načrtovanje postopka za izbiro govornih segmentov [Lévy04]. Izbira govornih segmentov je tisti del združevalne ali korpusne sinteze govora, pri katerem lahko najbolj vplivamo na hitrost celotnega postopka sinteze govora.

Potrebno je poiskati ugodno razmerje med velikostjo govorne zbirke in računsko zahtevnostjo postopka za izbiro govornih segmentov. Če bo postopek za izbiro govornih segmentov zelo poenostavljen in s tem zelo hiter, bomo lahko v istem času opravili izbiro segmentov pri večji govorni zbirki. Prekomerna poenostavitev postopka pa lahko povzroči izbiro neprimernih govornih segmentov in zato poslabša kvaliteto govora kljub uporabi večje govorne zbirke. Obratno lahko izbira kompleksnega postopka za izbiro govornih segmentov zagotovi optimalno izbiro segmentov, ki pa se zaradi časovnih omejitev lahko izvaja le na majhni govorni zbirki.

V nadaljevanju bomo predstavili postopek za izbiro govornih segmentov, pri katerem smo uspeli vplivati na hitrost postopka za izbiro govornih segmentov, vendar tako, da se to čim manj odraža na kvaliteti govora. To dosežemo tako, da poenostavimo izračun cene lepljenja in s tem ustvarimo pogoje, ki omogočajo specifično zgradbo algoritma za iskanje optimalne poti v grafu. Izbrani segmenti so še vedno optimalni, le cene lepljenja segmentov, na katerih temelji izbira, so manj natančne. Rezultate primerjave hitrosti postopkov iskanja govornih segmentov podajamo v zaključnem poglavju.

2. Izbira govornih segmentov pri polifonski združevalni sintezi govora

Navadno imajo sistemi za polifonsko oz. korpusno združevalno sintezo govora na razpolago obsežne govorne zbirke, ki lahko obsegajo na desetine ur posnetega, segmentiranega ter označenega govora, in zasedajo pomnilniški prostor, ki obsega več gigabajtov. V taki zbirki vsaka od osnovnih govornih enot oz. vsak od govornih segmentov, ki predstavlja določen niz osnovnih govornih segmentov ali polifon, nastopi večkrat, v različnih kontekstih ter z različnimi prozodičnimi lastnostmi. Naloga postopkov za izbiro govornih segmentov je poiskati najustreznejše govorne segmente iz zbirke, tako da bodo zlepljeni tvorili čim bolj kakovosten govorni signal.

Vhodni podatki, ki jih postopek za izbiro govornih segmentov sprejeme od modulov za jezikovno procesiranje v sintetizatorju govora, so zaporedja fonemov, ki jih je potrebno izgovoriti, pri čemer so za vsak fonem podani prozodični parametri za njegovo izgovorjavo. Ti parametri vsebujejo podatke o osnovni frekvenci in trajanju izgovarjave fonema.

Izhodni podatki, ki jih mora postopek za izbiro govornih segmentov posredovati modulu za združevanje segmentov v govorni signal, so zaporedja natanko določenih izsekov govorne zbirke, imenovana polifoni ali govorni segmenti, ki jih bo modul za združevanje segmentov moral združiti. Tudi ta zaporedja so lahko opremljena s prozodičnimi parametri za vsak izsek, s čemer je omogočeno, da modul za združevanje segmentov spremeni izvirne prozodične parametre segmentov iz govorne zbirke, tako da so čim bolj podobni želenim prozodičnim parametrom.

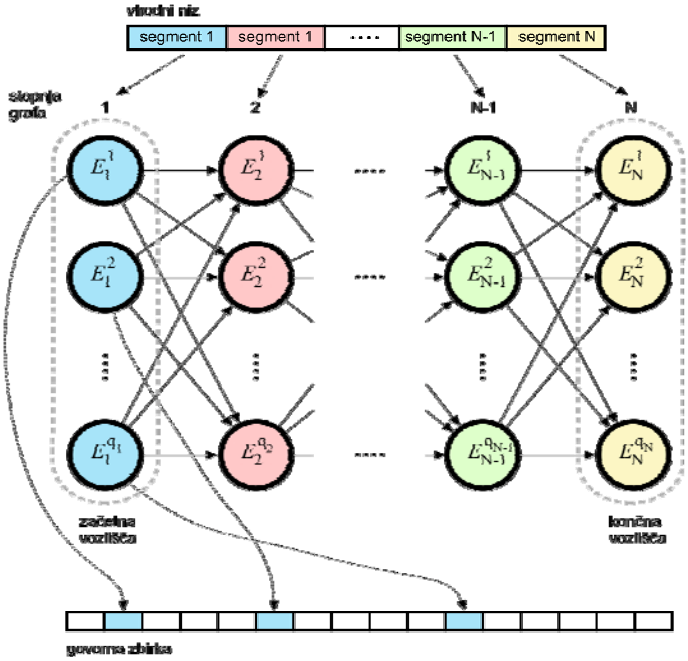
2.1. Tvorjenje grafa za iskanje optimalnega zaporedja govornih segmentov

Problem iskanja optimalnega zaporedja posnetih govornih segmentov za sintezo kakovostnega govornega signala lahko predstavimo kot iskanje optimalne poti v grafu. Takšna predstavitev nazorno prikaže problem izbire govornih segmentov, hkrati pa omogoča uporabo uveljavljenih postopkov za reševanje problema.

Vsako vozlišče grafa predstavlja en osnovni govorni segment iz govorne zbirke. Osnovni govorni segmenti so lahko alofoni, difoni, trifoni ali kakšne druge osnovne enote govora. Graf je razdeljen v posamezne nivoje. Na prvem nivoju grafa se nahajajo izhodiščna vozlišča. To so vsi osnovni govorni segmenti v govorni zbirki, ki

ustrezajo prvemu osnovnemu govornemu segmentu v vhodnem nizu znakov, ki ga je potrebno sintetizirati.

Povezava med vozlišči grafa določa možnost lepljenja osnovnih govornih segmentov, ki jih povezani vozlišči predstavljata. Pri lepljenju govornih segmentov, segment vozlišča, ki je na višji stopnji grafa, časovno sledi segmentu vozlišča na nižji stopnji grafa. Zato so povezave med vozlišči grafa usmerjene. Vozlišča grafa so medsebojno povezana tako, da je vsako vozlišče, ki se nahaja na stopnji n, povezano z vsemi vozlišči, ki se nahajajo na stopnji n+1.



Slika 1. Zgradba grafa za iskanje optimalnega zaporedja govornih segmentov.

V tako zgrajenem grafu lahko iskanje optimalnega zaporedja govornih segmentov definiramo kot iskanje najboljše poti med poljubnim izhodiščnim vozliščem grafa (prva stopnja grafa) in poljubnim končnim vozliščem grafa (zadnja stopnja grafa), pri čemer povezave med vozlišči grafa določajo možne poti.

Da bi lahko pričeli z iskanjem najboljše poti v grafu, moramo definirati kriterije, ki izražajo končni cilj izbire govornih segmentov, kot numerična razmerja med podatki, ki jih graf predstavlja. Končni cilj izbire govornih segmentov sta čim večja razumljivost in naravnost sintetičnega govora. Splošno so se uveljavili naslednji kriteriji, ki prispevajo k čim večji razumljivosti in naravnosti govora:

· čim manjše število lepljenj govornih segmentov,

· čim manjša nezveznost lepljenih segmentov na mestu lepljenja,

· čim boljše prileganje prozodičnih lastnosti lepljenih segmentov želeni prozodiji govora.

Prva dva kriterija ovrednotimo tako, da vsaki povezavi med vozlišči grafa priredimo ceno lepljenja, medtem ko zadnji kriterij ovrednotimo tako, da vsakemu vozlišču grafa priredimo ceno prileganja prozodičnih lastnosti. Cena posamezne poti v grafu je enaka vsoti cen vozlišč, skozi katera pot poteka, kateri je prišteta še vsota cen vseh povezav, ki jih pot vsebuje. Optimalna pot v grafu je pot z najnižjo ceno.

2.2. Cena prileganja prozodičnih lastnosti

Cena prileganja prozodičnih lastnosti izraža podobnost oziroma različnost med prozodičnimi lastnostmi točno določenega govornega segmenta iz govorne zbirke in želenimi prozodičnimi lastnostmi za del govornega signala, ki naj bi ga govorni segment tvoril. Cena prileganja prozodičnih lastnosti je navadno sestavljena iz obteženega rezultata primerjave trajanja govornega segmenta in želenega trajanja govornega segmenta ter iz obteženega rezultata primerjave poteka osnovne frekvence govornega segmenta in želenega poteka osnovne frekvence. Razmerje, v katerem na ceno vplivata trajanje segmenta in potek osnovne frekvence, je večinoma eksperimentalno določeno.

V grafu za iskanje optimalnega zaporedja govornih segmentov je cena prileganja prozodičnih lastnosti prirejena vsakemu vozlišču. Za vsako vozlišče je potrebno izračunati ceno prileganja prozodičnih lastnosti. Kljub temu, da so govorne zbirke lahko zelo obsežne, izračun cene prileganja prozodičnih lastnosti ne predstavlja numerične ovire pri iskanju optimalne poti v grafu.

2.3. Cena lepljenja

Govorni signal tvorimo tako, da združujemo oz. lepimo govorne segmente iz govorne zbirke. Pri postopku združevanja lahko nastanejo slišne nezveznosti v signalu. S ceno lepljenja poskušamo ovrednotiti vpliv nezveznosti v signalu na kvaliteto govora.

Možnih je več pristopov k vrednotenju vpliva lepljenja na kvaliteto govora. Najpreprosteje je, da lepljenju govornih segmentov, ki si v govorni zbirki neposredno sledijo, priredimo ceno "0", med tem ko vsem ostalim kombinacijam govornih segmentov priredimo ceno "1". Uporaba cene "0" pri segmentih, ki si neposredno sledijo v govorni zbirki, je logična, saj so ti segmenti že združeni in lepljenja ni potrebno izvajati. Z uporabo cene "1" pri segmentih, ki si ne sledijo v govorni zbirki, smo enako ovrednotili vsa lepljenja, ne glede na lastnosti segmentov, ki jih lepimo. Pri tako oblikovani ceni lepljenja bi postopek za iskanje optimalnega zaporedja govornih segmentov izbral tisto zaporedje, pri katerem je število lepljenj najmanjše, ne glede na to, katere govorne segmente lepimo.

Boljše ovrednotenje vpliva lepljenja na kvaliteto govora dosežemo, če je cena lepljenja govornih segmentov odvisna od alofonov, pri katerih izvajamo lepljenje. Lepljenju govornih segmentov, ki si v govorni bazi neposredno sledijo, priredimo ceno 0, tako kot pri prejšnjem pristopu.

Najnatančnejše ovrednotenje vpliva lepljenja na kvaliteto govora dosežemo, če pri izračunu cene lepljenja upoštevamo glasoslovne lastnosti obeh lepljenih segmentov. Pri tem lahko upoštevamo razlike v osnovni frekvenci segmentov, razlike v formantnih frekvencah, razlike v glasnosti, razlike v moči šuma, razlike v spektralnih lastnostih šuma, itn. Pri tem se je potrebno zavedati, da uporaba velikega števila parametrov zahteva določitev velikega števila uteži, ki ovrednotijo vpliv razlike v posameznem parametru na ceno lepljenja. Določanje teh uteži je lahko zelo zamudno in pogosto vključuje dolgotrajne preizkuse, empirične rešitve in predpostavke. Velika pomanjkljivost takšnega načina določanja cene lepljenja je numerična zahtevnost. Glede na to, da so cene lepljenja določene individualno za vsak

par osnovnih govornih segmentov iz govorne zbirke, je te cene nemogoče vnaprej izračunati.

Kompromisna rešitev, ki je bistveno hitrejša in kljub temu delno upošteva glasoslovne lastnosti lepljenih govornih segmentov, je vnaprejšnje določanje cene lepljenja za posamezne skupine osnovnih segmentov govorne zbirke. Pri tem pristopu razdelimo osnovne segmente govorne zbirke v skupine na podlagi njihovih glasoslovnih lastnosti, tako da so si govorni segmenti znotraj posamezne skupine glasoslovno čim bolj podobni. To dosežemo z uporabo postopkov za *rojenje vzorcev*. Cene lepljenja izračunamo vnaprej za vse kombinacije skupin vzorcev in jih shranimo.

2.4. Pregled postopkov za izbiro govornih segmentov

Optimalno pot v grafu bi zanesljivo lahko določili tako, da preiščemo vse možne poti v grafu in med njimi izberemo najboljšo. Število možnih poti med poljubnim izhodiščnim vozliščem grafa in poljubnim končnim vozliščem grafa je odvisno od števila nivojev grafa ter od števila pojavov osnovnih govornih segmentov v govorni zbirki.

Če upoštevamo, da se posnetek govornega segmenta v govorni bazi lahko pojavi tudi več tisoč krat in da so vhodni nizi lahko sestavljeni iz več deset osnovnih govornih segmentov, postane jasno, da je število možnih poti v grafu zelo veliko. Zaradi tega ne preiskujemo vseh možnih poti v grafu, temveč uporabljamo različne postopke, ki poenostavijo in pospešijo iskanje. Pri tem nekateri postopki ohranjajo optimalnost rešitve, med tem ko drugi žrtvujejo optimalnost rešitve na račun hitrejšega delovanja.

Optimalno zaporedje govornih segmentov se določa z minimiziranjem cene, ki odraža poslabšanje kvalitete sintetiziranega govora zaradi spektralnih razlik, razlik fonetičnega okolja ter medsebojnega lepljenja govornih segmentov. Sistem, ki je med prvimi uporabljal izbiro govornih segmentov spremenljive dolžine, je ATR-jev v-talk [Sagisaka92]. Poleg uporabe vseh do takrat uporabljanih parametrov je Hirokawa predlagal tudi uporabo prozodičnih razlik pri izbiri optimalnega zaporedja govornih segmentov [Hirokawa90]. Pri tem pristopu sintetiziran govor dobimo z lepljenjem izbranih govornih segmentov, ki jim po potrebi spremenimo prozodične lastnosti. Uporabo informacije o prozodiji, je pri izbiri govornih segmentov predlagal tudi Campbell [Campbell92], [Campbell94].

Postopek za minimizacijo vsote obeh cen uporablja iskanje, ki temelji na dinamičnem programiranju oziroma eni od njegovih izvedenk. Kot osnovna govorna enota za iskanje se navadno uporabljajo osnovni govorni segmenti – fonemi. Obstoječi sistemi za sintezo govora na podlagi lepljenja govornih segmentov iz obsežne govorne zbirke najpogosteje uporabljajo prav ta postopek. Na podlagi omenjenih metod je bil razvit sistem za sintezo govora CHATR [Black94], [Campbell96].

S povečanjem parametrov, ki se uporabljajo pri iskanju oz. izbiri govornih segmentov, se je povečal tudi obseg govorne zbirke. Ob dovolj obsežni govorni zbirki je mogoče iz nje izbrati take govorne segmente, ki so podobni zahtevanim vhodnim prozodičnim parametrom segmentov. V tem primeru pred medsebojnim lepljenjem izbranih govornih segmentov nad njimi ni potrebno izvesti sprememb prozodičnih lastnosti [Campbell97b].

Veliko novejših raziskav se ukvarja z izboljšanjem postopkov iskanja in definicijo parametrov, ki jih upoštevamo pri izračunu cene segmentov [Toda03], [Toda04], [Vepa04]. Modeliranje funkcij za izračun cen je kompleksen problem. Za razvoj zmogljivih funkcij je Campillo predlagal uporabo nevronskih omrežij [Campillo05].

Postopki iskanja si pri izbiri lahko pomagajo tudi z dodatnimi oznakami segmentov spremenljive dolžine, ki posebej označujejo kritične dele, kjer bi lepljenje prineslo potencialna popačenja v končni govorni signal [Breuer04].

Nov pristop k izbiri govornih segmentov predstavlja statistično modeliranje: FSM [Mohri00], GRM [Allauzen04], DCD [Allauzen03], Yi [Yi00], [Yi03a] ter Bulyko in Ostendorf [Bulyko01].

3. Postopek izbire govornih segmentov s poenostavljeno ceno lepljenja

V tem poglavju predlagamo nov, poenostavljen postopek izbire govornih segmentov, ki je zelo hiter in tako primeren za uporabo v implementacijah združevalnega sintetizatorja govora v vgrajenih sistemih.

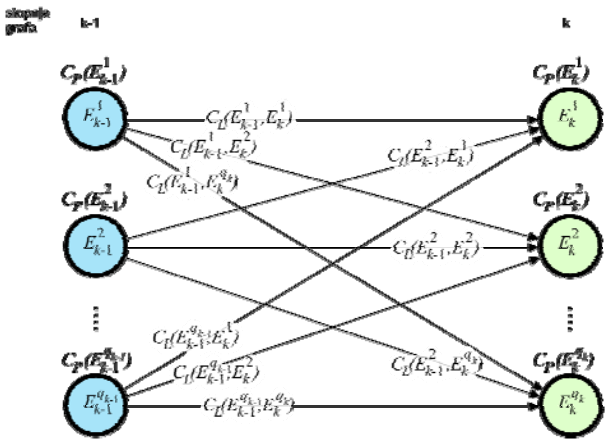
Osnovna poenostavitev pri tem postopku je, de je cena lepljenja dveh govornih segmentov odvisna le od fonemov, ki ju z lepljenjem združujemo. Če lepljenje izvajamo na sredini fonemov, tako kot npr. pri difonski sintezi, določimo za vsak fonem ceno lepljenja na sredini tega fonema.

Če lepljenje izvajamo na mejah med fonemi, moramo določiti ceno lepljenja za vsa zaporedja dveh fonemov, ki lahko nastopijo v govoru. Te cene lepljenja lahko določimo vnaprej in jih med postopkom sinteze ne računamo. Poleg tako določenih cen lepljenja upoštevamo še, da je cena lepljenja enaka 0, če si segmenta, ki ju lepimo neposredno sledita v govorni zbirki, ne glede na fonema, ki sta združena na mestu lepljenja.

Graf, s katerim si pomagamo pri izbiri govornih segmentov, tvorimo tako, kot je bilo opisano v prejšnjem poglavju. Graf vsebuje N stopenj, pri čemer vsaka stopnja ustreza natanko enemu izmed osnovnih govornih segmentov v vhodnem nizu, ki ga moramo sintetizirati. Na k-ti stopnji grafa, ki ustreza govornemu segmentu Sk, se nahaja q_k vozlišč, na k+1 stopnji grafa, ki ustreza govornemu segmentu S_{k+1}, se nahaja q_{k+1} vozlišč, in tako naprej.

Vsako izmed vozlišč na k-ti stopnji grafa E^i_k $(1 \le i \le q_k)$ predstavlja natanko določen posnetek govornega segmenta S_k v govorni zbirki. Za vsako vozlišče E^i_k izračunamo tudi ceno prileganja prozodičnih lastnosti govornega segmenta iz govorne zbirke, ki ga vozlišče predstavlja, in zahtevane prozodije za govorni segment S_k v vhodnem nizu. To ceno označimo z $C_P(E^i_k)$. Vozlišča povežemo tako, da vsako vozlišče E^i_k $(1 \le i \le q_k)$, ki se nahaja na stopnji k, povežemo z vsemi vozlišči E^j_{k-1} $(1 \le j \le q_{k-1})$, ki se nahajajo na stopnji k-1. Cena prehoda med vozliščema E^i_k in E^j_{k-1} je cena lepljenja segmentov govorne zbirke, ki jo vozlišči predstavljata. Označimo jo z $C_L(E^j_k\text{-}1, E^i_k)$.

Pri iskanju optimalne poti v grafu moramo ugotoviti, katera pot med poljubnim izhodiščnim vozliščem grafa E^i_1 $(1 \le i \le q_1)$ in poljubnim končnim vozliščem grafa E^i_N $(1 \le i \le q_N)$ je najcenejša.



Slika 2: Cene lepljenja govornih segmentov so prirejene prehodom med vozlišči grafa. Cene prileganja prozodičnih lastnosti so prirejene vozliščem grafa.

Ceno celotne poti računamo tako, da seštevamo cene lepljenja oz. cene povezav med vozlišči, po katerih se premikamo (C_L), in cene prileganja prozodičnih lastnosti oz. cene vozlišč, ki jih obiščemo (C_P). Na vsaki stopnji grafa k $(1 \le k \le N)$ moramo torej izbrati le eno izmed vozlišč E^i_k $(1 \le i \le q_k)$ oz. le enega izmed govornih segmentov v govorni zbirki, ki ga bomo uporabili pri sintezi govora. To vozlišče označimo z $E^{x(k)}_k$. Ceno optimalne poti v grafu lahko zapišemo kot:

$$C = \min_{x(1),x(2)..x(N)} \left(C_P(E^{x(1)}_1) + \sum_{k=2}^{N} \left(C_P(E^{x(k)}_k) + C_L(E^{x(k)}_k, E^{x(k-1)}_{k-1}) \right) \right).$$

Cena optimalne poti, kot funkcija izbire vozlišča *x(k)* na posamezni stopnji grafa, je razstavljiva funkcija. Če z $C_O(E^i_k)$ označimo ceno optimalne poti od izhodiščnih vozlišč grafa do vozlišča E^i_k na k-ti stopnji grafa in če s C_k označimo ceno optimalne poti od izhodiščnih vozlišč grafa do poljubnega vozlišča na na k-ti stopnji grafa, lahko zapišemo, da je:

$$C_k = \min_{x(k)} \left(C_o(E^{x(k)}_k) \right)$$

in

$$C_O(E^i_k) = C_P(E^i_k) + \min_{x(k-1)} \left(C_L(E^i_k, E^{x(k-1)}_{k-1}) + C_O(E^{x(k-1)}_{k-1}) \right).$$

Vidimo, da funkcijo cene lahko definiramo rekurzivno oz. da je cena poti do vozlišča E^i_k na k-ti stopnji grafa odvisna le od cene prozodičnega prileganja za vozlišče E^i_k in od cen optimalnih poti do vozlišč prejšnje stopnje grafa ($C_O(E^i_{k-1})$), ki jim prištejemo cene lepljenja.

Pri optimizaciji takšne funkcije, lahko za iskanje optimalne poti v grafu uporabimo dinamično programiranje. S to metodo poenostavimo iskanje optimalne poti v grafu tako, da ga razdelimo na iskanje delnih optimalnih poti za vsako stopnjo grafa.

Postopek je v praksi zasnovan tako, da vsakemu vozlišču grafa priredimo 4 parametre. Prvi parameter $I(E^i_k)$ je indeks osnovnega govornega segmenta v govorni zbirki, ki ga vozlišče predstavlja. Ta parameter priredimo vozlišču že na začetku postopka, ob ustvarjanju grafa. Drugi parameter je cena prileganja prozodičnih lastnosti $C_P(E^i_k)$, ki jo ravno tako izračunamo ob ustvarjanju grafa. Tretji parameter je najmanjša kumulativna cena oz. najmanjša cena poti med poljubnim izhodiščnim vozliščem in pričujočim vozliščem $C_O(E^i_k)$. To ceno vpišemo med potekom postopka za izračun optimalne poti. Četrti parameter je indeks vozlišča $P(E^i_k)$ iz prejšnje stopnje grafa, ki leži na optimalni poti med izhodiščnimi vozlišči in pričujočim vozliščem. Tudi ta parameter zapišemo med potekom postopka.

Postopek pričnemo tako, da najmanjši kumulativni ceni izhodiščnih vozlišč priredimo kar ceno prileganja prozodičnih lastnosti istih vozlišč:

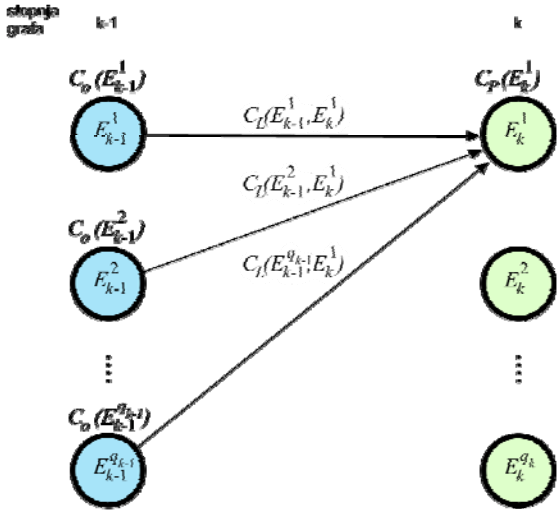
$$C_O(E^i_1) = C_P(E^i_1), \ (1 \le i \le q_1).$$

Pri izhodiščnih vozliščih postavimo kazalce na vozlišče iz prejšnje stopnje grafa na 0, saj izhodiščna vozlišča nimajo predhodnika. Nato določimo najmanjšo ceno poti do posameznih vozlišč na drugi stopnji grafa:

$$C_O(E^i_2) = C_P(E^i_2) + \min_{j=1}^{q_1} \left(C_L(E^i_2, E^j_1) + C_O(E^j_1) \right), \ (1 \le i \le q_k),$$

ter zapišemo tudi indeks (j) vozlišča na predhodni stopnji grafa, ki se nahaja na tej poti z najmanjšo ceno. Postopek sekvenčno ponavljamo za vse ostale stopnje grafa:

$$C_O(E^i_k) = C_P(E^i_k) + \min_{j=1}^{q_{k-1}} \left(C_L(E^i_k, E^j_{k-1}) + C_O(E^j_{k-1}) \right), \ (1 \le i \le q_k, 2 \le k \le N) \quad (1)$$



Slika 3: Cena optimalne poti do vozlišča E^i_k je odvisna od cen optimalnih poti do vozlišč predhodne stopnje grafa $C_O(E^i_{k-1})$, cen lepljenja $C_L(E^j_{k-1}, E^i_k)$ in cene prileganja prozodičnih lastnosti $C_P(E^i_k)$.

Cena optimalne poti je najmanjša izmed cen optimalnih poti do posameznih končnih vozlišč grafa:

$$C = \min_{j=1}^{q_N} \left(C_o(E^j_N) \right),$$

optimalno končno vozlišče pa je končno vozlišče z najmanjšo kumulativno ceno.

Po končanem postopku preberemo zaporedje vozlišč, ki ležijo na optimalni poti tako, da vzvratno sledimo indeksom vozlišč na predhodnih stopnjah grafa $P(E^i_k)$, ki smo jih sproti shranjevali.

Poenostavitev izračuna cene lepljenja, ki smo jo vpeljali pri tem postopku, omogoča, da cene lepljenja določimo vnaprej, tako da je cena lepljenja $C_L(E^j_{k-1}, E^i_k)$ odvisna le od vrste govornih segmentov S_k in S_{k-1}. To pa tudi pomeni, da so vse cene prehodov med vozlišči grafa E^j_{k-1} in E^i_k enake za poljuben j in i. To ne drži le v primeru, ko si govorna segmenta, ki ju predstavljata vozlišči E^j_{k-1} in E^i_k, neposredno sledita v govorni zbirki. V tem primeru je cena lepljenja enaka 0:

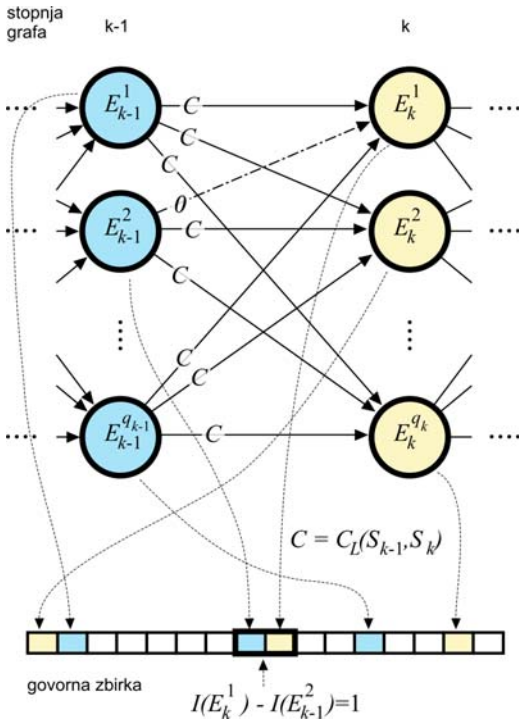
$$C_L(E_{k-1}^i, E_k^j) = \begin{cases} C_L(S_{k-1}, S_k) & ; I(E_k^j) - I(E_{k-1}^i) \neq 1 \\ 0 & ; I(E_k^j) - I(E_{k-1}^i) = 1 \end{cases}.$$

Pri tem $I(E^j_k)$ označuje indeks ali zaporedno mesto govornega segmenta, ki ga predstavlja vozlišče E^j_k v govorni zbirki.

To pomeni, da izračun najmanjše cene poti lahko dodatno poenostavimo. Rekurzivna enačba za izračun najmanjše cene poti od izhodišča do vozlišča E^i_k je zapisana v enačbi (1).

Če upoštevamo omenjene poenostavitve, lahko to enačbo zapišemo kot:

$$C_O(E_k^i) = C_P(E_k^i) + \min_{j=1}^{q_{k-1}} \begin{cases} C_L(S_{k-1}, S_k) + C_O(E_{k-1}^j) & ; I(E_k^j) - I(E_{k-1}^i) \neq 1 \\ C_O(E_{k-1}^j) & ; I(E_k^j) - I(E_{k-1}^i) = 1 \end{cases} \quad (2)$$



Slika 4: Cena lepljenja dveh govornih segmentov, ki si neposredno sledita v govorni zbirki, je enaka 0. Cene lepljenja vseh ostalih segmentov so odvisne le od vrste govornih segmentov, ki ju lepimo, in so zato enake za vse povezave med dvema nivojema grafa.

$C_L(S_{k-1}, S_k)$ je vedno pozitivno število. Zato lahko zgornjo enačbo zapišemo, kot sledi:

$$C_O(E_k^i) = C_P(E_k^i) + \begin{cases} \min\left(C_O(E_{k-1}^J), \min_{j=1}^{q_{k-1}} C_L(S_{k-1}, S_k) + C_O(E_{k-1}^j)\right), \\ \qquad \text{če } \exists J;\ I(E_k^i) - I(E_{k-1}^J) = 1 \qquad (3) \\ \min_{j=1}^{q_{k-1}} \left(C_L(S_{k-1}, S_k) + C_O(E_{k-1}^j)\right), \qquad \text{sicer} \end{cases}$$

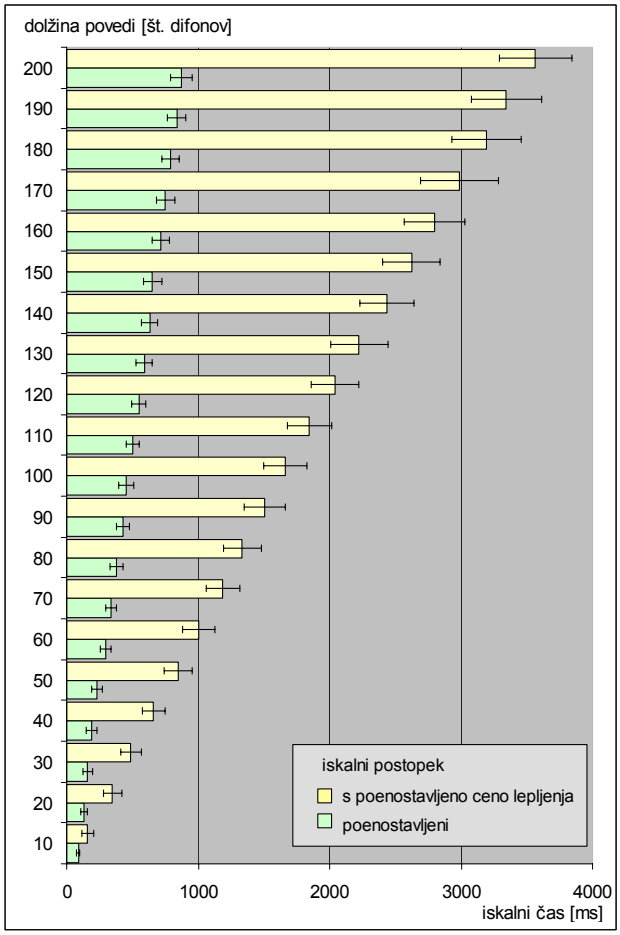
Ker izračun minimuma v zgornji enačbi ni odvisen od *i*, lahko ta izračun naredimo le enkrat za vsa vozlišča, ki pripadajo isti stopnji grafa S_k:

$$C'_O(S_k) = \min_{j=1}^{q_{k-1}} \left(C_L(S_{k-1}, S_k) + C_O(E_{k-1}^j)\right) = C_L(S_{k-1}, S_k) + \min_{j=1}^{q_{k-1}} \left(C_O(E_{k-1}^j)\right).$$

Enačbo (3) lahko sedaj zapišemo kot:

$$C_O(E_k^i) = C_P(E_k^i) + \begin{cases} \min\left(C_O(E_{k-1}^J), C'_O(S_k)\right) & \text{če } \exists J;\ I(E_k^i) - I(E_{k-1}^J) = 1 \\ C'_O(S_k) & \text{sicer} \end{cases}.$$

Ugotovimo, da je ob uporabi opisanega postopka izbire govornih segmentov s poenostavljeno ceno lepljenja potreben le po en izračun minimuma za vsako stopnjo grafa ter eno seštevanje in ena primerjava za vsako vozlišče grafa. Čas, potreben za izračun optimalne poti, narašča približno linearno z naraščanjem velikosti govorne zbirke.



Slika 6. Primerjava hitrosti obeh iskalnih postopkov za izbiro govornih segmentov. Poenostavljeni iskalni postopek je pričakovano hitrejši, saj ne preišče vseh možnih poti v grafu, ampak se omeji samo na najbolj obetavne. Hitrost iskanja narašča z dolžino povedi, za katero mora postopek v govorni zbirki najti ustrezno zaporedje govornih segmentov.

4. Rezultati

Dve različici vgrajene združevalne sinteze govora, ki uporabljata dva različna postopka izbire govornih segmentov, smo primerjali po kakovosti sintetiziranega govora ter po hitrosti delovanja. V prvem primeru smo pri izbiri govornih segmentov uporabili postopek izbire govornih segmentov s poenostavljeno ceno lepljenja, v drugem primeru pa poenostavljeni postopek izbire govornih segmentov [Mihelič06]. Kakovost sintetiziranega govora z uporabo slednjega je rahlo zaostajala za kakovostjo sintetiziranega govora prvega. Iskalni čas pri obeh uporabljenih postopkih izbire

govornih segmentov narašča linearno z dolžino povedi, ki jo je potrebno sintetizirati, in tudi linearno z velikostjo govorne zbirke, kar predstavlja napredek v primerjavi s klasičnimi postopki iskanja poti po grafu, ki se uporabljajo pri izbiri govornih segmentov pri združevalni ali korpusni sintezi govora. Kot je bilo pričakovati, sta oba postopka z daljšanjem povedi, za katere sta v govorni zbirki iskala segmente, potrebne za sintezo, delovala vse počasneje. Poenostavljeni iskalni postopek je hitrejši od postopka iskanja s poenostavljeno ceno lepljenja, ker je manj kompleksen. Segmente za sintezo krajših povedi najde enkrat hitreje, segmente za daljše povedi pa najde več kot štirikrat hitreje, kot to prikazuje slika 6.

5. Zaključek

V prispevku smo predstavili postopek za izbiro govornih segmentov pri polifonski združevalni sintezi govora, pri katerem smo s poenostavitvami postopkov iskanja poti po grafu vplivali na hitrost postopka za izbiro govornih segmentov tako, da se to čim manj odraža na kvaliteti govora. Izbrani segmenti so še vedno optimalni, le cene lepljenja segmentov, na katerih temelji izbira, so manj natančne. Postopek je primeren za uporabo v vgrajenih sintetizatorjih govora.

Postopek za polifonsko združevalno sintezo govora smo preskusili na vgrajeni napravi, ki smo jo razvili v ta namen. Sintetizator slovenskega govora smo vgradili v samodejni sistem za podajanje informacij o donosih medu na čebelarskih oprazovalnicah. Razumljivost in naravnost sintetiziranega govora smo ocenili z obsežnim preskusom, ki je bil pripravljen po mednarodnih priporočilih za preverjanje kakovosti sintetiziranega govora. Splošni vtis sintetizatorja govora je bil ocenjen z oceno 3.2, kar pomeni ‘dobro’, kar se ujema z ocenami splošnega vtisa za sintetizatorje govora za druge jezike, ki navadno prejemajo ocene okoli 3.5 po MOS lestvici. Poslušalci so ocenili, da je sintetični govor razumljiv, primerno razgiban in hiter, ter primeren za uporabo v samodejnih sistemih za podajanje informacij v govorni obliki preko telefona ali interneta.

6. Literatura

Allauzen, C., Mohri, M., Riley, M., (2003). DCD Library - Decoder Library, software collection for decoding and related functions, AT&T Labs - Research.

Allauzen, C., Mohri, M., Roark, B., (2004). A General Weighted Grammar Library, Proceedings of the Ninth International Conference on Automata (CIAA 2004), Kingston, Canada.

Black, A.W., Taylor, P., (1994). CHATR: a generic speech synthesis system, Proceedings of the COLING, Kyoto, Japan, str. 983-986.

Breuer, S., Abresch, J., (2004). Phoxsy: Multi-phone Segments for Unit Selection Speech Synthesis, Institute for Communication Research and Phonetics (IKP) University of Bonn, Proceedings of the Interspeech'04.

Bulyko, I., Ostendorf, M., (2001). Unit Selection for Speech Synthesis Using Splicing Costs with Weighted Finite State Transducers, Proceedings of the EUROSPEECH '01, Aalborg, Danmark. Zv. 2, str. 987–990.

Campbell, W.N., Wightman, C.W., (1992). Prosodic encoding of syntactic structure for speech synthesis. Proceedings of the ICSLP. Banff, Canada. str. 369-372.

Campbell, W.N., (1994). Prosody and the selection of units for concatenation synthesis, Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis. New York, ZDA, str. 61-64.

Campbell, W. N., (1996). CHATR: a high-definition speech resequencing system, Proceedings of the 3rd ASA/ASJ Joint Meeting, Havaji, ZDA, str. 1223-1228.

Campbell, W.N., (1997). Processing a speech corpus for CHATR synthesis, Proceedings of the ICSP, Seul, Koreja, str. 183-186.

Hirokawa, T., Hakoda, K, (1990). Segment selection and pitch modiffcation for high quality speech synthesis using waveform segments, Proceedings of the ICSLP'90, Kobe, Japan, str. 337-340.

Lévy, C., Linares, G., Nocera, P., Bonastre, J. F., (2004). Reducing Computational and Memory Cost for Cellular Phone Embedded Speech Recognition System, Proceedings of the ICASSP '04, Montreal, Canada, Zv. IV, str. 489-492.

Mihelič, A., (2006). Sistem za umetno tvorjenje slovenskega govora, ki temelji na izbiri in združevanju nizov osnovnih govornih enot, Doktorska disertacija v pripravi, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Mohri, M., Pereira, F. C. N., Riley, M., (2000). The Design Principles of a Weighted Finite-State Transducer Library, Theoretical Computer Science, Zv. 231, Št.1, str.17–32.

Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K, (1992). ATR ff-talk speech synthesis system, Proceedings of the ICSLP'92, Banff, Canada, str. 483-486.

Toda, T., (2003). High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion, Doktorska disertacija, Department of Information Processing, Nara Institute of Science and Technology.

Toda T., Kawa, H. , Tsuzak, M., (2004). Optimizing Sub-Cost Functions For Segment Selection Based On Perceptual Evaluations In Concatenative Speech Synthesis, Proceedings of the ICASSP'04, str. 657-660.

Vepa, J., King, S., (2004). Subjective Evaluation Of Join Cost Functions Used In Unit Selection Speech Synthesis, Proceedings of the INTERSPEECH'04, str. 1181-1184.

Yi, J., Glass, J., Hetherington, L., (2000). A flexible scalable finite-state transducer architecture for corpus-based concatenative speech synthesis, Proceedings of the ICSLP'00, Zv. 3. str. 322–325.

Yi, J. R. W., (2003). Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis, Doktorska disertacija, Massachusetts Institute Of Technology.

Nadgradnja sistema za razpoznavanje slovenskega tekočega govora UMB Broadcast News

Andrej Žgank, Marko Kos, Bojan Kotnik, Mirjam Sepesy Maučec, Tomaž Rotovnik, Zdravko Kačič

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor, Slovenija, andrej.zgank@uni-mb.si

Povzetek

V članku bomo predstavili nadgradnjo razpoznavalnika slovenskega tekočega govora za domeno dnevnoinformativnih oddaj. Sistem UMB Broadcast News trenutno predstavlja najkompleksnejši razpoznavalnik slovenskega govora. Zasnovan je na slovenski govorni in tekstovni bazi BNSI Broadcast News. V novo verzijo sistema UMB BN smo vključili več novih, kompleksnih modulov. Glavne spremembe so bile izpeljane na področju akustične segmentacije, izločanja značilk ter akustičnega modeliranja. Za vrednotenje sistema smo uporabili celoten testni nabor baze BNSI, ki vsebuje govorni signal v zelo različnih akustičnih okoljih (7 različnih f-razredov). Uporaba novih metod je uspešno izboljšala delovanje sistema za razpoznavanje slovenskega tekočega govora UMB BN.

Improved version of UMB Broadcast News Slovenian continuous speech recognition system

This paper presents the next version of a Slovenian continuous speech recognition system for the Broadcast News domain. The UMB Broadcast News system is currently the most complex Slovenian speech recognition system. It is built on Slovenian BNSI Broadcast News speech and text database. Several new, complex, modules were incorporated in the new UMB BN system. The major modifications were done in the area of acoustic segmentation, feature extraction and acoustic modeling. The system evaluation was performed using the complete BNSI database evaluation set, which contains spoken material in diverse acoustic conditions. The usage of new methods successfully improved the performance of the UMB BN Slovenian continuous speech recognition system.

1. Uvod

Kljub nenehnemu napredku na področju jezikovnih tehnologij predstavlja razpoznavanje tekočega govora zaradi svoje zelo velike kompleksnosti še vedno velik izziv. To še posebej velja za skupino močno pregibnih jezikov, med katere sodi tudi slovenski. Rezultati različnih raziskav kažejo, da uporaba metod, razvitih za zahodnoevropske jezike, ne zadostuje za doseganje dovolj visoke kvalitete.

Najkompleksnejši sistem za razpoznavanje slovenskega tekočega govora je trenutno UMB Broadcast News (BN), ki je namenjen avtomatskemu razpoznavanju govora v televizijskih dnevnoinformativnih oddajah. Osnovna zgradba sistema in prvi rezultati, doseženi na delu testnega nabora, ki je vključeval samo brani in spontani govor brez zvočnega ozadja, so bili predstavljeni v Žgank et al. (2006).

V članku[1] bomo predstavili nadgradnjo osnovne verzije sistema s kompleksnejšimi moduli ter podali prve rezultate, ki so bili doseženi na celotnem testnem naboru (vseh 7 f-razredov) slovenske govorne baze BNSI Broadcast News (Žgank et al., 2004). V postopek nadgradnje sistema smo vključili izboljšano avtomatsko segmentacijo govor/negovor, dva različna načina izločanja značilk, podporo za različne tipe akustičnih modelov glede na modeliranje konteksta in osnovnih enot, podporo za modeliranje mašil v spontanem govoru, izboljšane jezikovne modele ter razpoznavalnik govora s 64.000 besedami v slovarju.

2. Jezikovni viri

Ena izmed glavnih ovir, ki vplivajo na učinkovitost razvoja različnih sistemov razpoznavanja govora, je obstoj jezikovnih virov za ciljni jezik. Osnovni jezikovni viri, ki jih pri tem potrebujemo, so: govorna baza z ortografskimi transkripcijami za učenje akustičnih modelov, besedilni korpus za izdelavo jezikovnega modela ter slovar besed za razpoznavalnik govora.

2.1. Govorna baza BNSI Broadcast News

Za učenje akustičnih modelov sistema UMB BN uporabljamo govorni korpus slovenske baze BNSI Broadcast News (Žgank et al., 2004), ki obsega 36 ur transkribiranega govornega materiala iz obdobja 1999-2003. V korpus so vključene različne dnevnoinformativne oddaje RTV Slovenija (TV Dnevnik, Odmevi). Govorni posnetki so bili v celoti ročno segmentirani, označeni in transkribirani. 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, kolikor želimo doseči visoko kvaliteto razpoznavanja govora. V primerjavi z osnovno verzijo sistema UMB BN (Žgank et al., 2006) smo tako povečali učni korpus na 18 ur govornega materiala iz vseh f-razredov (Schwartz et al., 1997). Za vrednotenje vpeljanih metod smo tokrat uporabljali celoten testni nabor v skupni dolžini približno 3 ure govora, ki vsebuje posnetke v vseh različnih f-razredih. Za vrednotenje osnovne verzije sistema UMB BN smo namreč uporabljali samo posnetke iz razredov f-0 in f-1 (brani in spontani govor v studijskem okolju, brez zvočnega ozadja) (Žgank et al., 2006).

2.2. Besedilni korpusi

Pri učenju jezikovnih modelov so ključnega pomena uporabljeni besedilni korpusi, iz katerih se jezikovni model uči kontekstne odvisnosti med enotami modeliranja. Uporabili smo tri različne jezikovne vire: BNSI-Speech, BNSI-Text in Večer. BNSI-Speech

[1] Raziskovalno delo je bilo delno sofinancirano s strani ARRS po pogodbi št. J2-9742-0796-06 in pogodbi št. P2-0069.

vključuje transkripcije učnega govornega korpusa, BNSI-text je besedilni korpus baze BNSI in Večer korpus člankov časopisa Večer. Korpusi so si po strukturi zelo različni, zato smo jih uporabljali kot ločene jezikovne vire. Navedimo le nekaj bistvenih razlik. Prva dva vira predstavljata govorjeni jezik, zadnji pa pisni jezik. V korpusu govorjenega jezika označujemo izjave, v korpusih pisanega jezika pa povedi. Izjave so krajše in pogosto pomensko nezaključene, povedi v pisnem jeziku pa daljše in pomensko zaključene. V govoru zasledimo ponovitve, popravljanja, slovnična neujemanja in zelo svoboden besedni red. Teh pojavov v pisnem jeziku skorajda ne najdemo. Besedilni korpusi so tudi po obsegu zelo različni (Žgank et al., 2006), zato smo frekvence pojavljanja in sopojavljanja besed obravnavali ločeno po korpusih. Čeprav bi statistično najbolj zaupali korpusu Večer, ker je največji, sta za nas prva dva korpusa pomembnejša, ker je cilj razpoznavanje govorjenega jezika, in ne pisnega. Razen omenjenih treh korpusov sta bila v eksperimentih uporabljena še BNSI-Devel in BNSI-Eval. Po strukturi sta enaka korpusu BNSI-Speech in obsegata vsak po štiri oddaje. S pomočjo podatkov programske sheme RTV SLO smo zagotovili, da se korpusi po vsebini ne prekrivajo.

2.3. Slovar

Zadnji od potrebnih jezikovnih virov je slovar, ki mora v kar največji meri pokrivati testni nabor. V trenutni verziji sistema UMB BN uporabljamo slovar s 64.000 najpogostejšimi besedami, dobljenimi na osnovi statistike vključenih besedilnih korpusov. Kot poseben vnos v slovarju, ki je v jezikovnem modelu predstavljen le kot kontekst, smo uporabili različna mašila in onomatopeje (Žgank et al., 2008) ter številske vrednosti. Slovar zaenkrat še ne vsebuje različnih variant izgovorjav (narečne oblike) ter besed, ki niso bile izgovorjene do konca (redukcije).

3. Zasnova in opis sistema

Pri nadgradnji sistema UMB BN (Žgank et al., 2006) smo tudi v nadaljevanju uporabljali modularno zasnovo. Pri implementaciji metod smo praviloma upoštevali zahtevo po delovanju v »on-line« načinu. Predstavljena verzija sistema še vedno temelji na samo eni iteraciji razpoznavanja, saj trenutno v sistem še nismo implementirali segmentacije (Grašič et al., 2008) in adaptacije na govorca. V nadaljevanju poglavja bomo predstavili nove pristope, ki smo jih izvedli med zadnjo nadgradnjo sistema UMB BN.

3.1. Segmentacija govor/negovor

Za modeliranje GMM akustičnih razredov govor/negovor smo izbrali dvomodelni pristop. Za tega smo se odločili na osnovi preliminarnih rezultatov, ki so pokazali, da se za segmentacijo govor/negovor v bazi BNSI bolje obnese dvomodelni pristop kot večmodelni. Izkazalo se je, da je glavni razlog za to relativno majhna količina učnega materiala za negovor, kar posledično pomeni, da pride hitro do pretreniranosti modela, še posebej če je njegova kompleksnost večja (uporaba več Gaussovih porazdelitev na model). V našem primeru smo uporabili 512 Gaussovih porazdelitev.

Dvomodelni GMM govor/negovor pristop vsebuje le dva modela: enega za razred govor in enega za razred negovor. Za učenje slednjega je bil uporabljen učni material iz različnih akustičnih razredov, kot so glasba, najavne špice, tišina, šum (npr. hrup prometa, govor v ozadju), navijanje, zvoki okolja itn. Za učenje govornega modela je bil uporabljen čisti govor in govor z ozadjem. V bazi BNSI je čisti govor moč najti v razredih f-0, f-1 in f-2 (čisti telefonski govor), govor z zvočnim ozadjem pa v razredih f-3 in f-4.

Rezultat segmentacije govor/negovor predstavljajo odseki različnih dolžin, nekateri dolgi tudi nekaj 100 sekund. Tako dolgi odseki lahko predstavljajo težavo za dekodirni proces, zato smo pred samim dekodiranjem govora uporabili še modul VAD za zaznavanje govornega signala. Z njim smo razrezali na manjše dele vse govorne segmente, daljše od 30 sekund.

3.2. MFCC in PLP značilke govornega signala

Govorne segmente, katerih začetne in končne meje smo določili s pomočjo akustične segmentacije vhodnega signala, pretvorimo v nizkodimenzionalno zaporedje značilk, ki služijo kot osnovni vhodni parametri algoritma za avtomatsko razpoznavanje govora. V naši raziskavi smo opravili primerjavo dveh najpogosteje uporabljenih postopkov izločanja značilk. To so Mel frekvenčni kepstralni koeficienti (MFCC) (Davis, 1980) in koeficienti perceptivnega linearnega napovedovanja (PLP) (Hermansky, 1990). Nekateri avtorji poročajo, da je s PLP parametrizacijo moč doseči enako ali celo višjo uspešnost razpoznavanja kot z MFCC parametri (Woodland, 2001), zato smo se odločili, da opravimo primerjavo obeh tipov parametrizacij še v okviru sistema razpoznavanja slovenskega govora UMB Broadcast News.

MFCC parametri so bili prvič predstavljeni v zgodnjih osemdesetih letih prejšnjega stoletja (Davis, 1980). Kljub temu pa je ta način parametrizacije govornega signala še vedno uporabljen v številnih komercialno najbolj uspešnih sistemih avtomatskega razpoznavanja govora. V postopku izločanja MFCC značilk najprej izvedemo oknjenje vhodnega govornega signala s Hannovim oknom. Sledi določitev spektra moči posameznega okna na osnovi kratkočasovne Fourierjeve transformacije. Dobljeni spekter moči nato filtriramo (utežimo) z banko 24 polovično prekrivajočih se filtrov trikotne oblike, katerih centralne frekvence so razporejene ekvidistančno glede na melodično (Mel) frekvenčno skalo. Z opisanim postopkom Mel filtriranja poskušamo tako aproksimirati princip kritičnih pasov in amplitudno-frekvenčni odziv človekovega slušnega sistema. Zatem izvedemo logaritemsko komprimiranje izhodov filtrskih bank. V zadnjem koraku sledi izvedba postopka diskretne kosinusne transformacije (DCT), s pomočjo katere izvedemo sočasno dekorelacijo in zmanjšanje dimenzije končnega vektorja statičnih značilk. Končni vektor statičnih značilk po MFCC postopku tako vsebuje 13 elementov, od tega je 12 MFCC koeficientov, trinajsti element pa je podatek o energiji signala v pripadajočem oknu analiziranega vhodnega govornega signala. Kasneje, tik pred izvedbo iskalnega algoritma v postopku avtomatskega razpoznavanja govora, določimo že opisanim statičnim parametrom še njihove odvode prvega in drugega reda. Končno tako v iskalnem algoritmu uporabimo 39 elementov v vektorju dinamičnih značilk.

Postopek PLP parametrizacije (Hermansky, 1990) je bil prvič predstavljen leta 1990, nekoliko kasneje pa je že

bil uspešno uporabljen v postopku razpoznavanja tekočega govora z velikim slovarjem besed za angleški jezik (Woodland, 2001).

V osnovi je postopek določanja PLP značilk kombinacija prej opisanega postopka izločanja MFCC značilk in linearnega napovedovanja. Podobno kot pri MFCC tudi pri PLP postopku izvedemo filtriranje spektra moči s pomočjo Mel filtrske banke. Zatem sledi postopek določitve korelacijskih koeficientov s pomočjo inverzne diskretne Fourierjeve transformacije izhodov filtrskih bank. V skladu z zakonitostjo frekvenčno-amplitudne odvisnosti zaznane jakosti govornega signala pri človeku izvedemo uteževanje predhodno dobljenih korelacijskih koeficientov, ki jih v naslednjem koraku uporabimo pri postopku določitve parametrov linearnega napovedovanja (avtoregresivno modeliranje). Statični vektor PLP značilk sestavlja 13 koeficientov, s čimer dosežemo neposredno primerljivost z MFCC načinom parametrizacije govornega signala. Tudi v tem primeru nadgradimo vektor statičnih PLP značilk s pripadajočimi odvodi prvega in drugega reda.

3.3. Akustično modeliranje

Akustični modeli, ki so vključeni v sistem razpoznavanja govora UMB BN, temeljijo na tristanjskih levo-desnih prikritih modelih Markova s kombinacijami zveznih Gaussovih porazdelitev verjetnosti. Prvotne akustične modele smo zasnovali na grafemski osnovni enoti (Žgank in Kačič, 2005/1), kar pomeni, da je bilo v osnovnem naboru 25 grafemov ter dva dodatna modela za tišino. Z uporabo grafemov smo se izognili dodatni napaki, ki bi jo v sistem vnesla grafemsko-fonemska pretvorba. Eno izmed vmesnih verzij sistema smo zasnovali tudi na fonemskih osnovnih enotah, kjer smo število fonemov spreminjali od 39 do 27, vendar preliminarni eksperimenti niso pokazali statistično pomembnega izboljšanja rezultatov, se je pa bistveno povečala kompleksnost sistema. Za učenje prikritih modelov Markova smo uporabili prosto dostopno orodje HTK (HTK, 2008).

V prvem koraku učenja smo tvorili kontekstno neodvisne akustične modele na osnovi inicializacije z globalnimi vrednostmi. V drugem koraku smo ponovili postopek učenja in pri tem inicializacijo izvedli na osnovi ločenih vrednosti za vsak posamezen model. V vsakem izmed vmesnih korakov smo izvedli postopek prisilne poravnave, s katerim smo izboljšali fonetične transkripcije učnega korpusa.

Tretji korak učenja akustičnih modelov je namenjen gradnji kontekstno odvisnih akustičnih modelov (trigrafemov). Uporabili smo dva različna načina gradnje kontekstno odvisnih akustičnih modelov: notranjebesedne in medbesedne akustične modele (Odell, 1995). Pri prvem načinu se grafemski kontekst upošteva samo znotraj posamezne besede, medtem ko se pri drugem načinu grafemski kontekst razteza tudi preko meje besede. Na takšen način se izboljša modeliranje efekta koartikulacije na besednih mejah, kar pride še posebej do izraza pri modeliranju tekočega govora. Slabost takšnega pristopa je izrazito povečana kompleksnost iskalnega prostora v razpoznavalniku govora.

Uporaba trigrafemskih akustičnih modelov drastično poveča število prostih parametrov, ki jih je potrebno oceniti. Zato smo uporabili postopek vezave stanj z odločitvenim drevesom (Young et al., 1994). S pomočjo tega postopka vežemo stanja, ki so si akustično dovolj podobna med seboj, in tako združimo razpoložljiv učni material. Inicializacijo odločitvenih dreves smo izvedli s podatkovno tvorjenimi grafemskimi razredi (Žgank et al., 2005/2). Število Gaussovih porazdelitev verjetnosti na stanje v kontekstno odvisnih akustičnih modelih smo korakoma povečevali do 16.

3.4. Jezikovno modeliranje

Na osnovi besedilnih korpusov smo zgradili standardni bigramski model. V model smo vključili vse bigrame, tudi tiste, ki so se pojavili samo enkrat. Posledično je nastal relativno velik jezikovni model, ki vsebuje 7.37M bigramov. Jezikovni model je sestavljen iz treh komponent: prvo komponento smo zgradili na korpusu BNSI-Speech, drugo na korpusu BNSI-Text in tretjo na korpusu Večer. Interpolacijske koeficiente komponent smo določili tako, da smo minimizirali perpleksnost jezikovnega modela na BNSI-Devel. Interpolacijski koeficienti po komponentah so bili: 0.26, 0.29 in 0.45. Perpleksnost interpoliranega jezikovnega modela na BNSI-Eval je znašala 410, delež besed izven slovarja (OOV) pa 4.22 %.

3.5. Dekodirnik

Uporabili smo sinhroni iskalni algoritem, ki za zmanjšanje računske zahtevnosti vključuje Viterbijevo aproksimacijo. Za zmanjšanje iskalnega prostora smo sinhroni iskalni algoritem podprli s tehnikami, ki statično in dinamično zmanjšujejo število vozlišč v iskalnem prostoru in bistveno pripomorejo pri pohitritvi razpoznavanja. Uporabili smo sledeče pristope: drevesno oblika slovarja (statično zmanjšanje števila vozlišč iskalnega prostora), snopovno omejevanje (dinamično zmanjšanje aktivnih vozlišč), pogled naprej v jezikovnih modelih (dinamično zmanjšanje aktivnih vozlišč) in omejevanje števila aktivnih hipotez (dinamično zmanjšanje aktivnih vozlišč).

4. Rezultati

Za uspešno izvedbo testiranja je potrebno pripraviti referenčne transkripcije ter orodje za vrednotenje. V našem primeru smo uporabili prosto dostopno orodje Sclite (Sclite, 2008), ki je bilo razvito za potrebe vrednotenj HUB v domeni NIST. Orodje omogoča vrednotenje in časovno poravnavo rezultatov na nivoju besednih mej. Rezultate bomo podajali 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 vrednotenja smo izvedli primerjavo med notranjebesednimi in medbesednimi trigrafemi (tabela 1) z uporabo MFCC značilk.

Modeli	Št. stanj	Razpoznanih (%)	Pravilnih (%)
not.bes.	3868	65,6	63,1
med.bes.	4550	69,0	65,7

Tabela 1: Primerjava med notranjebesednimi in medbesednimi trigrafemskimi akustičnimi modeli.

Oba tipa akustičnih modelov v tem eksperimentu sta uporabljala MFCC značilke. Prva opazna razlika nastopi že pri številu stanj v akustičnem modelu po uporabi združevanja z odločitvenim drevesom. Kljub približno enakemu številu vseh možnih trigrafemov (~17k) imajo medbesedni akustični modeli za dobrih 17 % več stanj. Ta podatek kaže na večjo raznolikost akustičnih modelov. Z uporabo medbesednih akustičnih modelov smo povečali delež pravilnih besed s 63,1 % na 65,7 %. Do izboljšanja rezultatov je verjetno prišlo zaradi boljšega modeliranja koartikulacije na meji med besedami.

V drugem koraku smo primerjali oba postopka izločanja značilk ob uporabi medbesednih trigrafemskih akustičnih modelov. Rezultati so podani v tabeli 2.

Značilke	Št. stanj	Razpoznanih (%)	Pravilnih (%)	xRT
MFCC	4550	69,0	65,7	5,07
PLP	4719	69,6	66,0	4,33

Tabela 2: Primerjava izločanja značilk na osnovi MFCC in PLP.

Primerjava akustičnih modelov, naučenih z MFCC in PLP postopkom, je pokazala, da je število stanj po združevanju z odločitvenim drevesom podobno. Z uporabo PLP metode se je delno izboljšal delež pravilnih besed – s 65,7 % na 66,0 %. Do večje razlike pri rezultatih je prišlo pri hitrosti delovanja. Hitrost delovanja smo merili na računalniku s procesorjem Intel Core2 Quad 2,4 GHz in pri procesiranju vedno uporabljali samo eno jedro. Sistem s PLP značilkami je deloval za dobrih 17 % hitreje kot sistem z MFCC značilkami. Kljub podrobnejši analizi rezultatov in spremljanju delovanja sistema nismo mogli ugotoviti natančnega vzroka za takšno pohitritev.

V zadnjem koraku eksperimentov smo izvedli analizo vpliva nove avtomatske segmentacije govor/negovor na rezultate razpoznavanja govora. Primerjava je predstavljena v tabeli 3.

Segmentacija	Razpoznanih(%)	Pravilnih(%)	xRT
ročna	69,6	66,0	4,33
avtomatska VAD1	68,0	64,2	5,21
avtomatska VAD2	66,7	63,1	5,20

Tabela 3: Analiza vpliva avtomatske segmentacije govor/negovor na razpoznavanje govora.

Primerjali smo vpliv ročne in avtomatske segmentacije testnega nabora na rezultate razpoznavanja govora. Pri avtomatski segmentaciji govor/negovor smo govorne segmente dodatno razrezali na krajše odseke z uporabo modula VAD. Pri tem smo v primeru VAD1 tvorili nekoliko daljše odseke kot v primeru VAD2. Z uporabo avtomatske segmentacije se je delež pravilnih besed zmanjšal s 66,0 % na 64,2 % (VAD1) oziroma 63,1 % (VAD2). Poslabšanje rezultatov z uporabo avtomatske segmentacije je bilo pričakovano, saj takšen postopek vnaša v sistem dodatno napako. Do razlike med daljšimi (VAD1) in krajšimi (VAD2) odseki je verjetno prišlo zato, ker pride v primeru daljših odsekov bolj do izraza učinkovitost modeliranja z jezikovnim modelom.

5. Zaključek

V članku smo predstavili najnovejšo verzijo sistema za razpoznavanje tekočega govora UMB BN. Vpeljali smo različne nove pristope, ki omogočajo učinkovito delovanje sistema. Nekatere izmed analiz preliminarnih rezultatov so že pokazale, da se takšen sistem s samo eno iteracijo razpoznavanja govora bliža fazi »pretreniranosti«, kar pomeni, da bo v prihodnje potrebno v sistem vključiti dodatne iteracije, ki bodo podprle tudi adaptacijo na govorca.

6. Literatura

Davis S. B. et al. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoust., Speech, and Signal Proc.*, vol. 28, pp. 357–366.

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 (2008). 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.

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

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* (v tisku), Trondheim, Norveška.

Vpliv predhodne segmentacije govor/negovor na segmentacijo govorcev

Matej Grašič, Marko Kos, Zdravko Kačič

Inštitut za elektroniko in telekomunikacije,
Fakulteta za elektrotehniko, računalništvo in informatiko,
Univerza v Mariboru
Smetanova 17, 2000 Maribor
matej.grasic@uni-mb.si, marko.kos1@uni-mb.si, kacic@uni-mb.si

Povzetek

Prispevek obravnava vpliv predhodne segmentacije in klasifikacije govor/negovor na segmentacijo govorca oz. na pravilno zaznavo menjave govorca v zvočnem nizu. V članku je predstavljena statistična metoda segmentacije govor/negovor zasnovana na GMM,. Prav tako je predstavljen postopek zaznavne menjave govorca BIC, ki se pogosto uporablja v razmerah, kjer predhodna informacija o govorcu/ih ni na voljo. Uspešnost metod je bila ocenjena v okviru domene BroadCast News, ob uporabi slovenske govorne baze BNSI.

Impact of prior speech/non speech segmentation on speaker segmentation

This paper addresses the impact of speech/non speech pre-segmentation on the performance of speaker turn detection/segmentation. In the article a GMM approach for speech/non segmentation and classification is presented. For the purpose of speaker segmentation e.g. speaker turn detection the BIC segmentation approach was used. The methods were evaluated within the Broadcast News domain, where the Slovenian BNSI database was used.

1. Uvod

Segmentacijo govorcev pogosto uporabljamo na področju, kot je sledenje govorcev, avtomatsko indeksiranje in avtomatska razpoznava govora (ASR – automatic speech recognition). Namen segmentacije govorcev v sistemu avtomatskega razpoznavanja govora je zaznava menjave govorca. To informacijo uporabljamo za prilagoditev akustičnih modelov na določenega govorca v sistemih avtomatskega razpoznavanja govora. Segmentacijo pogosto uporabljamo tudi pri označevanju velikih količin govornega materiala, ko moramo posnetke ustrezno urediti in arhivirati.

Glavni namen segmentacije govorcev je poiskati meje med govorci v zvočnem nizu oziroma posnetku. Na splošno poznamo tri glavne pristope segmentacije govorcev (Zochová, Radová, 2005). Eden prvih postopkov temelji na metričnem pristopu in uporablja merjenje razdalje med dvema sosednjima oknoma v zvočnem nizu. Z metodo lahko tvorimo krivuljo razdalj, na kateri lokalni maksimumi predstavljajo kandidate za meje med različnimi govorci.

V primeru, ko je potrebna večja natančnost in je material za učenje na voljo, običajno uporabljamo principe, ki temeljijo na modelih (npr. Gaussovi modeli - GMM). Vendar pa ta princip ni primeren, ko akustični material vnaprej ni na voljo in ko se učno in testno okolje akustično precej razlikujeta.

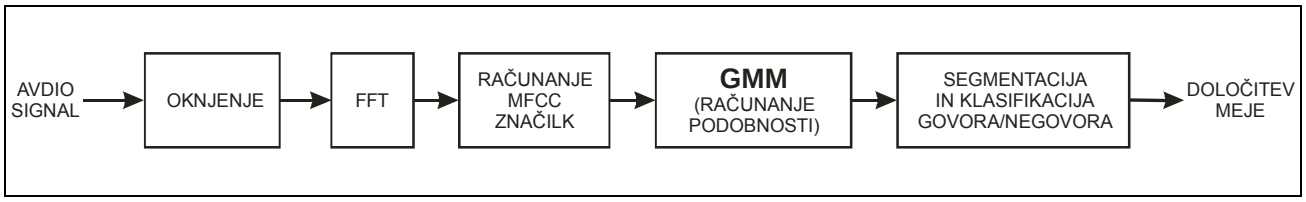
Tretji princip temelji na razpoznavalniku govora, kjer se v prvem koraku izvede razpoznavanje govora, v drugem koraku pa se v premorih skušajo identificirati potencialna mesta za meje med govorci.

Alternativo pristopom, ki temeljijo na modelih, predstavlja BIC (Bayesian Information Criterion). Ta postopek definira segmentacijo kot postopek izbire optimalnega modela. BIC je najpogosteje uporabljen postopek na področju avtomatskega določanja menjave govorca.

Segmentacija zvočnega niza na govorne in negovorne segmente je ena izmed osnovnih oblik akustične segmentacije zvoka. Govorne segmente enostavno definiramo kot segmente s prisotnim govorom, medtem ko so negovorni segmenti segmenti brez prisotnosti govora. Kot del akustične segmentacije je segmentacija govor/negovor navadno prvi od niza korakov akustične segmentacije. Na splošno je akustična segmentacija členitev zvočnega niza na homogene odseke po nekem vnaprej določenem pravilu. Tako lahko predhodno segmentirane govorne dele še naprej segmentiramo glede na spol govorca, zvočno ozadje, pasovno širino posnetka (npr. studio, telefon), itn. Segmentacija zvočnega materiala lahko uporabljamo na področjih, kot je nadzor, povzemanje sestankov in indeksiranje dnevnoinformativnih oddaj (Broadcast News).

Ker vsebujejo oddaje Broadcast News segmente z najrazličnejšimi akustičnimi lastnostmi, je akustična segmentacija takšnih posnetkov precej zahtevna naloga. Zvočni zapis je lahko posnet v studijskem okolju ali pa preko telefonske linije. Vsebuje lahko čist govor, govor z glasbo ali čisto glasbo. V oddajah je veliko število različnih govorcev: poročevalci, voditelji oddaj, novinarji na terenu, gosti, politiki, športniki, znane osebnosti, tuji govorci itn. Pogosto se tudi zgodi, da se isti govorec pojavi v zvočnem zapisu večkrat, vendar v različnih akustičnih pogojih. To je le nekaj razlogov, zakaj je akustična segmentacija oddaj Broadcast News danes še vedno precej zahtevna naloga.

Za akustično segmentacijo so bile do sedaj uporabljene različne metode. Podobno kot pri segmentaciji govorcev tudi tu poznamo metode, ki temeljijo na modelih (Gaussovi modeli, nevronske mreže)



Slika1. Blokovna shema sistema za segmentacijo in klasifikacijo govor/negovor.

(Meignier et. al., 2004; Meinedo, Neto, 2005). Pri nalogah, kjer zvočnega materiala nimamo na voljo vnaprej, uporabljamo nenadzorovane tehnike segmentacije (BIC, cumulative sum – CUSUM) (Yunfeng et. al., 2007; Omar, Chauduri, Ramaswamy, 2005). Kljub pozitivni lastnosti, da ne potrebujeta zvočnega materiala vnaprej, pa sta BIC in CUSUM računsko zelo zahtevni metodi. Za BIC je tudi znano, da ima težave z zaznavanjem kratkočasovnih segmentov. Avtorjem v (Sainath, Kanevsky, Iyengar, 2007) je uspelo s pomočjo razširjenega Baum-Welschevega algoritma zmanjšati računsko zahtevnost in izboljšati zaznavo kratkočasovnih odsekov, ob tem pa ohraniti uspešnost na visokem nivoju.

2. Segmentacija govor/negovor

V članku bomo predstavili vpliv predhodne segmentacije govor/negovor na segmentacijo govorcev za domeno Broadcast News. Pri segmentaciji govorcev tako uporabimo le govorne dele zvočnega zapisa, kar pripomore k boljšemu rezultatu segmentacije. Vrednotenje postopkov je bilo izvedeno s slovensko bazo Broadcast News BNSI.

2.1. Osnovna struktura sistema za segmentacijo govor/negovor

Slika 1 kaže blokovno predstavitev govor/negovor segmentacijskega in kasifikacijskega sistema. Vhodni zvočni signal je vzorčen s frekvenco vzorčenja 16 kHz in kvantiziran s 16-bitno ločljivostjo. Oknjenje je izvedeno s Hannovim oknom dolžine 512 otipkov, kar znaša 32 ms. Korak pomika okna je 10 ms (160 otipkov). Za vsako okno izvedemo Fourierjevo transformacijo dolžine 512. Temu koraku sledi izračun koeficientov MFCC (MEL frequency cepstral coefficients). Vektor značilk sestavlja 12 MFCC značilk, ki so razširjene z normalizirano energijo Tako vektor značilk vsebuje 13 koeficientov. V naslednjem koraku za vsak vektor izračunamo logaritem verjetja za posamezen model GMM. Na osnovi vrednosti logaritmov verjetja in v skladu z vnaprej določenimi pravili (npr. minimalni čas trajanja negovornega segmenta) segmente klasificiramo in določimo meje. Rezultat segmentacije uporabimo v nadaljnjem postopku segmentacije govorcev.

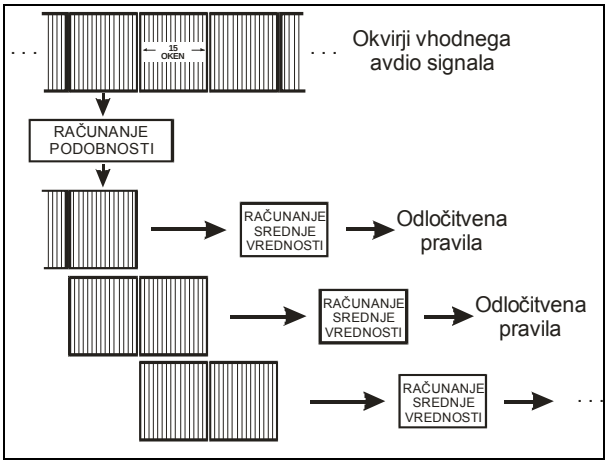
2.2. Modeli GMM in akustični razredi

Za modeliranje akustičnih razredov govor/negovor smo izbrali pristop z dvema modeloma. Za dvomodelni pristop smo se odločili na podlagi ugotovitev in rezultatov, do katerih smo prišli predhodno na tem področju. Rezultati so pokazali, da se za segmentacijo govor/negovor v domeni Broadcast News bolje obnese dvomodelni kot večmodelni pristop. Izkazalo se je, da je glavni razlog za to relativno majhna količina učnega materiala za negovor, kar posledično pomeni, da pride hitro do pretreniranosti modela, še posebej, če je njegova kompleksnost večja (uporaba več Gaussovih porazdelitev na model). V našem primeru smo uporabili 512 Gaussovih porazdelitev.

Dvomodelni GMM je osnovni model govor/negovor, ki vsebuje le dva modela: enega za govorni razred in enega za negovorni razred. Za učenje negovornega razreda smo uporabili učni material iz več akustičnih razredov, kot so glasba, uvodni napevi, tišina, šum (npr. hrup prometa, človeški šum itn.), navijanje, zvoki okolja itn. Za učenje govornega modela smo uporabili čisti govor in govor z ozadjem. V bazi je čisti govor moč najti v F-razredu F0, F1 in F2 (čisti telefonski govor), govor z zvočnim ozadjem pa v razredih F3 in F4.

2.3. Postopek segmentacije in klasifikacije

Po izračunavanju koeficientov MFCC smo za vsako okno izračunali vrednost logaritma verjetja vektorja za posamezen model. Te vrednosti smo nato shranili v medpomnilnik vrednosti verjetnosti in jih nato uporabili za izračun povprečne vrednosti za posamezen model. Izračun povprečne vrednosti smo izvedli vsakih 15 oken, kar predstavlja en okvir. Izračun povprečne vrednosti logaritma verjetja smo izvedli vsaka dva okvirja, torej nad tridesetimi okni (300 ms). Po izračunu povprečne vrednosti logaritma verjetja smo te vrednosti uporabili za določitev meje med segmenti s pomočjo vnaprej določenih pravil in na osnovi najvišje povprečne vrednosti podobnosti. Postopek je prikazan na sliki 2.



Slika 2. Princip okvirjenja vhodnega signala in računanja logaritma verjetja.

Po računanju povprečne vrednosti sledi določanje maksimalne srednje vrednosti, ki določi, ali trenutni okvir spada v razred govor ali v razred negovor. Za zaznavanje kratkih območij tišine si pomagamo z uporabo kratkočasovnega energijskega VAD-a. Meje za govorne in negovorne odseke računamo glede na pravila minimalnega trajanja govornega in negovornega segmenta. Minimalni čas trajanja negovornega segmenta je 1500 ms, saj je to določeno s pravili označevanja za bazo BNSI. Minimalni čas trajanja govornega segmenta pa smo nastavili na 600 ms. Razlog za to je, da se veliko govornih segmentov v bazi začne z uvodnim pozdravom voditelja (600-700 ms), ki mu sledi kratek premor (okoli 300 ms). Na ta način uspešno zaznamo pravi začetek govora.

3. Segmentacija s pomočjo postopka BIC

Bayesov informacijski kriterij uporabljamo kot kriterij za selekcijo oz. izbiro modela, ki v danem primeru najbolje predstavlja izbrane podatke (Tritschler, Gopinath, 1999). Kriterij pogosto uporabljamo v postopkih statističnega modeliranja; kriterij je prvi predstavil Schwarz (1978).

Naj bo nabor podatkov $X = \{x_j \in \Re^d : j = 1,\ldots,N\}$ in $\lambda = \{\lambda_i : i = 1,\ldots, K\}$ nabor kandidatov za parametrični model ter β_i število parametrov v modelu λ_i. Izraz BIC sedaj določimo z:

$$BIC_i = \log P(X \mid \lambda_i) - \alpha \frac{1}{2} \beta_i \log N, \quad (1)$$

kjer je $\log P(X|\lambda_i)$ logaritmična verjetnost učnih podatkov X za model λ_i in α utež drugega člena (Nishida, Kawahara, 2005).

3.1. Postopek segmentacije

Da lahko najdemo menjavo govorca znotraj segmenta, opravimo postopek izbire modela, kjer je model M_1 definiran kot $X = \{x_j \in \Re^d : j = 1,\ldots, N\}$ in določen s samo eno Gaussovo porazdelitvijo, model M_2 definiran kot $X = \{x_j \in \Re^d : j = 1,\ldots, N\}$ in določen z dvema Gaussovima porazdelitvama, kjer je $X_1 = \{x_j \in \Re^d : j = 1,\ldots, i\}$ določen s prvim polnokovariančnim Gaussom in $X_2 = \{x_j \in \Re^d : j = i+1,\ldots, N\}$ določen z drugim polnokovariančnim Gaussom (Tritschler, Gopinath, 1999).

Ker je $x_j \in \Re^d$ in ker je model M_1 definiran z enim in M_2 z dvema Gaussoma, lahko določimo, da ima model M_2 dvakrat več parametrov ($k_2 = 2k_1$):

$$k_1 = d + \frac{d(d+1)}{2}. \quad (2)$$

S pomočjo (1) lahko naredimo test preverjanja modelov, kjer izberemo M_1 pred M_2, če je pogoj $\Delta BIC = BIC_1 - BIC_2$ pozitiven.

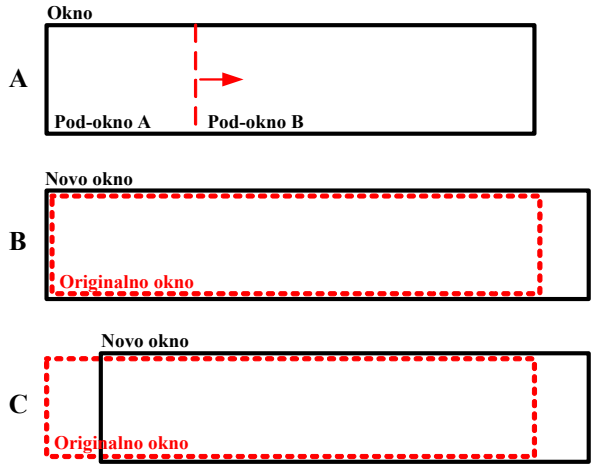
Če uporabimo (3) velja naslednje.

$$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_\omega| + \frac{i}{2}\log|\Sigma_f| + \frac{n-i}{2}\log|\Sigma_s| + \frac{1}{2}\lambda\left(d + \frac{d(d+1)}{2}\right)\log n, \quad (3)$$

V primeru, da je izraz ΔBIC_i negativen, je i-ti okvir primeren kandidat za točko menjave govorca, kjer je $|\Sigma_\omega|$ determinanta kovariančne matrike testiranega okna (model M_1), $|\Sigma_f|$ je determinanta kovariančne matrike prvega podokna A (pripadajoče modelu M2: prvi Gauss) in $|\Sigma_s|$ je determinanta kovariančne matrike drugega podokna B (pripadajoče modelu M2: drugi Gauss).

3.2. Strategija izbire okna

Pravilna izbira okna nad katerim opravimo postopek BIC, je izredno pomembna za doseganje visokega odstotka pravilne razpoznave. Izbiro okna je potrebno opraviti na način, ki preprečuje mešanje zvočnih podatkov različnih govorcev v enem podoknu. Dodatna pozornost je prav tako potrebna pri izbiri podokna, saj majhno število podatkovnih vektorjev znotraj podokna lahko povzroči nepopolno predstavitev govorca znotraj modela. V takšnih primerih se lahko pojavijo nepravilno vrinjene oz. izbrisane meje menjave govorcev.



Slika 1: BIC strategija izbira okna

V obravnavanem sistemu je bila izbrana strategija spremenljivega okna (Ajmera et al., 2004). Ta ima v primerjavi s fiksno izbiro okna prednost, da lahko zajamemo oz. upoštevamo večjo količino podatkov govorca, kar pripomore k izboljšanju detekcije menjave govorca. Strategija izbire podoken je prikazana na sliki 1. Slika 1 A prikazuje izbiro podokna znotraj okna. Pri tem postopka BIC ne opravimo na samih robovih, saj v teh primerih ni na voljo dovolj podatkov o govorcu za izgradnjo modela govorca. Proces, prikazan na sliki 1 A, je potrebno ponoviti po ponovni izbiri glavnega okna. Slika 1 B in slika 1 C prikazujeta postopek izbire okna, kjer velikost okna povečujemo do MAXWINSIZE. Parameter MAXWINSIZE je določen s pomočjo razvojne baze. Če je velikost okna enaka ali pa večja od MAXWINSIZE, se velikost okna ponovno izbere glede na strategijo na sliki 1 C.

4. Testno okolje

4.1. Slovenska govorna baza BNSI Broadcast News

Slovenska baza BNSI Broadcast News je bila posneta v sodelovanju med Fakulteto za elektrotehniko, računalništvo in informatiko Univerze v Mariboru in slovensko nacionalno televizijo RTV Slovenija (Žgank, 2005). Vsebuje dve vrsti TV-oddaj. Prvi tip oddaj so večerne novice,. drugi tip oddaj pa predstavljajo glavni dnevni dogodki, ki so v oddajah podrobneje analizirani. Baza je sestavljena iz 42 informativnih oddaj, kar predstavlja 36 ur govornega materiala. Ta material je nadalje razdeljen na tri dele: učni, razvojni in testni del. Učni del predstavlja 30 ur govornega materiala, medtem ko razvojni in testni del obsegata vsak po 3 ure. Dva najpogostejša razreda v bazi BNSI sta F0 (brani studijski govor, 36,6 %) in F4 (brani ali spontani studijski govor z zvočnim ozadjem, 37,6 %). 16,2 % govora v bazi je spontanega in v studijskem okolju (F1), medtem ko je 6 % govora z glasbo v ozadju (F3). Skupaj baza vsebuje govor 1565 različnih govorcev. Glavnino (1069) predstavljajo moški govorci, medtem ko je žensk 477. Spol preostalih 19 govorcev je označen kot neznan.

4.2 Priprava značilk in nastavitev parametrov

Za samo testiranje in vrednotenje smo uporabili vektor značilk s 13 elementi (12 MFCC + logE – energija). Vektor značilk smo izračunavali vsakih 10 ms znotraj okvirja 32 ms zvoka. Vektorji so bili zapisani v binarno datoteko z dodano oznako govor/negovor. Na ta način smo se izognili iskanju menjav govorca v negovornih področjih. Klasifikacijo govor/negovor smo opravili s pomočjo ročnih referenčnih transkripcij in s pomočjo postopka, predstavljenega v poglavju 3. Za BIC smo uporabili že omenjen variabilni način izbire okna, kjer je bila konstanta MAXWINSIZE nastavljena na 20 s z začetno velikostjo okna 4 s.

5. Rezultati

Pri označevanju menjave govorca se lahko pojavita dva tipa napak: napaka zaznavanja in napaka zgrešitve. Napaka zaznavanja se zgodi, ko je zaznana menjava govorca, čeprav se ta v resnici ni zgodila. Takšen tip napake se običajno meri s tako imenovano mero natančnosti (precision measure). V primeru, da menjava govorca ni bila zaznana, pa govorimo o napaki zgrešitve. Takšen tip napake tipično izrazimo s tako imenovano mero priklica (recall measure). Meri recall in precision sta podani kot:

$$\text{Precision}=\frac{\text{št. pravilno zaznanih prehodov}}{\text{skupno število najdenih prehodov}} \quad (6)$$

$$\text{Recall}=\frac{\text{št. pravilno zaznanih prehodov}}{\text{skupno število pravilnih prehodov}}. \quad (7)$$

Da lahko pravilno vrednotimo uspešnost in da lahko rezultate primerjamo z ostalimi sistemi, uporabimo skupno mero F, definirano kot:

$$F=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (8)$$

Pri vrednotenju za vsako zaznano menjavo govorca uporabljamo interval zaupanja, znotraj katerega je zaznana menjava izbrana kot pravilna oz. nepravilna. Uporabili smo interval zaupanja velikosti 1 s.

5.1. Rezultati vrednotenja

Tabela 1 prikazuje uspešnost zaznave menjave govorca na slovenskem testnem naboru BNSI s postopkom BIC.

Tabela 1. Primerjava uspešnosti zaznave menjave govorca z avtomatsko segmentacijo govor/negovor in brez nje.

Izbrana metoda	*Segmentacija govor/negovor*	*Recall (%)*	*Precision (%)*	*F (%)*
BIC ($\lambda = 3,8$)	iz referenčnih transkripcij	82,80	67,60	74,45
BIC ($\lambda = 3,8$)	avtomatska s pomočjo GMM modela	77,9	65,8	71,34
BIC ($\lambda = 3,8$)	brez	83,90	53,20	65,11

Iz rezultatov je razvidno, da je po pričakovanju postopek BIC z določitvijo področij govor/negovor iz referenčnih transkripcij dosegel najboljši rezultat uspešnosti določitve menjave govorca. V primeru, ko smo za segmentacijo govor/negovor uporabili postopek GMM, je odstotek pravilno razpoznanih menjav nekoliko padel. Razlogi za to so predvsem slabša uspešnost segmentacije v pomembnih trenutkih, kot je menjava govorca. Pri menjavi govorcev se namreč pogosto pojavijo različni negovorni zvoki v ozadju, ki lahko povzročijo nepravilno določitev segmenta kot govor/negovor. Težava se pojavi tudi pri vrednotenju, saj se v bazi pojavljajo segmenti govora/negovora, ki so daljši od 1.5 s, vendar niso označeni, postopek segmentacije in klasifikacije govor/negovor pa jih zazna. Če je torej takšen prehod zaznan pri menjavi govorca, ga postopek BIC ne bo označil.

6. Zaključek

V članku smo predstavili vpliv segmentacije govor/negovor na uspešnost avtomatske zaznave menjave govorca v govorni domeni tipa Broadcast News. Ugotovili smo, da je avtomatska segmentacija govor/negovor prispevala k izboljšanju uspešnosti zaznave menjave v primerjavi z nesegmentiranim postopkom. Nadalje smo ugotovili, da je uspešnost zaznave menjave govorca le nekoliko slabša ob uporabi avtomatske segmentacije govor/negovor v primerjavi s segmentacijo z referenčnimi transkripcijami.

7. Literatura

Ajmera, J., McCowan, I. and Bourlard, H., 2004. Robust Speaker Change Detection, *IEEE Signal Processing Letters*, 649-651, Vol. 11, no. 8

Meignier, S., et al., 2004. Benefits of prior acoustic segmentation for automatic speaker segmentation, *Proc. ICASSP '04*, 397 - 400.

Meinedo, H. and Neto, J., 2005. A Stream-based Audio Segmentation, Classification and Clustering; Pre-processing System for Broadcast News using ANN Models, *Proc. Eurospeech '05*, Lisbon, Portugal.

Nishida, M. and Kawahara, T., Speaker Model Selection Based on the Bayesian Information Criterion Applied to Unsupervised Speaker Indexing, *IEEE Transactions on Speech and Audio Processing*, 583 – 592, vol. 13, no. 4, July 2005

Omar, M., Chauduri, U. and Ramaswamy, G., 2005. Blind Change Detection for Audio Segmentation, *Proc. ICASSP '05*, 501 - 504.

Sainath, T.N., Kanevsky, D. and Iyengar, G., 2007. Unsupervised Audio Segmentation using Extended Baum-Welch Transformations, *Proc. ICASSP 2007*, 209 – 212.

Schwarz, G., 1978. Estimating the deminsion of a model, *The annals of statistics*, 461-464, Vol. 6

Tritschler, A. and Gopinath, R., 1999. Improved speaker segmentation and segments clustering using the Bayesian information criterion, *Eurospeech*, 679–682,

Yunfeng, D., et al., 2007. Audio Segmentation via Tri-Model Bayesian Information Criterion, *Proc. ICASSP '07*, 205 - 208.

Zochová, P. and Radová, V., 2005. Modified DISTBIC Algorithm for Speaker Change Detection, *Interspeech*, 3073-3076

Žgank, A., Verdonik D., Markuš, A. Z. and Kačič, Z., 2005. BNSI Slovenian Broadcast News Database – Speech and Text Corpus, *Proc. Interspeech*, Lisbon, Portugal.

Označevanje vrste diskurznih označevalcev

Darinka Verdonik

Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor
darinka.verdonik@uni-mb.si

Povzetek

Razvoj zahtevnejših jezikovnotehnoloških aplikacij in uporaba korpusov v pragmatičnih in diskurznih raziskavah spodbujata označevanje različnih diskurznih in pragmatičnih pojavov v jezikovnih virih. V tem prispevku obravnavam označevanje diskurznih označevalcev in predlagam ločevanje štirih vrst: ideacijskih označevalcev, interakcijskih označevalcev, označevalcev procesov tvorjenja in interpretacijskih označevalcev. Predstavljena shema je osnova za nadaljnjo korpusno podprto analizo diskurznih označevalcev in za evalvacijo tako označenih virov v zahtevnejših jezikovnotehnoloških aplikacijah.

Annotating discourse marker type

With the demand for more powerful NLP applications and for the use of corpora in pragmatic and discourse studies comes a need for discourse and pragmatic attributes in language resources. In this paper, I focus on the annotation of discourse markers. I propose a classification of discourse markers which consists of four categories, ideational markers, interactional markers, markers of production processes and interpretation markers. The classification is a foundation for further corpus based analysis of discourse markers and for the evaluation of language resources in NLP applications.

1. Uvod

Diskurzna in pragmatična raven postajata vse bolj pogosto predmet označevanja v jezikovnih virih, zlasti ko skušamo razvijati zahtevnejše jezikovnotehnološke aplikacije, kot so razpoznavanje spontanega govora, prevajanje govora, dialog ipd. Najdemo lahko vedno več poskusov, kjer skušajo razviti sheme za označevanje teh ravni v jezikovnih virih (Carlson et al., 2003; Mitkov et al., 2000; Muller et al., 2002; Byron et al., 1997; Heeman, Allen, 1999; Miltsakaki et al., 2002). Čeprav se diskurz in pragmatika uvrščata tudi v pregledne publikacije s področja jezikovnih tehnologij (npr. Mitkov, 2003), pa je na področju diskurznega in pragmatičnega označevanja jezikovnih virov še veliko prostora za nove raziskave in razprave, saj še ni jasno niti, ali bosta diskurz in pragmatika dve različni ravni označevanja ali ena sama oziroma ali in kako naj se standardno označujejo različni diskurzni in pragmatični pojavi, kot so navezovanje, kohezija, diskurzni označevalci, struktura pogovora, govorna dejanja, struktura informacije, retorična razmerja, namernost idr.

V tem prispevku se bom osredotočila na označevanje samo enega od navedenih pojavov, diskurznih označevalcev. V preteklosti je že bilo predstavljenih nekaj shem in poskusov označevanja diskurznih označevalcev (Heeman, Allen, 1999; Miltsakaki et al., 2002; Verdonik et al., 2007), vendar nadaljnje raziskave tovrstnih izrazov, ki so v analizi diskurza zelo aktivne, prinašajo vedno nova spoznanja in vedno širši in raznovrstnejši nabor diskurznih označevalcev (npr. Schiffrin, 1987; Redeker, 1990; Fraser, 1996; Blakemore, 2002; Overstreet, 2005; Swerts, 1998; Clark, Fox Tree, 2002; od slovenskih raziskav npr. Gorjanc, 1998; Smolej, 2004; Pisanski Peterlin, 2005; Schlamberger Brezar, 2007). Kot bomo videli v nadaljevanju, funkcije, ki jih opravljajo posamezni diskurzni označevalci, pokrivajo zelo širok spekter in segajo na različne ravni diskurza, prav tako so izrazi v vlogi diskurznih označevalcev kar se da raznovrstni. Zato se že dalj časa kaže potreba po njihovi nadaljnji razvrstitvi.

V tem prispevku predlagam podkategorijo vrst pri označevanju diskurznih označevalcev v jezikovnih virih in predstavim eno od možnih razvrstitev. V shemah za označevanje jezikovnih virov se skušata zagotavljati enostavnost in nedvoumnost, kar omogoča homogeno, hitro in v kasnejših fazah zadovoljivo uspešno avtomatsko označevanje. Zato predstavljena razdelitev vrst diskurznih označevalcev teži k preprostosti in poenostavitvam, ki temeljijo na najbolj izrazitih skupnih lastnostih. Shema ni mišljena kot natančna jezikoslovna tipologija diskurznih označevalcev, ampak kot osnova za nadaljnjo korpusno podprto analizo diskurznih označevalcev v jezikoslovju, na področju tehnologij pa pomeni osnovo za evalvacijo uporabnosti atributa diskurznih označevalcev.

2. Definicija diskurznih označevalcev in metoda

Obstajajo različne interpretacije diskurznih označevalcev, tj. izrazov, kot so npr. *ja, mhm, aha, no, dobro, v redu, glejte, eee, in, torej* in še številni drugi. Med temeljne pristope lahko štejemo interakcijsko-sociolingvističnega (Schiffrin, 1987; Redeker, 1990), slovnično-pragmatičnega (Fraser, 1996) ter relevantnostnega (Blakemore, 2002). Če jih primerjamo, ugotovimo, da avtorji praviloma ločujejo med dvema temeljnima ravnema diskurza: predstavno, ideacijsko oz. propozicijsko ter proceduralno, pragmatično oz. komunikacijsko. Pri tem so diskurzni označevalci vedno predstavljeni kot izrazi, ki ne prispevajo veliko k vsebini in ne vplivajo bistveno na pomen sporočila, ampak opravljajo predvsem komunikacijske, pragmatične oz. proceduralne funkcije. To sicer zelo ohlapno definicijo lahko štejemo kot skupno izhodišče preučevanja diskurznih označevalcev.

V nadaljevanju predlagam, da tako definirane diskurzne označevalce delimo na štiri vrste: ideacijske, interakcijske, označevalce procesov tvorjenja in interpretacijske označevalce. Delitev skušam utemeljiti na treh ravneh: z označevanjem in analizo avtentičnega gradiva, s sintezo spoznanj tujih raziskav diskurznih označevalcev ter s teoretičnim izhodiščem, po katerem je diskurz zgrajen iz več temeljnih ravni.

V nadaljevanju najprej utemeljujem delitev s teoretičnimi izhodišči in sintezo tistih spoznanj drugih raziskav, ki podpirajo predstavljeno razdelitev, v drugem delu članka pa s predstavitvijo rezultatov označevanja vrst diskurznih označevalcev v korpusnem gradivu.

3. Teoretična izhodišča

Številne raziskave diskurza in jezikovne rabe se osredotočajo na različne ravni diskurza oz. jezikovne rabe, pri čemer lahko ločimo vsaj naslednje: predstavno oz. ideacijsko raven, interakcijsko oz. sociološko raven ter kognitivno oz. mentalno raven.

Kot kažejo lastne predhodne raziskave (npr. Verdonik, 2006; Verdonik et al., 2007), na vseh teh ravneh učinkujejo tudi diskurzni označevalci; običajno učinkujejo na več ravneh hkrati, vendar je, kot ugotavlja že Schiffrin (1987), praviloma ena od ravni bolj poudarjena in zato primarna. Glede na raven diskurza, na kateri primarno učinkujejo, ločim ideacijske, interakcijske in kognitivne diskurzne označevalce, slednje pa naprej na označevalce procesov tvorjenja in interpretacijske označevalce.

3.1. Ideacijski in interakcijski označevalci

Schiffrin (1987) ločuje pet ravni diskurza, na katerih delujejo diskurzni označevalci: menjavanje vlog, struktura dejanj, okvir sodelovanja, predstavna struktura, informativnost. Vendar na koncu ugotavlja, da označevalci s semantičnim pomenom, kot so vezniki in časovni deiksi, delujejo predvsem na ravni predstavne strukture, in tisti brez semantičnega pomena na ostalih ravneh. Takšne ugotovitve vodijo k zaključku, da obstaja večja razlika med označevalci predstavne strukture (npr. *in, pa, torej, in sicer* ...) in vsemi ostalimi označevalci (npr. *mhm, ja, no, dobro, glejte* ...). Tako sklepa tudi Redekerjeva (1990), ki z nadgradnjo modela Schiffrinove (1987) loči označevalce ideacijske strukture in označevalce pragmatične strukture. Tukaj za označevalce pragmatične strukture uporabljam izraz interakcijski označevalci, saj lastne predhodne raziskave (Verdonik et al., 2007; Verdonik et al., v tisku) kažejo, da je njihova osrednja lastnost vzpostavljanje in razvijanje odnosa med sogovorniki, izraz pragmatičen pa je lahko zelo široko in različno razumljen. Označevalci, ki imajo primarno povezovalno vlogo in so usmerjeni predvsem v besedilo diskurza, manj pa v medosebne odnose, so večinoma vezniškega in prislovnega izvora in jih imenujem ideacijski. Poudarjena povezovalna vloga različnih vezniških in prislovnih, pa tudi drugih besed v slovenskem jeziku je bila raziskana npr. že v Gorjanc (1998), vendar nas tukaj v nasprotju z navedeno raziskavo zanima povezovalna vloga teh izrazov samo na ravni diskurza in v skladu z definicijo diskurznih označevalcev.

3.2. Označevalci procesov tvorjenja in interpretacijski označevalci

Pri nekaterih diskurznih označevalcih je bolj kot interakcijska ali predstavna poudarjena kognitivna raven. Sem spadajo večinoma izrazi, ki tradicionalno (npr. v Schiffrin, 1987; Fraser, 1999; Blakemore, 2002) niso obravnavani kot diskurzni označevalci.

Prva podskupina kognitivnih označevalcev so označevalci procesov tvorjenja. Sem spadajo predvsem t. i. zapolnjevalci vrzeli oz. mašila (*eee, mmm* idr.), s katerimi govorec opozarja, da išče ustrezno besedo, se odloča, kaj bo rekel v nadaljevanju, pa tudi skuša ohraniti vlogo oz. prevzeti vlogo (Clark, Fox Tree, 2002; Swerts, 1998; Verdonik, 2007). Prav tako je kazanje na raven tvorjenja primarna funkcija različnih diskurznih izrazov z glagoli rekanja in vedenja (*bom rekel, ne vem* ...). Čeprav tudi označevalci procesov tvorjenja do neke mere delujejo na interakcijski ravni, pa je bolj poudarjena funkcija razkrivanja procesov tvorjenja, ki za interakcijske označevalce ni značilna.

Druga podskupina kognitivnih označevalcev, interpretacijski označevalci, ima primarno vlogo na ravni interpretacije. Gre predvsem za izraze, ki so bili v angleščini raziskovani pod termini general extenders, co-ordination tags, set markers, discourse extenders idr. in se rabijo praviloma na koncu izjave, začenjajo pa se ali z besedico *in* oz. *pa* ali z *ali* (npr. *in tako naprej, pa tako, ali pa kaj*). Overstreetova (2005) definira naslednje osnovne funkcije teh izrazov: signalizirajo predpostavko, da naslovnik ve, kaj ima tvorec v mislih, in da zato nadaljnje tvorjenje v nakazani smeri ni potrebno; spodbujajo naslovnika k solidarnosti, naj se vživi situacijo, ki jo tvorec opisuje; nakazujejo, da bi lahko rekli še veliko več o predmetu pogovora oz. da bi lahko rekli še več, ampak je tisto nepomembno; opozarjajo, da to, kar je bilo rečeno, ni povsem natančno; ublažijo izjave, ki bi lahko prizadele naslovnika; poudarjajo povedano in spodbujajo odgovor. Sklenemo lahko, da je skupna primarna funkcija teh izrazov, da usmerjajo naslovnika pri interpretaciji, zato jih štejem za interpretacijske označevalce.

4. Gradivo

Gradivo vključuje dva pogovorna žanra, ki se razlikujeta v stopnji spontanosti in formalnosti ter prenosniku. Prvi žanr predstavljajo telefonski pogovori med stranko in informatorjem v turistični agenciji, turistični pisarni in hotelski recepciji. Gradivo je izbrano iz korpusa *Turdis* (Verdonik, Rojc, 2006) in poimenovano ***Turdis-2***. Natančnejši podatki so v tabeli 1.

	Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
		minute	besede	minute	besede
Agencija	38	3,40	525	129,23	19936
TIC	12	3,63	529	43,58	6350
Hotel	15	2,78	417	41,68	6261
Skupaj	**65**	**3,30**	**501**	**214,49**	**32547**

Tabela 1: Število in dolžina pogovorov v Turdis-2.

Drugi žanr predstavljajo televizijski intervjuji o aktualnih dogodkih v dnevnoinformativni oddaji, v katerih sodelujejo novinar ter en ali dva intervjuvanca, iz obdobja 1999 do 2005. Gradivo je izbrano iz baze *BNSI Broadcast News* (Žgank et al., 2004) in poimenovano **BNSIint**. Natančnejši podatki so v tabeli 2.

Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
	minute	besede	minute	besede
30	**6,61**	**1041**	**198,35**	**31236**

Tabela 2: Število in dolžina pogovorov v BNSIint.

5. Rezultati korpusne analize

V korpusnem gradivu so bile predstavljene vrste diskurznih označevalcev ročno označene in s tem tudi razdvoumljene v primerih, ko je lahko isti izraz v vlogi diskurznega označevalca ali ne. V nadaljevanju podajam pregled izrazov v vlogi diskurznih označevalcev po vrstah in ločeno za vsak žanr ter podatke o pogostosti rabe v številu konkordanc (št. rab) in v odstotkih glede na število vseh besed v korpusih (% besed).

5.1. Ideacijski označevalci

Ideacijski označevalci so predvsem nekateri priredni vezniki in prislovi. Najpogostejši ideacijski označevalci so bili *in, pa, torej* in *tako da*. Več podatkov je v tabeli 3.

	Turdis-2		BNSIint	
Ideacijski	št. rab	% besed	št. rab	% besed
in	73	0,224	151	0,483
pa	121	0,372	1	0,003
torej	4	0,012	44	0,141
tako da	63	0,387	1	0,006
ampak	9	0,028	3	0,010
in sicer	16	0,098	1	0,006
namreč	2	0,006	16	0,051
potem	60	0,184	0	0,000
pol	13	0,040	0	0,000
sicer	6	0,018	0	0,000
vendar (pa)	0	0,000	4	0,016
SKUPAJ	**367**	**1,370**	**221**	**0,717**

Tabela 3: Ideacijski označevalci v Turdis-2 in BNSIint.

Ideacijski označevalci so v naših korpusih v primerjavi z drugimi vrstami diskurznih označevalcev dokaj redki, saj predstavljajo le okoli 1 % vseh besed. V telefonskih pogovorih v turizmu so nekoliko pogostejši kot v televizijskih intervjujih, vendar je razlika veliko manjša kot pri interakcijskih označevalcih.

5.2. Interakcijski označevalci

Med interakcijske označevalce sodijo večinoma izrazi, ki so tudi tradicionalno obravnavani kot diskurzni označevalci. Z njimi sogovorniki signalizirajo, da se poslušajo in razumejo, da se strinjajo oz. ne strinjajo in se dogovarjajo o poteku diskurza. Najpogostejši interakcijski označevalci v našem gradivu so bili oporni signali, *ja, (a/ali) ne?, no, dobro/v redu/okej/prav* ter *aha*. Podrobnejši podatki so v tabeli 4.

	Turdis-2		BNSIint	
Interpretacijski	št. rab	% besed	št. rab	% besed
ja	603	1,853	68	0,218
aha	234	0,719	0	0,000
aja	14	0,043	0	0,000
mhm	97	0,298	7	0,022
(a/ali) ne?	604	1,902	50	0,179
dobro/v redu/ okej/prav	227	0,977	22	0,074
no	92	0,283	109	0,349
(po)(g)lejte/ (g)lej	79	0,243	48	0,154
(a) veste	27	0,095	6	0,022
zdaj	208	0,639	3	0,010
tako	54	0,166	6	0,019
oporni signali*	1048	3,220	34	0,109
SKUPAJ	**3287**	**10,437**	**353**	**1,156**

* Z opornimi signali udeleženci v pogovoru signalizirajo, da poslušajo, da razumejo ali da se strinjajo z govorcem, ne uvajajo pa (daljše) menjave vlog niti udeleženci z njimi ne nakažejo namena, da prevzamejo vlogo. V gradivu so oporni signali posebej označeni.

Tabela 4: Interakcijski označevalci v Turdis-2 in BNSIint.

Interakcijski označevalci so v našem gradivu najpogostejši diskurzni označevalci. Opazimo pa veliko razliko v pogostosti rabe: v telefonskih pogovorih predstavljajo več kot 10 % vseh besed, v televizijskih intervjujih jih je skoraj 10-krat manj in le nekaj več kot ideacijskih označevalcev. Podrobnejša analiza razlogov za te razlike je bila predstavljena v Verdonik et al. (v tisku).

5.3. Označevalci procesov tvorjenja

Označevalci procesov tvorjenja so predvsem različni izrazi, ki so pogosto imenovani mašila ali zapolnjevalci vrzeli. Podaljšani polglasnik, podaljšani fonem *m* ali *n* ipd. so najpogostejši, poleg teh pa še *mislim*, različni večbesedni izrazi z glagoli rekanja ter z glagolom *vedeti*. Več podatkov je v tabeli 5.

	Turdis-2		BNSIint	
Označevalci tvor.	št. rab	% besed	št. rab	% besed
eee/mmm/ eeem ...	1264	4,047	1293	4,139
Mislim	24	0,074	2	0,006
(kako) bi rekel/-la	0	0,000	15	0,099
bom (jaz) rekel/-la	7	0,046	13	0,083
moram reči/rečt	0	0,000	9	0,064
da/če (tako) rečem	1	0,006	4	0,038
kaj (jaz) vem	6	0,052	1	0,010
ne vem	23	0,141	3	0,019
SKUPAJ	**1325**	**4,203**	**1340**	**4,460**

Tabela 5: Označevalci tvorjenja v Turdis-2 in BNSIint.

Označevalci procesov tvorjenja so edina skupina diskurznih označevalcev, ki so pogosteje rabljeni v televizijskih intervjujih kot v telefonskih pogovorih, vendar je razlika minimalna. Po pogostosti rabe sodijo med bolj pogoste, saj dosegajo skoraj 5 % vseh besed.

5.4. Interpretacijski označevalci

Interpretacijski označevalci so večinoma izrazi, ki se začenjajo ali z besedico *in*, z besedico *pa* (predvsem v manj formalnih konverzacijah), redkeje tudi z besedico *ali*. Podrobneje so predstavljeni v tabeli 6.

	Turdis-2		BNSIint	
Interpretacijski označevalci	št. rab	besed v %	št. rab	besed v %
in tako/tko naprej	3	0,028	24	0,231
in tako dalje	3	0,028	0	0,000
in podobno/-ega	4	0,025	3	0,019
in te/takšne stvari	1	0,009	2	0,013
in to/tega	8	0,049	0	0,000
in vse/vsega	2	0,012	0	0,000

in vse to	0	0,000	1	0,010
pa tako/tko naprej	1	0,009	0	0,000
pa te/take stvari	3	0,028	0	0,000
pa to/tega/temi/teh	5	0,031	0	0,000
pa tako/tko/tak	15	0,092	0	0,000
pa teh zadev	1	0,009	0	0,000
pa vse skupaj	2	0,018	0	0,000
ali (pa) kaj	6	0,046	0	0,000
ali (pa) kaj/kej takega/takšnega/tazga	6	0,068	0	0,000
ali (pa) (ne)kaj podobno/-ega	5	0,052	1	0,010
ali (pa) kakorkoli (že)	1	0,012	1	0,006
SKUPAJ	**66**	**0,516**	**32**	**0,288**

Tabela 6: Interpretacijski označevalci v Turdis-2 in BNSIint.

Interpretacijski označevalci so v našem gradivu najmanj pogosti: v telefonskih pogovorih obsegajo komaj pol odstotka vseh besed, v televizijskih intervjujih pa še enkrat manj. Pri tem moramo upoštevati tudi, da so večinoma večbesedni in že zaradi tega dosežejo še nekoliko večjo pogostost. Tudi tukaj velja, da so v (manj formalnih) telefonskih pogovorih bolj pogosti kot v televizijskih intervjujih.

6. Zaključek

V članku sem predstavila shemo za označevanje vrste diskurznih označevalcev v jezikovnih virih, in sicer ideacijskih označevalcev, interakcijskih označevalcev, označevalcev procesov tvorjenja in interpretacijskih označevalcev. Shema izhaja iz empiričnih analiz in je podprta s teoretičnim ločevanjem temeljnih ravni diskurza. Zaradi potreb označevanja jezikovnih virov je kar se da robustna, zato upošteva le najbolj pogoste in splošno prepoznane značilnosti diskurznih označevalcev in se izogiba razvrščanju istega izraza v več vrst.

V korpusnem gradivu se je pokazalo, da so najpogostejši interakcijski označevalci, zatem označevalci procesov tvorjenja, manj pogosti so bili ideacijski označevalci, najmanj pogosti pa interpretacijski. Vendar so bile velike razlike med obema uporabljenima korpusoma: v splošnem so bili diskurzni označevalci v televizijskih intervjujih veliko manj pogosti (skupaj 4,748 % vseh besed) kot v telefonskih pogovorih v turizmu (skupaj 16,526 % vseh besed).

Predstavljena shema je podlaga za nadaljnjo korpusno podprto analizo diskurznih označevalcev v jezikoslovju, na področju tehnologij pa pomeni osnovo za evalvacijo uporabnosti atributa diskurznih označevalcev.

7. Literatura

Blakemore, D. (2002) Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers. Cambridge: Cambridge University Press.

Byron, D. K., P. A. Heeman (1997). Discourse marker use in task-oriented spoken dialog. 5th European Conference on Speech Communication and Technology (Eurospeech), Rhodes, Greece.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski, 2003. Current Directions in Discourse and Dialogue, chapter Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Kluwer Academic Publishers.

Clark, H.H., Fox Tree, J.E.: Using uh and um in spontaneous speaking. Cognition 84, 73–111 (2002)

Fraser, B. (1996) 'Pragmatic markers', *Pragmatics,* 6/2, 167-190.

Gorjanc, Vojko, 1998: Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija* 46/4. 367−388.

Heeman, Peter, James Allen (1999). Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. Computational Linguistics, 25(4).

H. van den Heuvel, D. Iskra, E. Sanders in F. de Vriend. 2008. Validation of spoken language resources: an overview of basic aspects. *Language Resources and Evaluation,* 42: 41–73.

Miltsakaki, E., R. Prasad, A. Joshi, B. Webber (2002). *The Penn Discourse Treebank. Proceedings of 4th LREC*, Lisbon, Portugal.

Mitkov, Ruslan, 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. In Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000), Lancaster, U.K..

Mitkov, 2003...

Muller, Christoph, Stefan Rapp, and Michael Strube, 2002. Applying co-training to reference resolution. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Philadelpha PA.

Overstreet, M.: And stuff und so: Investigating pragmatic expressions in English and German. Journal of Pragmatics 37, 1845–1864 (2005)

Pisanski Peterlin, A. (2005) 'Text-organising metatext in research articles: An English-Slovene contrastive analysis', *English for Specific Purposes* 25: 307–319.

Redeker, G. (1990) 'Ideational and pragmatic markers of discourse structure', *Journal of Pragmatics* 14: 367-381.

Schiffrin, D. (1987) *Discourse Markers.* Cambridge: Cambridge University Press.

Schlamberger Brezar, M. (2007) 'Vloga povezovalcev v govorjenem diskurzu', *Jezik in slovstvo* 52(3-4): 21-32.

Smolej, M. (2004) 'Členki kot besedilni povezovalci', *Jezik in slovstvo* 49(5): 45-57.

Swerts, M., 1998: Filled pauses as markers of discourse structure. Journal of Pragmatics 30, 485–496.

Verdonik, D., Rojc, M. (2006) 'Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems', in *Proceedings of 5th LREC*, Genoa, Italy.

D. Verdonik, M. Rojc in M. Stabej. 2007a. Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian. *Language Resources and Evaluation,* 41, 147–180.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. 2007b. Diskurzni označevalci v dveh pogovornih žanrih. *Jezik in slovstvo,* 52/6, 19–32.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. V tisku. The impact of context on discourse marker use in two conversational genres. *Discourse Studies.*

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', in *Proceedings of 4th LREC*, Lisbon, Portugal, pp. 2103-2106.

Validacija označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint

Darinka Verdonik[1], Andrej Žgank[1], Agnes Pisanski Peterlin[2]

[1]Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru
Smetanova 17, SI-2000 Maribor
{darinka.verdonik, andrej.zgank}@uni-mb.si
[2]Filozofska fakulteta Univerze v Ljubljani
Aškerčeva 2, SI-1000 Ljubljana
agnes.pisanski@guest.arnes.si

Povzetek

Označevanje diskurznih označevalcev v korpusnem gradivu je lahko včasih odvisno od interpretacije označevalca. Da bi ocenili, koliko so rezultati korpusne analize diskurznih označevalcev odvisni od interpretacije označevalca korpusnega gradiva in natančnost uporabljene sheme za označevanje diskurznih označevalcev v slovenščini, smo izvedli validacijo označenosti reprezentativnega vzorca uporabljenega korpusnega gradiva. Rezultati so pokazali, pri katerih diskurznih označevalcih se pojavlja večja variabilnost označevanja in s katerimi diskurznimi označevalci bi bilo mogoče shemo nadgraditi.

Validating the annotation of discourse markers in Turdis-2 and BNSIint corpora

The annotation of discourse markers in a corpus may sometimes depend on annotator interpretation. To assess to what extent the results of a corpus analysis of discourse marker use depends on annotator interpretation and to evaluate the precision of the annotation scheme used in the annotation of discourse markers in Slovene, a validation of the annotation of a representative sample of the corpus material used was carried out. The results showed which discourse markers show greater variability and which discourse markers could be used to upgrade the annotation scheme.

1. Uvod

Diskurzni označevalci so zadnji dve desetletji v pragmatičnem jezikoslovju zelo aktualna tema (Schiffrin, 1987; Redeker, 1990; Fraser, 1999; Schourup, 1999; Blakemore, 2002; Fox Tree, 2006; idr.; v slovenističnem jezikoslovju pa npr. Gorjanc, 1998; Smolej, 2004; Verdonik, 2006; Pisanski Peterlin, 2005; Schlamberger Brezar, 2007; Verdonik et al., 2007a) in v splošnem pomenijo številne predvsem pragmatične izraze, ki v diskurzu ne prispevajo (pomembno) k vsebini, kot npr. v naslednjem segmentu odgovora informatorke v turistični agenciji stranki (diskurzni označevalci označeni s krepkim tiskom):

zdaj pa *zlo pomembno kar je | tudi pri* ***eee*** *pri* ***eee*** *tejle kar se vstopnici tiče* ***ne?***

Na področju jezikovnih tehnologij najdemo vedno več poskusov označevanja diskurznih označevalcev v jezikovnih virih (Carlson et al., 2003; Mitkov et al., 2000; Muller et al., 2002; Byron et al., 1997; Heeman, Allen, 1999; Miltsakaki et al., 2002). Tovrstne poskuse spodbuja predvsem potreba po dodajanju vse več metajezikovnih podatkov v jezikovne vire, ki izhaja iz razvoja zahtevnejših jezikovnotehnoloških aplikacij, kot so razpoznavanje spontanega govora, strojno prevajanje, prevajanje govora, zahtevnejši sistemi dialoga ipd. Na primeru slovenskega jezika je bila shema za označevanje diskurznih označevalcev predstavljena v Verdonik et al. (2007a), na njeni podlagi pa so bili diskurzni označevalci označeni v dveh govornih korpusih omejenega obsega (vsak okoli 30.000 pojavnic), Turdis-2 in BNSIint (Verdonik et al., 2007b; Verdonik et al., v tisku). Toda označevanje diskurznih označevalcev je v veliki meri odvisno od interpretacije označevalca: isti izrazi lahko namreč opravljajo ali funkcijo diskurznega označevalca ali propozicijske vsebine, vloge pa niso vedno jasno razmejene; poleg tega nabor izrazov v vlogi diskurznih označevalcev v Verdonik et al. (2007a) nikakor ni končen. Zato lahko domnevamo, da bi različne osebe lahko bolj ali manj različno interpretirale in označevale diskurzne označevalce v korpusnem gradivu. Temu se seveda želimo kolikor mogoče izogniti oziroma moramo to upoštevati pri rezultatih korpusne analize.

Ker smo korpusa Turdis-2 in BNSIint uporabljali za jezikoslovne raziskave diskurznih označevalcev in ker želimo pred označevanjem obsežnejšega gradiva zagotoviti kar se da homogeno označevanje, smo validirali označenost navedenih korpusov[1]. Namen validacije je bil dvojen: (1) oceniti, do kolikšne mere so rezultati korpusne analize diskurznih označevalcev, ki smo jih uporabljali v jezikoslovnih raziskavah, odvisni od interpretacije označevalca korpusa; ker smo v raziskavah uporabljali samo skupne kvantitativne podatke (Verdonik et al., 2007b; Verdonik et al., v tisku), nas je tudi pri validaciji zanimala predvsem skupna kvantitativna razlika v številu označenih diskurznih označevalcev; (2) preliminarno oceniti, ali je shema za označevanje diskurznih označevalcev, predstavljena v Verdonik et al. (2007a), dovolj natančna ali pa jo je treba dopolniti in v katerih segmentih.

V nadaljevanju najprej predstavimo oba validirana korpusa, Turdis-2 in BNSIint, nato opišemo validacijski postopek in predstavimo rezultate validacije.

2. Gradivo

Korpusa Turdis-2 in BNSIint zajemata vsak približno 30.000 pojavnic.

Turdis-2 vključuje telefonske pogovore med stranko in informatorjem v turistični agenciji, turistični pisarni

[1] Delo enega izmed soavtorjev je bilo delno sofinancirano s strani ARRS po pogodbi št. J2-9742-0796-06.

oziroma hotelski recepciji. Gradivo je izbrano iz korpusa *Turdis* (Verdonik, Rojc, 2006) tako, da obsega okoli 30.000 pojavnic. Na tak obseg smo se omejili zaradi vzporednih jezikoslovnih raziskav (Verdonik et al., 2007b; Verdonik et al., v tisku), v katere smo zajeli gradivo dveh različnih pogovornih žanrov v primerljivem obsegu. 30.000 pojavnic je dvakrat več gradiva kot v naših predhodnih raziskavah (Verdonik, 2006; Verdonik et al., 2007a). Gradivo je omejeno zaradi časovne zahtevnosti ročnega označevanja korpusov in razpoložljivosti primernih govornih virov.

Gradivo za korpus *Turdis* je bilo posneto na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru spomladi 2004 v sodelovanju z lokalnimi turističnimi organizacijami in njihovimi zaposlenimi. Korpus je bil ročno ortografsko transkribiran. Natančnejši podatki o izboru *Turdis-2* so predstavljeni v tabeli 1.

	Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
		Minute	Pojavnice	Minute	Pojavnice
Turistična agencija	38	3,40	525	129,23	19936
Turistična pisarna	12	3,63	529	43,58	6350
Hotelska recepcija	15	2,78	417	41,68	6261
Skupaj	**65**	**3,30**	**501**	**214,49**	**32547**

Tabela 1: Število in dolžina pogovorov v Turdis-2.

Korpus BNSIint vključuje televizijske intervjuje o aktualnih dogodkih v večerni dnevnoinformativni oddaji nacionalne televizije iz obdobja 1999-2005. V intervjujih sodelujejo novinar ter en ali dva intervjuvanca. Gradivo je izbrano iz baze *BNSI Broadcast News* (Žgank et al., 2004), ki je nastajala v sodelovanju Fakultete za elektrotehniko, računalništvo in informatiko v Mariboru ter RTV Slovenija in vključuje dnevnoinformativne oddaje in informativne pogovorne oddaje, zajete iz arhiva RTV Slovenija. Oddaje so bile ročno ortografsko transkribirane in segmentirane. Natančnejši podatki o korpusu BNSIint so v tabeli 2.

Št. pog.	**Povprečna dolžina**		**Skupna dolžina**	
	Minute	Pojavnice	Minute	Pojavnice
30	**6,61**	**1041**	**198,35**	**31236**

Tabela 2: Število in dolžina intervjujev v BNSIint.

3. Validacija

Preverjanje kakovosti je bistven element gradnje jezikovnih virov. Naš namen je bil validirati kakovost označevanja diskurznih označevalcev v Turdis-2 in BNSIint. V ta namen smo pripravili validacijski korpus; ta je obsegal približno 10 % gradiva iz korpusov Turdis-2 in BNSIint. Podrobnejši podatki o obsegu validacijskega korpusa (število pogovorov, dolžina v minutah, število pojavnic) so predstavljeni v tabelah 3 (za Turdis-2) in 4 (za BNSIint).

	Št. pogovorov		**Dolžina v min.**		**Št. pojavnic**	
	Skupaj	%	Skupaj	%	Skupaj	%
Turistična agencija	4	10,5	13,78	10,7	1965	9,6
Turistična pisarna	2	17,6	4,97	11,4	638	10,1
Hotelska recepcija	2	13,3	3,88	9,3	578	9,2
Skupaj	**8**	**12,3**	**22,63**	**10,6**	**3181**	**9,8**

Tabela 3: Število pogovorov ter dolžina v minutah in številu pojavnic validacijskega korpusa skupno ter v odstotkih od celotnega gradiva Turdis-2.

Št. pogovorov		**Dolžina v min.**		**Št. pojavnic**	
Skupaj	%	Skupaj	%	Skupaj	%
3	**10,0**	**18,62**	**9,39**	**3157**	**10,11**

Tabela 4: Število pogovorov ter dolžina v minutah in številu pojavnic validacijskega korpusa skupno ter v odstotkih od celotnega gradiva BNSIint.

V korpusih Turdis-2 in BNSIint so bili diskurzni označevalci ročno označeni skladno s shemo, predstavljeno v Verdonik et al. (2007a). Validacija je potekala tako, da sta diskurzne označevalce v validacijskem gradivu na novo označila dva zunanja strokovnjaka, ki nista bila povezana s snovanjem sheme za označevanje (Verdonik et al., 2007a) in označevanjem obeh korpusov, ju pa označevanje diskurznih označevalcev v korpusih zanima z različnih uporabnostnih vidikov: prvi validator je bil namreč strokovnjak s področja uporabnega jezikoslovja, drugi s področja govornih tehnologij.

Validacijsko označevanje diskurznih označevalcev je bilo opravljeno skladno s shemo za označevanje, ki je predstavljena v Verdonik et al. (2007a). V tej shemi so diskurzni označevalci definirani kot izrazi, ki k vsebini diskurza ne prispevajo nič ali skoraj nič, pojavljajo pa se v naslednjih pragmatičnih funkcijah:

- vzpostavljanje povezave z vsebino prejšnjega oziroma sledečega diskurza,
- vzpostavljanje in razvijanje odnosa med sogovorniki,
- izražanje odnosa govorca do prejšnje oziroma sledeče vsebine diskurza,
- organiziranje poteka diskurza na ravni prehodov med temami pogovora, menjavanja vlog in strukture izjave.

V Verdonik et al. (2007a) so nadalje najpogostejši tovrstni izrazi tudi našteti in obravnavani, in sicer so to: *ja, mhm, aha, aja, no, eee* v različnih izgovornih variantah (*eeem, eeen, nnn, mmm ...*), *ne?/a ne?/ali ne?/jel?, dobro/v redu/okej/prav, glejte/poglejte, veste/a veste/veste* + vprašalni zaimek (npr. *veste kaj*), *mislim, zdaj* in oporni signali. Slednji zaradi načina transkripcije gradiva niso mogli biti validirani.

	Turdis-2						BNSIint					
	K	V1		V2		E_t%	K	V1		V2		E_b%
DO	***F***	***F1***	***E1 %***	***F2***	***E2 %***		***F***	***F1***	***E1%***	***F2***	***E2 %***	
glejte	5	6	-	5	-	**-**	15	14	+6,7	15	0,0	**±3,4**
ja	66	64	+3,0	70	-6,1	**±4,6**	13	13	0,0	13	0,0	**±0,0**
ne?	65	62	+4,6	65	0,0	**±2,3**	16	11	+31,3	18	-12,5	**±21,9**
dobro idr.	36	36	0,0	36	0,0	**±0,0**	5	5	-	5	-	**-**
mislim	4	2	-	4	-	**-**	1	1	-	3	-	**-**
zdaj	16	15	+6,3	20	-25,0	**±15,7**	1	1	-	1	-	**-**
SKUPAJ	**192**	**185**	**+3,6**	**200**	**-4,2**	**±3,9**	**51**	**45**	**+11,8**	**55**	**-7,8**	**±9,8**

Tabela 5: Rezultati validacijskega označevanja in vrednotenja označenosti gradiva.

Validatorja se o odprtih vprašanjih nista smela posvetovati ne med seboj ne z avtorji sheme za označevanje.

Po končanem validacijskem označevanju smo primerjali označenost validacijskega gradiva v korpusih Turdis-2 in BNSIint z označenostjo validacijskega gradiva pri prvem in drugem validatorju ter na podlagi tega ocenili, pri katerih izrazih je označevanje diskurznih označevalcev najbolj variiralo. Ker smo v raziskavah uporabljali samo skupne kvantitativne podatke (Verdonik et al., 2007b; Verdonik et al., v tisku), nas je tudi pri validaciji zanimala samo skupna kvantitativna razlika v številu označenih diskurznih označevalcev in smo opazovali le skupno razliko v številu označenih diskurznih označevalcev, ne pa tudi razlik pri označevanju posameznih pojavnic. Rezultati so predstavljeni v nadaljevanju.

4. Rezultati

Kot ugotavljajo Verdonik et al. (2007a), so nekateri izrazi vedno v vlogi diskurznega označevalca. Validatorja sta se strinjala, da so takšni izrazi *aha, aja, mhm, no* in *eee/eeem/eeen/nnn/mmm* ipd. Te lahko zato avtomatsko označimo in nadaljnja validacija označevanja ni smiselna.

Rezultati validacije za ostale diskurzne označevalce, obravnavane v Verdonik et al. (2007a), so prikazani v tabeli 5. Na levi strani tabele so podatki za korpus Turdis-2, na desni za korpus BNSIint. V stolpcih K je število diskurznih označevalcev v validiranih korpusih, v stolpcih V1 so podatki za gradivo validatorja 1 ter v stolpcu V2 za gradivo validatorja 2. Za vsako validacijsko gradivo je v stolpcih F1 in F2 navedena pogostost rabe. V stolpcih E1 % in E2 % je ovrednoteno odstopanje po enačbi E1 = (F – F1) / F * 100 oz. E2 = (F – F2) / F * 100. Odstopanje je izračunano tudi skupno za vsak korpus po enačbi E_t = (|E1| + |E2|) / 2 oz. E_b = (|E1| + |E2|) / 2.

Če je bilo v validacijskem gradivu manj kot 10 primerov rabe posameznega diskurznega označevalca, odstopanj v odstotkih nismo računali. To se je dogajalo predvsem pri tistih diskurznih označevalcih, ki tudi v celotnih korpusih Turdis-2 in BNSIint niso bili rabljeni pogosto (manj kot 20-krat sta npr. rabljena *mislim* v obeh korpusih in *zdaj* v BNSIint).

Kot lahko sklepamo na podlagi podatkov v tabeli 5, je najbolj nedvoumno označevanje diskurznih označevalcev *dobro/v redu/okej/prav*. Pri teh v rezultatih ni bilo odstopanj. Odstopanja do 5 %, kar je običajna tolerančna meja v validaciji, zasledimo pri diskurznih označevalcih *ja* in *glejte*. Pomembna odstopanja pa se pojavijo pri označevanju *zdaj* v Turdis-2 (v BNSIint je *zdaj* rabljen le trikrat v celotnem korpusu, zato se tam ne pokažejo pomembnejše razlike) ter pri označevanju *ne?* v BNSIint. Deloma lahko visok odstotek odstopanja pri slednjih pripišemo nizki pogostosti, vseeno pa to kaže, da je označevanju teh dveh diskurznih označevalcev treba posvetiti več pozornosti.

Medtem ko lahko za nabor izrazov, ki so že v Verdonik et al. (2007a) definirani kot diskurzni označevalci z veliko pogostostjo, ugotavljamo v splošnem dokaj homogeno označevanje tako med validatorjema kot v izvornih korpusih, pa so velike razlike pri označevanju ostalih, "novih" izrazov, ki po presoji validatorjev in označevalca korpusa tudi lahko opravljajo vlogo diskurznega označevalca, pa v referenčni raziskavi (Verdonik et al., 2007a) niso bili obravnavani. Kateri od teh izrazov so bili označeni v validacijskih korpusih in kolikokrat, prikazuje tabela 6. Podatki v stolpcu K so za validirana korpusa, v stolpcih V1 za validacijski korpus prvega validatorja in stolpcih V2 za validacijski korpus drugega validatorja.

DO	Turdis-2			BNSIint		
Medmeti	**K**	**V1**	**V2**	**K**	**V1**	**V2**
hm	1	1	1	1	1	1
ma	1	1	1	1	1	0
a	1	0	2	0	0	0
he	0	0	0	1	1	1
vvv	0	0	0	0	0	1
evo	1	0	0	0	0	0
da	0	0	0	2	0	0
SKUPAJ	**4**	**2**	**4**	**5**	**3**	**3**
Ostalo						
in	0	0	0	3	0	26
torej	0	0	0	0	6	12
pa	0	0	12	0	0	0
namreč	0	0	0	0	0	7
pač	1	0	3	0	0	4
pol	0	4	3	0	0	0
pol pa	0	1	0	0	0	0
potem	0	3	0	0	0	0
seveda	0	0	0	0	1	5
saj	0	2	1	0	0	0
ampak	0	0	1	0	0	1
odlično	0	0	1	0	0	0
važi	0	1	1	0	0	0
v bistvu	0	1	1	0	0	0
tako	0	1	0	0	0	0
vem	0	1	0	0	0	0
da	0	0	0	0	0	2
naj	0	0	0	0	0	1
skratka	0	0	0	0	0	1
bom rekel	0	0	0	0	0	1

moram reči	0	0	0	0	0	1
SKUPAJ	**1**	**14**	**23**	**3**	**7**	**61**

Tabela 6: Novi izrazi v vlogi diskurznega označevalca.

Iz tabele 6 vidimo, da niti med obema validatorjema ni skupnega mnenja, kateri "novi" izrazi ustrezajo definiciji diskurznih označevalcev in kako pogosto so v tej vlogi. Zlasti velike razlike so pri izrazih *in, torej, pa, namreč* ipd., katerih konektorska vloga je sicer že bila prepoznana v domači (npr. Gorjanc, 1998) literaturi, za tuje tem sorodne izraze (npr. v angl. *and, so*) pa je prepoznana tudi pragmatična vloga (Schiffrin, 1987). Razlike so nedvomno v veliki meri posledica tega, da validatorja nista imela na voljo vira, ki bi podrobneje obravnaval pragmatične vloge teh izrazov, brez dvoma pa lahko sklenemo, da kaže razširiti raziskavo diskurznih označevalcev na nekatere priredne veznike (*in, torej, pa, namreč* ...), členke (*pač, seveda* ...) in prislove (*pol, potem* ...). Prav tako so v tej vlogi očitno nekateri, sicer redko rabljeni medmeti (v našem primeru *hm, ma*).

5. Zaključek

V prispevku smo predstavili validacijo označevanja diskurznih označevalcev v korpusih Turdis-2 in BNSIint ter skušali oceniti, do kolikšne mere so rezultati korpusne analize diskurznih označevalcev, ki smo jih uporabljali v jezikoslovnih raziskavah, nevtralni, ter preliminarno oceniti, ali je shema za označevanje diskurznih označevalcev, predstavljena v (Verdonik et al., 2007a), dovolj natančna oziroma v katerih segmentih jo je treba dopolniti.

Ugotovili smo, da je v korpusu Turdis-2 večja variabilnost pri označevanju diskurznega označevalca *zdaj* ter v korpusu BNSIint pri označevanju *ne?*. Za ostale diskurzne označevalce ugotavljamo, da so večinoma homogeno označeni.

Pri nadaljnjem razvoju označevanja diskurznih označevalcev v jezikovnih virih bi tako kazalo dodatno pozornost nameniti predvsem *zdaj* in *ne?*, pa tudi diskurznima označevalcema *ja* in *glejte*, pri katerih zaznamo variabilnost označevanja do 5 %. Pri vseh teh bi bilo tudi smiselno dopolniti validacijsko analizo s primerjanjem položajev, v katerih so bili označeni kot diskurzni označevalci.

Največ pozornosti pa je treba posvetiti izrazom, ki v shemi, na kateri sta temeljila označevanje in validacija (Verdonik et al., 2007a), niso eksplicitno obravnavani. Kot kažejo rezultati validacije, so to predvsem nekateri (priredni) vezniki, pa tudi členki, prislovi in medmeti. V tej smeri bi bile potrebne dodatne natančne raziskave pragmatičnih vlog teh izrazov, na katerih bi temeljila nadaljnja nadgradnja sheme za označevanje diskurznih označevalcev.

6. Literatura

D. Blakemore. 2002. *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers.* Cambridge: Cambridge University Press.

D. K. Byron, P. A. Heeman. 1997. Discourse marker use in task-oriented spoken dialog. *5th European Conference on Speech Communication and Technology (Eurospeech)*, Rodos, Grčija.

L. Carlson, D. Marcu, in M. E. Okurowski. 2003. *Current Directions in Discourse and Dialogue*, poglavje Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. Kluwer Academic Publishers.

J. E. Fox Tree. 2006. Placing *like* in telling stories. *Discourse Studies* 8(6): 723-743.

B. Fraser. 1996. Pragmatic markers. *Pragmatics,* 6/2, 167-190.

B. Fraser. 1999. What are discourse markers? *Journal of Pragmatics* 31: 931-952.

V. Gorjanc. 1998. Konektorji v slovničnem opisu znanstvenega besedila. *Slavistična revija* 46/4. 367–388.

P. Heeman, J. Allen. 1999. Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog. *Computational Linguistics,* 25(4).

E. Miltsakaki, R. Prasad, A. Joshi in B. Webber. 2002. The Penn Discourse Treebank. *Language Resources and Evaluation Conference'04*, Lizbona, Portugalska.

R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones, in V. Sotirova. 2000. Coreference and anaphora: developing annotating tools, annotated resources and annotation strategies. *Proc. of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2000)*, Lancaster, Vel. Britanija.

C. Müller, St. Rapp, in M. Strube. 2002. Applying co-training to reference resolution. *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Philadelpha, ZDA.

A. Pisanski Peterlin. 2005. Text-organising metatext in research articles: An English-Slovene contrastive analysis. *English for Specific Purposes* 25: 307–319.

G. Redeker. 1990. Ideational and pragmatic markers of discourse structure. *Journal of Pragmatics* 14: 367-381.

D. Schiffrin. 1987. *Discourse Markers.* Cambridge: Cambridge University Press.

M. Schlamberger Brezar. 2007. Vloga povezovalcev v govorjenem diskurzu. *Jezik in slovstvo* 52(3-4): 21-32.

L. Schourup. 1999. Discourse markers. *Lingua* 107: 227-265.

M. Smolej. 2004. Členki kot besedilni povezovalci. *Jezik in slovstvo* 49(5): 45-57.

D. Verdonik, M. Rojc. 2006. Are you ready for a call? – Spontaneous conversations in tourism for speech-to-speech translation systems. *Proc. of 5th LREC*, Genova, Italija.

D. Verdonik, M. Rojc in M. Stabej. 2007a. Annotating discourse markers in spontaneous speech corpora on an example for the Slovenian. *Language Resources and Evaluation,* 41: 147–180.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. 2007b. Diskurzni označevalci v dveh pogovornih žanrih. *Jezik in slovstvo,* 52/6, 19–32.

D. Verdonik, A. Žgank in A. Pisanski Peterlin. V tisku. The impact of context on discourse marker use in two conversational genres. *Discourse Studies.*

A. Žgank, T. Rotovnik, M. Sepesy Maučec, D. Verdonik, J. Kitak, D. Vlaj, V. Hozjan, Z. Kačič, B. Horvat. 2004. Acquisition and annotation of Slovenian Broadcast News database. *Proc. of 4th LREC*, Lizbona, Portugalska.

Črpanje primerov za japonsko-slovenski slovar iz vzporednega korpusa

Kristina Hmeljak Sangawa,[*] Tomaž Erjavec[†]

* Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
kristina.hmeljak@guest.arnes.si
† Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Na FFUL in IJS med urejanjem spletnega japonsko-slovenskega slovarja za študente japonščine, ki nastaja ob spremljanju potreb porabnikov in njihovem sodelovanju, skušamo z uporabo jezikovnih tehnologij slovar čimbolj učinkovito dopolnjevati in dograjevati. Iz vzporednega korpusa, ki nastaja v okviru vaj iz prevajanja med slovenščino in japonščino, ter vzporednih besedil, ki smo jih s pomočjo iskalnikov in regularnih izrazov črpali s spleta, smo zbrali stavčne primere k vsem slovarskim geslom, ki jih vzporedni korpus vsebuje. Pri pogostih besedah, ki so se v korpusu največkrat pojavljale, smo število primerov omejili in v korpusu zaradi lažjega branja izbrali najkrajše povedi z izbranimi gesli. Projekt lahko služi kot model za učinkovito ustvarjanje referenčnega učnega gradiva z uporabo prosto dostopnih orodij in besedil.

Extracting examples from a parallel corpus for a Japanese-Slovene dictionary

A Japanese-Slovene learners' dictionary is being produced in cooperation between the Faculty of Arts of the University of Ljubljana and the Jožef Stefan Institute. The process of building the dictionary relies on users' collaboration and feedback and on using language technologies and resources. The paper reports on the augmentation of the dictionary with example sentences, automatically extracted from a Japanese-Slovene parallel corpus that was built for this purpose. The corpus was compiled using Japanese-Slovene text produced at the Faculty of Arts as part of student course-work and parallel texts collected from the Web. We present the compilation and annotation of the parallel corpus, the method of selecting examples to be included in the dictionary, and give an informal evaluation of the results. The methodology presented can serve as a model for low-cost production of lexicographic material.

1. Uvod

Na Oddelku za azijske in afriške študije Filozofske fakultete Univerze v Ljubljani že od nastanka oddelka postopoma razvijamo japonsko-slovenski slovar za slovensko govoreče študente (Hmeljak 2001). Slovar je bil leta 2003 predelan v obliko XML po priporočilih TEI (Erjavec idr. 2003) in postavljen na strežnik Instituta Jožef Stefan, kjer je od takrat prosto dostopen (http://nl.ijs.si/jaslo/). Število uporabnikov je tudi za slovenske razmere majhno, zato je projekt izrazito netržen. Ob sodelovanju študentov japonologije, ki so glavni uporabniki slovarja, gradivo postopoma nadgrajujemo in objavljamo na spletu. Na osnovi analiz iskalnih nizov, ki jih beležimo na spletnem iskalniku, slovar postopoma dopolnjujemo, pri tem pa skušamo čim bolj učinkovito izkoriščati obstoječe vire, metode in smernice. Tako smo na primer po ugotovitvi, da uporabniki velikokrat iščejo japonske besede in izraze v latinici, s pomočjo obstoječega orodja dodali latinizirano obliko vsem slovarskim geslom, v isti nadgradnji pa dodali tudi podatke o težavnostni stopnji besedišča po seznamu besedišča za osrednji izpit iz znanja japonščine kot tujega jezika JLPT (Erjavec idr. 2006). Slovarsko gradivo je bilo tudi vključeno v orodje za podporo branju japonskih besedil Reading Tutor (Kawamura 2005). Slovar je trenutno zastavljen kot enosmerni japonsko-slovenski slovar, a omogoča tudi iskanje po slovenskih prevodnih ustreznicah in primerih. Iz analize iskalnih nizov, ki jih je zabeležil strežnik, smo ugotovili, da so uporabniki v več kot polovici primerov iskali slovenske besede, kar nakazuje potrebo po slovensko-japonskem slovarju, ki je šele v načrtu.

Trenutna, 3. izdaja slovarja, ki je bila objavljena leta 2006, ima skoraj 10.000 japonskih gesel s pribl. 25.000 slovenskimi prevodnimi ustreznicami, a samo 2.375 primerov rabe. Ker so primeri rabe nedvomno velikega pomena v slovarju za začetnike in ker se je v zadnjih letih na oddelku nabrala manjša zbirka japonsko-slovenskih vzporednih besedil (Hmeljak Sangawa 2007), smo se odločili to gradivo dograditi in urediti v manjši vzporedni korpus, da bi iz njega črpali primere besedne rabe za dvojezični slovar. V sledečih razdelkih opisujemo zbiranje gradiva za korpus, urejanje korpusa in črpanje primerov ter ocenjujemo uporabnost tako dobljenih primerov rabe.

2. Gradnja vzporednega korpusa

2.1. Zbiranje vzporednih besedil

Za pogoste jezikovne pare, še posebej za kombinacijo angleščine z drugim jezikom, obstajajo dandanes velike količine vzporednih besedil v elektronski obliki, velikokrat prosto dostopnih na spletu, kar je spodbudilo že več projektov in metod za avtomatsko črpanje vzporednih korpusov s spleta (npr. Resnik in Smith 2003; za japonščino Fry 2005; Tanaka 2001 idr.). Veliko je že tudi javno dostopnih urejenih zbirk stavčno poravnanih vzporednih besedil oz. baz prevajalskih spominov. Tako je npr. za slovenščino že od leta 2002 na spletu dostopna zbirka prevodov zakonodaje EU (Željko 2002), ki je danes dostopna na več lokacijah, v obsegu pribl. 56 milijonov besed (http://evrokorpus.gov.si/) oz. 27,7 milijonov slovenskih besed (Steinberger et al. 2006), iz iste zakonodaje pa so nastali tudi vzporedni korpusi, ki združujejo slovenščino z nemščino, francoščino, španščino in italijanščino. Tudi za japonščino obstajajo

velike količine japonsko-angleških besedil, iz katerih so tudi že črpali slovarsko ponazarjalno gradivo (Breen 2003; Tanaka 2001).

Za jezikovni par japonščina-slovenščina je jezikovnih virov bistveno manj: med obema jezikovnima področjema je razmeroma malo tako kulturnih kot gospodarskih in drugih stikov, do pred 10 leti pa skoraj ni bilo prevajalcev sposobnih prevajanja med japonščino in slovenščino. Zato v primerjavi z drugimi jezikovnimi pari obstaja razmeroma malo prevedenih besedil, še manj pa takih v lahko dostopni elektronski obliki. Zato smo skušali izkoristiti vse možne vire, do katerih smo imeli dostop, čeprav niso vsi vzorni primeri prevodov: od internega gradiva katedre za japonologijo FFUL do posrednih prevodov, t.j. parov prevodov istega (angleškega ali drugega) besedila v japonščino in slovenščino. Zbrali smo naslednje štiri vrste besedil.

2.1.1. Vrste besedil v korpusu

a) *Gradivo z vaj iz prevajanja na katedri za japonologijo*. Študenti 3. letnika japonologije prevajajo v slovenščino japonska besedila, večinoma o japonski družbi, običajih ali dogodkih, študenti 4. letnika pa prevajajo v japonščino slovenska besedila o slovenskih krajih in znamenitostih, zato besedila vsebujejo poleg splošnega tudi besedišče, ki je specifično za obe deželi. Izhodiščna besedila črpamo s spleta, tako da so že v elektronski obliki; študentske prevode preverimo in popravimo v razredu, nato jih objavimo v stavčno poravnani obliki kot študijsko gradivo na fakultetnem strežniku (http://e-ucenje.ff.uni-lj.si/), s čimer je bilo gradivo že pripravljeno za vključitev v korpus.

b) *Izročki in gradivo japonskih gostujočih profesorjev na FFUL*. Na Oddelku za azijske in afriške študije Filozofske fakultete Univerze v Ljubljani vsako leto gostuje eden ali več japonskih profesorjev, ki predavajo v japonščini. Del gradiva na oddelku prevedemo v slovenščino, da je dostopno tudi študentom nižjih letnikov oz. drugih študijskih smeri. Tako izvirna besedila kot prevodi so že bili v elektronski obliki, avtorji in prevajalci pa so se strinjali s tem, da besedila objavimo na spletu, tako da smo lahko tudi to gradivo vključili v korpus.

c) *Prevedeno leposlovje*. Ker so bili tako izročki gostujočih profesorjev kot tudi večina vaj iz prevajanja strokovna, vsebinsko in jezikovno zahtevna besedila, smo se odločili vključiti v korpus nekaj leposlovja s preprostimi vsakdanjimi dialogi, krajšimi povedmi in osnovnejšim besediščem. Zbrali smo odlomke iz dveh ducatov romanov in novel, ki so bili doslej prevedeni iz japonščine v slovenščino, ter prvega in doslej edinega knjižnega prevoda slovenske leposlovne proze, ki je izšel letos. Odlomke smo preslikali, pretvorili v besedilne datoteke s programom za razpoznavanje znakov, ročno preverili in poravnali. Leposlovna dela seveda krijejo avtorske pravice, ki omejujejo razmnoževanje in objavljanje besedil, toda menimo, da citiranje posameznih povedi kot ponazarjalni primer v slovarju zaradi omejene količine citiranega besedila ne krši avtorskih pravic.

d) *Spletno gradivo*. Da bi še dodatno razširili zbirko besedil, smo s pomočjo splošnih iskalnikov (google, najdi.si) poiskali vzporedna japonska in slovenska besedila. Za druge jezikovne pare, ki vključujejo angleščino, obstajajo že uspešni poskusi samodejnega črpanja vzporednih besedil v dveh jezikih s spleta (Resnik 1998, Resnik in Smith 2003, Fry 2005); spletnih strani, ki so prevedene iz raznih jezikov v angleščino, je veliko, vendar za jezikovni par japonščina-slovenščina nismo pričakovali velikih količin vzporednih besedil, zato smo za začetek ročno poiskali besedila in sproti preverjali njihovo relevantnost in kakovost.

Najprej smo poiskali strani v japonščini na domeni .si, tako s pomočjo iskalnikove funkcije za določanje jezika kot tudi z iskanjem znakov zlogovnice hiragane, ki se uporablja samo za zapis japonščine. Na ta način smo dobili 34 strani, od katerih sta bili 2 uporabni (predstavitvi podjetij), ostale pa so bile ali japonski povzetki slovenskih daljših besedil, ali nekakovostni strojni prevodi preko angleščine, ali strani v drugih jezikih, ki jih je iskalnik pomotoma razpoznal kot japonske.

Z obratnim postopkom smo poiskali tudi strani v slovenščini na domeni .jp, tako z omejevanjem iskanja na zadetke v jeziku "slovenščina" kot z iskanjem pogostih slovenskih besed. Na ta način smo dobili nekaj sto strani, a skoraj nobena ni bila v slovenščini: večinoma je šlo za napako iskalnika pri določanju jezika, tako da sta bili samo dve besedili (predstavitvi proizvodov) uporabni.

Največ uporabnih vzporednih besedil smo našli tako, da smo kot iskalni niz vnesli "Japanese" in "Slovene" oz. "Slovenian", ali "日本語" in "slovenščina". Tako smo dobili strani na portalih, kjer je glavno besedilo (večinoma v angleščini, a tudi v francoščini, ruščini in esperantu) prevedeno v več tujih jezikov, vključno z japonščino in slovenščino: od bolj tehničnih strani, kot so Googlove strani podpore, priročniki programske opreme, meniji in vmesniki raznih portalov, do bolj preprostih in razumljivih strani verskih skupin (Jehovove priče ipd.), jezikovnih priročnikov, turističnih informacij in podobnih besedil na težavnostni ravni, ki je primerna tudi za začetnike.

Na prvi pogled idealni vir besedil za naš korpus je tudi urejena zbirka vzporednih besedil projekta Opus (Tiedemann 2007, Tiedemann in Nygaard 2004), v katerem so zbrana prosto dostopna in prosto objavljiva besedila, ki so že stavčno poravnana s prevodi v več jezikih (priročnik KDE je npr. preveden v kar 61 jezikov). Podkorpusa KDE in OpenSubtitles vključujeta tako japonščino kot slovenščino, vendar prvi vsebuje veliko zelo tehničnega izrazja in težko razumljivega besedila, v drugem pa so fraze sicer vsakdanje in razumljive tudi za začetnika, toda velikokrat se prevod sploh ne ujema, saj so podnapisi po svoji naravi povzetki govorjenega besedila, pri čemer se v japonskem in slovenskem podnapisu iste (največkrat angleške) govorne enote včasih ohranijo različni deli le-te, kar veliko pa je bilo tudi napak v prevodih, ki jih prispevajo prostovoljci. Zato tega gradiva nismo vključili v tokratni korpus.

2.2. Vzporejanje besedil na ravni povedi in izgradnja korpusa

Dobljena besedila v raznih formatih smo nato normalizirali v besedilno obliko, poenotili vse kodiranje v UTF-8 in poravnali na ravni povedi.

Besedila iz prve skupine so bila že poravnana, pri ostalih pa smo uporabili orodje +Tools Align iz programskega paketa za pomnilnike prevodov Wordfast (http://www.wordfast.net); poravnavo smo v celoti ročno preverili in pri tem izločili povedi z očitnimi napakami in povedi, kjer je prevod manjkal.

Tako dobljena besedila smo pretvorili v obliko XML. Japonski del besedil smo lematizirali s prosto dostopnim

orodjem za morfološko analizo japonskih besedil Chasen (http://sourceforge.jp/projects/chasen-legacy/), slovenski del besedila pa lematizirali z lastnim orodjem totale (Erjavec idr. 2005).

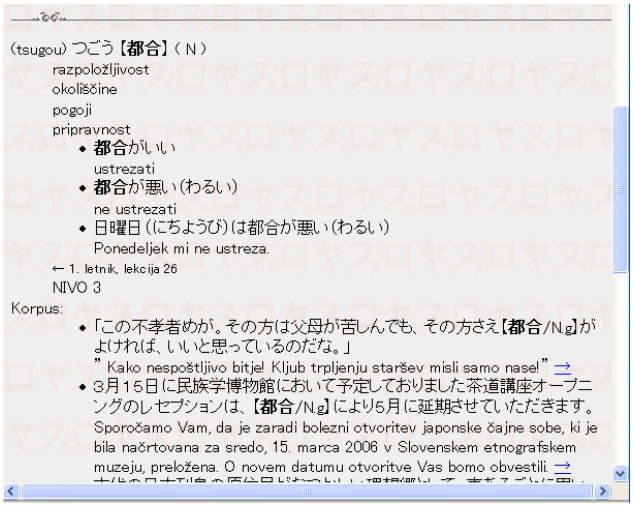
Tako smo dobili vzporedni lematizirani korpus s 7914 prevodnimi enotami (povedmi), v japonskem delu 226.220 besed, v slovenskem delu pa 171.261 besed. Razmerje med vrstami gradiv je prikazano v tabeli 1.

tip besedila	delež v korpusu
vaje iz prevajanja	24,5%
gradivo gost. profesorjev	13,5%
leposlovje	15,7%
spletno gradivo (od tega neposrednih prevodov 1,4% posrednih prevodov 44,9%)	46,3%

Tabela 1: Razmerje med vrstami gradiv.

3. Črpanje in vnašanje primerov v slovar

Vsa gesla japonsko-slovenskega slovarja smo avtomatsko poiskali v korpusu in tako dobljene povedi dodali k ustreznim geslom, s čimer je 4.648 japonskih gesel dobilo nove primere. Pri osnovnem besedišču, kjer so v geslih že bili primeri rabe, smo primere iz korpusa grafično ločili od obstoječih slovarskih primerov, kot je prikazano v sliki 1. Na koncu vsakega primera iz korpusa puščica označuje povezavo na stran z bibliografskimi podatki o besedilu, iz katerega je primer vzet: vir (naslov spletne strani ali podatki o knjigi) originalnega besedila in prevoda ter jezik in avtor le-teh.



Slika1. Geslo s prvotnimi primeri in primeri iz korpusa

Pri pogostih besedah, ki so se v korpusu pojavile v velikem številu povedi, smo v slovar vnesli samo 6 najkrajših primerov. Leksikografsko bolj utemeljeni kriteriji za izbor primerov, ki jih v tokratnem delu še nismo uspeli upoštevati, so predstavljeni v razdelku 4.

Ker so primeri kljub temu razmeroma dolgi, je iztočnica v japonskem primeru grafično poudarjena. Za uporabnika bi bila še bolj pregledna grafična izpostavitev tistega dela slovenskega primera (besede ali zveze), ki odgovarja japonski iztočnici, toda samodejno vzporejanje besedila na besedni ravni presega zmožnosti tokratnega projekta.

4. Ocena samodejno pridobljenih primerov

Z opisanim postopkom smo polovici gesel v slovarju uspeli dodati dejanske primere rabe z ustreznim prevodom. Primeri rabe v slovarju so velikega pomena pri učenju jezika, saj lahko uporabnik iz njih pridobi informacije o pomenskih, skladenjskih, kolokacijskih, pragmatičnih in slogovnih lastnostih določene besede. To je še posebej pomembno pri japonščini, ki ima v primerjavi z evropskimi jeziki slogovno in pragmatično razčlenjeno besedišče, saj se za isti pojem, ki se v slovenščini izraža z eno samo besedo, velikokrat uporabljajo različni sinonimi glede na formalnost situacije, govorčevo starost in spol, sobesedilo idr. Pri takih besedah so v geslih našega slovarja kazalke na bolj ali manj vljudne idr. sinonime, ravno iz primerov pa je najbolj nazorno in intuitivno razvidno tipično sobesedilo določene besede.

Idealni slovarski primer rabe naj bi bil obenem razumljiv, informativen in tipičen. Primeri iz korpusa seveda niso deležni običajnega slovaropisnega postopka analize korpusnih primerov, sinteze geselskega članka z definicijo oz. prevodom in ureditve oz. priredbe navedenih primerov, zato se razlikujejo od običajnih slovarskih (prirejenih ali izmišljenih) primerov v tem, da njihovo besedišče ni omejeno na osnovno (torej so lahko ostale besede v primeru težje od iztočnice, ki jo primer ponazarja), skladnja ravno tako ni nujno preprosta, večinoma so vezani na sobesedilo, tako da niso vedno razumljivi, ko jih iz sobesedila iztržemo, pri primerih iz vzporednega korpusa pa se zgodi tudi to, da v slovenskem prevodu sploh ni prevoda japonske iztočnice, ki naj bi jo primer ponazarjal, ker se pojavlja v širšem sobesedilu. Te pomanjkljivosti so delno rešljive: k vsem besedam v primerih bi lahko avtomatsko dodali povezave k ustreznim geslom, da lahko uporabnik lažje preveri pomen neznanih besed v primeru – glede na to, da je v japonskem delu korpusa 13.083 pojavnic, bi morali zato ustrezno dopolniti tudi gesla samega slovarja, ki obsega 10.000 gesel. Težave na skladenjskem nivoju so delno rešljive z implementacijo algoritma za izbiranje najlažjih primerov, v končni fazi pa bo še vedno potrebno opozorilo uporabniku, da gre za realne primere rabe s prevodnimi ustreznicami, ki niso nujno tako splošne kot ustreznice v jedru gesla.

Zaradi razmeroma majhnega obsega našega korpusa pridobljeni primeri pričakovano ne pokrivajo vseh podpomenov geselskih iztočnic, tudi ne ponazarjajo vseh njenih najbolj tipičnih skladenjskih, kolokacijskih, slogovnih in pragmatičnih vzorcev. Vendar smo opazili, da korpusni primeri prispevajo tudi nove prevodne ustreznice, ki v slovarju še niso bile navedene, še posebej pri frazemih. Tako je npr. iztočnica **あわせる** (*awaseru*) v zadnji različici slovarja bila prevedena samo kot »nastaviti« oz. »sešteti«, v korpusnih primerih pa smo našli tudi daljše prevodne enote: **顔をあわせる** (*kao wo awaseru* - dobesedno *postaviti skupaj obraza*) s prevodoma »srečati, videti se«, ter **声を合わせて歌う** (koe wo awasete utau - dobesedno *peti in združiti glasove*) s prevodom »peti skupaj«. Glede na to, da so ravno kolokacije ena od težjih leksikalnih vidikov učenja tujega jezika (Gorjanc in Jurko 2004), bi naš slovar potreboval dopolnitev kolokacijskih informacij, še posebno glede na to, da se ga uporabniki poslužujejo tudi za aktivno rabo, z

iskanjem slovenskih besed. Ravno korpusni primeri z večbesednimi leksikalnimi enotami in njihovimi prevodi v kontekstu lahko tu nudijo koristno gradivo tako za uporabnike kot tudi za urejevalce slovarja pri postopnem dopolnjevanju le-tega.

Pri dveh podobnih projektih za samodejno črpanje primerov iz korpusa za vnos v japonski slovar (Yoshihashi in Nishina 2007, Mizuno idr. 2008) se pri izboru primerov poleg kriterija dolžine uporabljajo tudi drugi kriteriji: oba projekta izbirata primere glede na težavnost besedišča, pismenk in slovničnih vzorcev, ki se merijo po standardih izpita iz japonščine kot tujega jezika (Japan Foundation 2004), Yoshihashi in Nishina pa poleg tega na osnovi samodejne skladenjske analize izbirata skladenjsko najmanj razvejane povedi. V projektu za črpanje angleških slovarskih primerov iz korpusa (Kilgarriff idr. 2008) se poleg kriterijev za izbiranje čim bolj razumljivih primerov uporabljajo tudi kriteriji za določanje najbolj tipičnih primerov, na osnovi merjenja kolokacijskih vzorcev. Tudi pri našem projektu v prihodnosti načrtujemo dopolnitev kriterijev za izbiranje primerov, v trenutni fazi pa ocenjujemo, da so primeri s prevodom kljub težavnosti lahko koristni za uporabnika slovarja, ki se bo zaradi omejenih človeških virov šele počasi dopolnjeval. Ob tem se mora seveda uporabnik zavedati, da korpusni primer nudi prevodno rešitev v konkretnem sobesedilu, ki je ne moremo posplošiti tako kot pri izmišljenih, čim bolj splošno veljavnih primerih v slovarjih.

5. Zaključek

Predstavili smo metodo zbiranja vzporednega korpusa in črpanja primerov rabe za vključitev v dvojezični slovar. Pridobljeni korpus in primeri so se izkazali za koristne, ker so prispevali ilustrativno gradivo za polovico slovarskih iztočnic, obenem pa bi večji in bolj uravnotežen korpus omogočil boljše kritje tako števila iztočnic kot tudi podpomenov in posebnosti rabe pri vsaki iztočnici, zato načrtujemo nadaljevanje začetega dela.

Literatura

Breen, Jim W. 2003. Word Usage Examples in an Electronic Dictionary. PAPILLON-2003 Workshop on Multilingual Lexical Databases, Hokkaido University, 12pp. {http://www.csse.monash.edu.au/~jwb/papillon/dicexamples.html}

Erjavec, Tomaž , Kristina Hmeljak Sangawa, I. Srdanović, 2003. An XML TEI encoding of a Japanese-Slovene learner's dictionary. V V. Rajkovič idr. (ur.). *Zbornik B 6. mednarodne multi-konference Informacijska družba IS 2003*. Ljubljana: Institut Jožef Stefan.

Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen, Ralf Steinberger. 2005. Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland. pp. 32-36.

Erjavec, Tomaž , Kristina Hmeljak Sangawa, Irena Srdanović Erjavec, 2006. JaSlo, a Japanese-Slovene learners' dictionary: methods for dictionary enhancement. *Proceedings XII EURALEX international congress*, Torino: Edizioni Dell'orso. pp. 611-616.

Fry, John. 2005. Assembling a parallel corpus from RSS news feeds, in Proceedings of the Workshop on Example-Based Machine Translation, MT Summit X, Phuket, Thailand, September 2005. {http://www.ai.sri.com/pubs/files/1181.pdf}

Gorjanc, Vojko, Primož Jurko. 2004. Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49, št. 3-4, str. 49-62

Hmeljak Sangawa, Kristina, 2002. Slovar japonskega jezika za slovenske študente japonščine. V Tomaž Erjavec in Jerneja Gros (ur.), *Jezikovne tehnologije: zbornik konference*. Ljubljana: Institut Jožef Stefan.

Hmeljak Sangawa, Kristina, 2007. Hon'yaku no jugyou ni okeru taiyaku koopasu no kouchiku to sono katsuyou no kanousei. V C.Tsuchiya in A. Bekeš (ur.), *Nihongo kyouiku renraku kaigi ronbunshuu*, vol. 19, Ljubljana/Aichi, str. 59-63

Japan Foundation. 2004. Japanese Language Proficiency Test: Test contents specifications. Tokyo: Bonjinsha.

Kawamura, Yoshiko idr, 2005. Implementation and Evaluation of a Web-based Multilingual Editing System for the Reading Tutorial Dictionary Tool. V *Dai10kai Yooroppa nihongokyouiku shimpojiumu - Leuven*. {http://language.tiu.ac.jp/aje2005.pdf}

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell, Pavel Rychlý, 2008. GDEX: Automatically finding good dictionary examples in a corpus. V *Proceedings of 13th Euralex International Congress*.

Mizuno, Junta idr. 2008. Nihongo dokkai shien no tame no gogi goto no yourei chuushutsu shisutemu no kouchiku. V *Proceedings of the Workshop on Natural Language Processing for Education - The 14th Annual Meeting of the Association for Natural Language Processing.* Tokyo: Gengoshorigakkai. str. 63-66.

Resnik, Philip, Noah A. Smith. 2003. The Web as a parallel corpus. Computational Linguistics, Volume 29 , Issue 3 (September 2003). str. 349 - 380. {http://citeseer.ist.psu.edu/resnik03web.html}

Srdanović Erjavec, Irena, Tomaž Erjavec, Adam Kilgarriff, 2008. A web corpus and word sketches for Japanese. *Shizen gengo shori*, 2008, vol. 15, no. 2, str. 137-159.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga, 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*. Genoa, Italy, 24-26 May 2006.

Tanaka, Yasuhito. 2001. Compilation of a multilingual parallel corpus. PACLING 2001, Sep 2001, Japan. {http://afnlp.org/pacling2001/pdf/tanaka.pdf}

Tiedemann, Jörg. 2007. Building a Multilingual Parallel Subtitle Corpus. In Proceedings of CLIN 17, Leuven, Belgium, 2007.

Tiedemann, Jörg, Lars Nygaard. 2004. The OPUS corpus - parallel & free. V *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04*). Lisbon, Portugal, May 26-28.

Yoshihashi, Kenji, Kikuko Nishina, 2007. Gakushuusha ni awaseta reibun hyouji tsuuru. V *CASTEL-J in Hawaii 2007 Proceedings*. str. 223-226.

Željko, Miran, 2002. Pripomočki na spletu za prevajalce zakonodaje EU. *Zbornik mednarodne konference Informacijska družba 2002 – jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

Predstavitev in analiza slovenskega wordneta

Darja Fišer*, Tomaž Erjavec†

* Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana
darja.fiser@guest.arnes.si
† Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

V prispevku predstavljamo prvi prosto dostopen slovenski semantični leksikon z imenom sloWNet, ki je bil izdelan avtomatsko s pomočjo prosto dostopnih korpusnih in leksikalnih virov. Pri gradnji smo večpomenske besede razdvoumili s pomočjo besedno vzporejenega večjezičnega korpusa in drugih že obstoječih wordnetov, enopomenske besede in besedne zveze pa smo pridobili iz dvojezičnih virov. Predstavljena različica wordneta vsebuje skoraj 20.000 različnih literalov oz. 17.000 sinsetov, ki so večinoma samostalniški. Težišče prispevka je na analizi wordneta glede na skupine konceptov in domene, v katere generirani sinseti spadajo, iz katerega vira so bili sinseti ustvarjeni in kakšne relacije veljajo med njimi. Podrobneje se posvetimo hipernimiji, ki je v wordnetu najpogostejša, in izračunamo dolžino hipernimskih verig. V drugem delu analize besedišče iz wordneta primerjamo s korpusom jos100k. Proučimo pokritje samostalnikov, ki se pojavljajo v korpusu, in na vzorcu večpomenskih samostalnikov s pomočjo konkordanc ročno preverimo zastopanost njihovih pomenov v wordnetu.

Presentation and analysis of Slovene wordnet

The paper presents the first freely available Slovene semantic lexicon called sloWNet which was developed automatically from already existing freely available corpus and lexical resources. In the construction process, polysemous words were disambiguated with a word-aligned multilingual parallel corpus and already existing wordnets for these langauges. On the other hand, translations for monosemous words were obtained from bilingual sources. SloWNet contains almost 20,000 literals or 17,000 synsets which are mostly nominal. The paper focuses on the analysis of wordnet with respect to what kind of concepts are found in sloWNet, which domains they belong to, what resource they were created from and what relations hold between them. We take a closer look at hypernymy, the most common relation in wordnet, and compute the length of hyponymy chains. The second part of the analysis compares the wordnet vocabulary with the jos100k corpus by examining to what extent nouns from the corpus are covered in sloWNet and how well the senses of polysemous words are represented in sloWNet.

1. Uvod

V prispevku predstavljamo slovenski wordnet, pojmovno zasnovano leksikalno zbirko, v kateri so besede (literali) z istim pomenom združene v množice sinonimov (sinsetov), ti pa so med seboj povezani z leksikalnimi in semantičnimi relacijami. Prva tovrstna zbirka za angleški jezik je začela nastajati pred dobrima dvema desetletjema na Univerzi v Princetonu (Fellbaum 1998) in je kmalu postala eden najbolj priljubljenih pripomočkov pri avtomatskem razumevanju naravnega jezika, kot so pridobivanje in luščenje informacij, razdvoumljanje in strojno prevajanje. Vendar WordNeta raziskovalci niso samo veliko uporabljali, temveč so začeli ustvarjati podobne zbirke tudi za druge jezike. Tako so pod okriljem mednarodnih projektov EuroWordNet (Vossen 1998) in BalkaNet (Tufis in dr. 2000) nastali wordneti za številne evropske jezike, s čimer je wordnet pridobil pomembno večjezično razsežnost. Od takrat naprej pa družina wordnet samo še raste; združenje Global WordNet Association[1] na svojih spletnih straneh trenutno poroča o obstoju wordnetov v 50 različnih jezikih.

Gradnje wordnetov se raziskovalci lotevajo na različne načine, ki so odvisni od leksikalnih virov, ki jih imajo na voljo. Nedvomno je ročna gradnja neodvisnih wordnetov za vsak jezik posebej najbolj zanesljiv pristop, saj zagotavlja najboljše rezultate. Vendar je ta podvig preveč dolgotrajen in predrag, da bi si ga za večino jezikov sploh lahko privoščili. Zato so ga številni raziskovalci skušali poenostaviti in pospešiti s polavtomatskimi ali avtomatskimi pristopi, ki izkoriščajo že obstoječe vire. Mednje sodijo eno- in dvojezični slovarji (Rigau idr. 1998, Knight in Luk 1994) ter taksonomije in druge obstoječe ontologije (Yao in dr. 2007).

Tudi slovenski wordnet smo skušali izdelati s pomočjo že obstoječih virov, kar so za slovenski jezik večinoma korpusi, saj za slovenščino prosto-dostopnih in obsežnih leksikalnih virov za splošno besedišče praktično ni.

Po vzoru uspešno izdelanih wordnetov za številne jezike, kot so na primer italijanščina, romunščina, in korejščina, smo sledili t.i. razširitvenemu modelu (Vossen 1998), ki prevzema strukturo in relacije iz angleškega WordNeta. Prednost tega modela je, da zagotavlja najvišjo možno stopnjo ujemanja med različnimi jeziki. Pristop vključuje tudi visoko stopnjo avtomatizacije, kar raziskovalcem prihrani ogromno časa in finančnih sredstev. Poleg prednosti razširitveni model prinaša tudi negativne posledice, med katerimi je najpomembnejša velika odvisnost od leksikalne in konceptualne strukture izvornega (angleškega) jezika. Če nismo dovolj pozorni, je lahko izdelana leksikalna baza arbitrarna in z dejansko organizacijo in leksikalizacijo konceptov v nekem jeziku nima veliko skupnega (Orav in Vider 2004, Wong 2004). Vendar obstajajo učinkoviti mehanizmi tudi za spopadanje z leksikalnimi vrzelmi in denotacijskimi razlikami (Bentivogli 2000), zato preprostost pristopa in možnost avtomatizacije prepričljivo odtehtata njegove slabosti.

[1] http://www.globalwordnet.org/gwa/wordnet_table.htm [10.05.2007]

Prispevek predstavlja wordnet za slovenščino, ki smo ga razvili v skladu z zgornjimi načeli. V drugem razdelku opisujemo postopek gradnje, v tretjem analiziramo vsebino izdelane zbirke, njeno pokritje pa v četrtem razdelku preverjamo na korpusu jos100k. Prispevek sklenemo s sklepi in načrti za prihodnje delo.

2. Izdelava slovenskega wordneta

Pri delu smo izhajali iz predpostavke, da so prevodi verodostojen semantični vir in da je semantično relevantne informacije mogoče izluščiti iz različnih že obstoječih virov. Z idejo, da je semantične informacije mogoče pridobiti iz prevodne relacije, so se že ukvarjali Diab (2002), Ide idr. (2002) ter Resnik in Yarowsky (1997).

Osnovni nabor sinsetov smo pridobili z avtomatskim prevajanjem srbskega wordneta s pomočjo slovensko-srbskega slovarja, ki smo jih nato tudi ročno pregledali in popravili (glej Erjavec in Fišer 2006).

Nadaljnji razvoj je izhajal iz Princeton WordNeta (PWN) in je potekal v dveh delih. Prevodne ustreznice za literale, ki imajo v PWN samo en pomen in jih torej ni potrebno razdvoumljati, smo izluščili iz Wikipedije in sorodnih virov, podobno, kot so to storili Declerck idr. (2006) in Casado idr. (2005). V PWN je enopomenskih več kot 80 % literalov, kar pomeni, da smo lahko zgolj z dvojezičnim pristopom pridobili veliko število zanesljivih prevodov.

S pomočjo večjezičnih vzporednih korpusov in wordnetov za druge jezike pa smo se spopadli še z večpomenskimi literali. Besedno vzporejene paralelne korpuse so za iskanje sinonimov oz. ločevanje pomenov besed uporabili že van der Plas in Tiedemann (2006) ter Dyvik (2002).

2.1. Večpomenski literali

Pri tem pristopu smo uporabili večjezični vzporedni korpus SEE-ERA.NET[2] (Tufis in dr. 2008), ki je podkorpus korpusa JRC-Acquis (Steinberger in dr. 2006). Vsebuje približno 1 milijon besed v osmih jezikih (mi smo poleg slovenščine uporabili še angleščino, romunščino, češčino in bolgarščino) in je že stavčno poravnan. Z uporabo različnih prosto dostopnih orodij smo korpus v vseh petih jezikih besednovrstno označili in ga lematizirali ter ga nato z orodjem Uplug (Tiedemann 2003) še vzporedili na ravni besed. Na podlagi tega smo izluščili dvojezične leksikone, ki smo jih nato združili v večjezične.

Dobljene leksikone smo primerjali z že obstoječimi wordneti za te jezike. Za angleščino smo uporabili PWN, za češki, romunski in bolgarski jezik pa wordnete iz projekta BalkaNet. Če smo našli ujemanje med leksikonskim vnosom v enem izmed jezikov in sinsetom v ustreznem wordnetu, smo si za to besedo zapomnili ID sinseta, v katerem se je pojavila. To smo ponovili za vse jezike, nato pa smo poiskali presek med vsemi pripisanimi ID-ji v različnih jezikih in rezultat pripisali slovenski ustreznici v leksikonu. Na koncu smo vse slovenske besede, ki so od ostalih jezikov podedovale isti ID, združili v isti sinset.

[2] http://dcl.bas.bg/ssbc/home.html [15.06.2008]

2.2. Enopomenski literali

Za iskanje prevodnih ustreznic angleških literalov z enim samim pomenom smo dvojezične leksikone izluščili iz naslednjih prosto dostopnih zbirk (glej (Fišer in Sagot 2008):

- Eurovoc[3] je večjezični tezaver, ki ga v EU uporabljajo za klasifikacijo dokumentov. Uporabili smo različico 4.2, ki vsebuje 6,802 deskriptorjev in njihove prevode 21 evropskih jezikov.
- Wikipedia[4] je spletna enciklopedija, ki jo sestavljajo in dopolnjujejo prostovoljci. Angleško-slovenski leksikon smo iz nje izluščili na podlagi povezav med članki na isto temo v obeh jezikih, ki jih člankom dodajajo uporabniki, zato ne vsebujejo veliko napak. Leksikon smo nadgradili še s preprosto analizo besedila člankov, s katero smo normalizirali velike in male začetnice iztočnic, izluščili sinonime, za nekatere vnose pa tudi definicijo.
- Angleški in slovenski Wiktionary[5] sta nastala na podlagi iste iniciative kot Wikipedija, vsebujeta pa definicije izrazov in njihove prevodne ustreznice v številne druge jezike.
- Wikispecies[6] je taksonomija živih bitij v latinščini, za pogoste živali in rastline pa vsebuje tudi prevode v različne jezike, nekaj jih je tudi v slovenščini.

2.3. Združevanje rezultatov

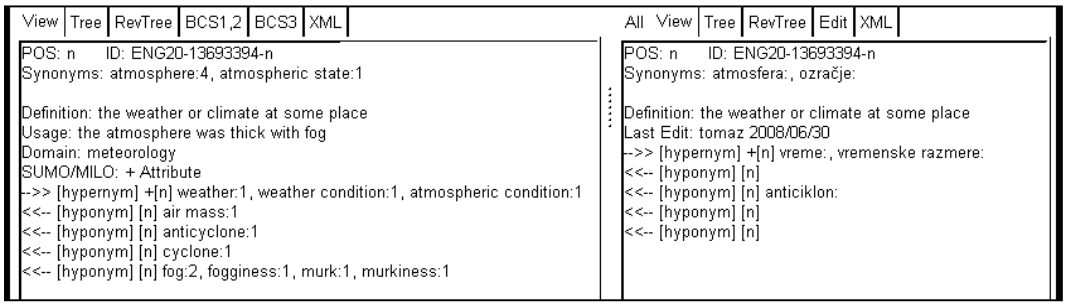
Na koncu smo rezultate vseh pristopov združili. Sinsete, ki smo jih pregledali in popravili ročno, smo prevzeli iz prejšnje različice. Avtomatsko generirane sinsete pa smo združili tako, da smo upoštevali vse pridobljene literale, pri čemer smo ohranili informacijo o njihovem izvoru. Na podlagi tega smo združen wordnet še filtrirali glede na zanesljivost in raznolikost virov, ki so posamezni literal prispevali.

Ker z avtomatskim pristopom nismo mogli izdelati celotnega wordneta, smo se strukturnim luknjam v mreži, ki aplikacijam otežujejo uporabo wordneta, izognili tako, da smo za manjkajoče sinsete iz angleškega wordneta prevzeli njihovo strukturo in relacije. Zavedamo se, da bo prazne sinsete potrebno čim prej zapolniti, do takrat pa bodo aplikacijam v pomoč pri iskanju splošnejšega ali bolj specifičnega sinseta oziroma drugih semantičnih relacij.

2.4. Struktura slovenskega wordneta

Wordnet je sestavljen iz sinsetov, v katerih so združene besede in besedne zveze (literali), ki označujejo isti koncept. Vsak sinset ima svojo identifikacijsko kodo, na podlagi katere je mogoče najti ekvivalenten sinset v wordnetih za vse ostale jezike, ki uporabljajo kode PWN. Sinset vsebuje še informacije o besedni vrsti, skupini konceptov, ki jim pripada, področno oznako, povezavo na ontologijo SUMO ter semantične in leksikalne relacije, ki kažejo na druge sinsete v mreži. Zbirko smo oblikovali v formatu XML, ki ga zahteva pregledovalnik in urejevalnik VisDic in je prikazan v sliki 1 (Horak in Smrž 2000).

[3] http://europa.eu/eurovoc [15.03.2008]
[4] http://www.wikipedia.org [15.03.2008]
[5] http://www.wiktionary.org [15.03.2008]
[6] http://species.wikimedia.org [15.03.2008]



Slika 1. Primer sinseta v VisDicu.

3. Analiza slovenskega wordneta

V tem razdelku analiziramo izdelano leksikalno zbirko. Najprej naštejemo nekaj kvantitativnih podatkov o wordnetu, nato preverimo, kakšne sinsete smo z opisanimi metodami dobili. Zanima nas njihova besedna vrsta, skupine konceptov in domene, v katere generirani sinseti spadajo, ter vir, iz katerega so bili ustvarjeni. Prav tako preverimo, kakšne relacije veljajo med njimi, pri čemer se najbolj posvetimo hipernimiji, ki je najpogostejša.

3.1. Osnovni podatki o wordnetu

S kombinacijo metod, opisanih v prejšnjem razdelku, smo dobili 16.886 sinsetov oz. 19.582 različnih literalov. Močno prevladujejo sinseti, ki vsebujejo samo en literal (11.099), sinsetov z več literali je razmeroma malo (4.146). Popvprečna dolžina sinseta je 1,16 literala, najdaljši sinset ima 16 literalov (ENG20-02498705-v) in izhaja iz prve različice wordneta, kjer so bili angleški sinseti prevedeni s pomočjo dvojezičnega slovarja, nato pa ročno pregledani in popravljeni. Slovenski wordnet vsebuje tako enobesedne (11.099) kot večbesedne literale (8.483). Enobesedne literale smo pridobili predvsem iz korpusa, večbesedne pa iz Wikivirov, Eurovoca in z ročnim pregledom avtomatsko generiranih sinsetov prve različice slovenskega wordneta.

3.2. Analiza sinsetov glede na besedno vrsto in skupine konceptov

V tabeli 1 prikazujemo rezultate analize sinsetov glede na besedno vrsto in skupine konceptov, v katere sodijo. Zaradi virov in metod, ki smo jih za izdelavo wordneta uporabili (Wikiviri večinoma opisujejo samostalniške iztočnice, vzporejanje korpusa na ravni besed pa prav tako najbolje deluje za samostalnike), je v izdelanem wordnetu največ ravno samostalnikov. Sledi jim nekaj glagolov in pridevnikov, prislovov pa nam zaenkrat še ni uspelo pridobiti. Za lažjo gradnjo novih wordnetov in njihovo medsebojno primerjavo so na pobudo projekta BalkaNet koncepti v wordnetu ločeni na osnovne in specifične, osnovni pa so nadalje razvrščeni v tri skupine (glej Tufis in dr. 2000).

Bes. vrsta	BCS 1	BCS 2	BCS 3	Specifični	Skupaj
sam.	950	1.611	902	11.943	15.406
prid.	0	37	90	290	417
gl.	251	506	158	146	1.061
Skupaj	1.201	2.154	1.150	12.379	16.884

Tabela 1: Sinseti glede na besedno vrsto in skupine konceptov.

Čeprav so nekatere odločitve o razvrščanju konceptov v skupine sporne, načeloma velja, da čim splošnejši kot je koncept in višje kot je v hierarhiji wordneta, tem bolj je v jeziku pomemben. Tako je v skupini najosnovnejših konceptov na primer koncept tekočina, med specifične pa spada industrijska revolucija. Osnovni koncepti iz prvih dveh skupin so v slovenskem wordnetu zelo dobro zastopani, saj so večinoma podedovani iz prejšnje različice, v kateri smo se osredotočili ravno nanje, z avtomatskim pristopom pa nam je uspelo pridobiti še nekaj konceptov iz tretje skupine in veliko število specifičnih konceptov, tako da slovenski wordnet vsebuje četrtino vseh konceptov iz PWN.

3.3. Analiza literalov glede na domene in vire, iz katerih so bili ustvarjeni

Wordnet vsebuje tudi področne oznake za posamezne koncepte (Bentivogli in dr. 2004), zato smo za boljšo vsebinsko predstavo o slovenskem wordnetu pregledali, katere domene so v njem. Sinseti v PWN so razvrščeni v približno 200 domen, slovenski pa jih vsebuje 144. V Tabeli 2 je naštetih deset domen, ki so v slovenskem wordnetu najpogosteje pripisane literalom skupaj z viri, na podlagi katerih so bili ti ustvarjeni. Najpogostejša je najsplošnejša domena *faktotum*, ki so jo pripisali vsem sinsetom, za katere ni bilo mogoče določiti nobene bolj specifične domene. Sledijo ji koncepti iz domen *zoologija*, *botanika* in *biologija*. Največ konceptov smo pridobili iz Wikipedije in sorodnih virov, ki so prispevali največ literalov ravno za koncepte s področja biologije. Ročno ustvarjeni koncepti so predvsem splošni, podobno velja za koncepte, generirane iz korpusa in Eurovoca.

	Vir					
Domena	seera	euro	ročno	wiki	več	Skupaj
factotum	1.386	71	3.246	310	16	5.029
zoology	38	9	63	3.160	6	3.276
botany	40	8	73	2.368	3	2.492
biology	19	4	56	1.390	4	1.473
admin.	33	58	79	502	169	841
chemistry	59	32	66	446	49	652
geography	10	38	65	225	39	377
anatomy	26	2	139	172	11	350
religion	2	6	47	235	2	292
economy	93	46	121	17	4	281
drugo (134)	1.225	511	2.170	4.015	169	8.090
Skupaj	2.931	785	6.125	12.840	472	23.153

Tabela 2: Zastopanost domen in viri, iz katerih so bili literali pridobljeni[7].

[7] Večpomenski literali, ki se pojavljajo v več sinsetih, so šteti za vsak pomen posebej.

Kljub temu, da korpus SEE-ERA.NET ni splošen, je korpusni pristop dal toliko splošnega besedišča, ker smo leksikon razdvoumljali s pomočjo wordnetov iz BalkaNeta, ki so ob zaključku projekta pokrivali predvsem osnovne koncepte. Zanimiv je podatek, da večjega prekrivanja med viri ni. To pomeni, da smo za izdelavo wordneta izbrali raznolike vire, ki so prispevali raznoliko besedišče in tako prispevali k bogatosti wordneta.

3.4. Analiza relacij med sinseti

Glede na to, da so sinseti v wordnetu med seboj povezani v mrežo, nas je zanimalo, katere relacije med njimi so najpogostejše. Tabela 3 vsebuje relacije med sinseti. Upoštevane so samo tiste relacije, ki izhajajo iz zapolnjenega sinseta in kažejo na drug zapolnjen sinset. To pomeni, da prazni sinseti, ki smo jih zaradi ohranjanja celovitosti mreže prevzeli iz PWN, niso upoštevani. V povprečju je skoraj vsak samostalnik povezan z enim sinsetom, vsak glagol z dvema, tretjina pridevnikov pa neposredno ni povezana z nobenim slovenskim sinsetom. Najpogostejša relacija je hipernimija, s tem pa tudi njena inverzna relacija hiponimija. Holonimija je razdeljena na tri relacije (pripadnik, del in kos). Relacije, kot so derivacija, antonimija in glagolska skupina, niso povsem jezikovno neodvisne, zato jih bo za slovenščino v prihodnosti potrebno preveriti in potrditi.

	Kaže neposredno na slovenski sinset			
Relacija	prid.	sam.	gl.	sl. skupaj
hypernym	0	7.340	729	8.069
holo_member	0	4.466	0	4.466
eng_derivative	0	1.066	1.066	2.132
holo_part	0	1.051	0	1.051
drugo (12)	294	916	411	1.621
Skupaj	294	14.839	2.206	17.339

Tabela 3: Relacije v slovenskem wordnetu.

Podrobneje nas je zanimala hipernimija, ki predstavlja 46 % vseh relacij med slovenskimi sinseti. Ker je ta relacija najpogostejša ravno pri samostalnikih (91 %), smo preverili, kako dolge so posamezne hipernimske verige za samostalniške sinsete od vsakega sinseta do vrhnjega koncepta in koliko praznih sinsetov vsebujejo. Razveseljivo je, da je vseh 9 vrhnjih sinsetov v slovenščini: *abstrakcija, dejanje, dogodek, entiteta, lastnina, pojav, psihološka značilnost, skupina* in *stanje*. Kot prikazuje tabela 4, ima večina verig do 10 sinsetov, več kot to jih ima samo 7 % verig, pri čemer imajo najdaljše tri 16 vozlišč (npr. veriga med *telica* ↔ *entiteta*). 46 % vseh verig je neprekinjenih, 52 % jih vsebuje manjše število praznih sinsetov (večinoma po enega), samo 2 % verig je takih, ki vsebujejo po pet ali več lukenj.

	Dolž. verige				
Dolž. luknje	vrh	<5	<10	≥10	Skupaj
vrh	9	-	-	-	9
0	-	4.206	2.861	273	7.340
<5	-	3.384	4.227	617	8.228
>5	-	0	63	285	348
Skupaj	9	7.590	7.151	1.175	15.925

Tabela 4. Hipernimske verige.

4. Pokritje besedišča v korpusu jos100k

Izdelan wordnet smo evalvirali avtomatsko in ročno. Avtomatska evalvacija je bila opravljena s primerjavo generiranega wordneta in ročno izdelanega prototipnega wordneta (glej Erjavec in Fišer 2006), pri kateri smo izmerili 70 % natančnost. Evalvacija vzorca ročno pregledanih sinsetov pa je pokazala, da je približno 7 % literalov, ki jih v prototipnem wordnetu ni, kljub temu pravilnih, nadaljnjih 5 % pa jih je tesno povezanih s sinsetom, v katerem se pojavljajo. Evalvacijo podrobneje opisujemo v Fišer in Sagot (2008), v tem prispevku pa se želimo posvetiti primerjavi besedišča iz wordneta in korpusa jos100k (Erjavec in Krek, 2008), ki je podkorpus, vzorčen iz korpusa Fida+. Vsebuje 100.000 besed in je označen z ročno preverjenimi oblikoslovnimi oznakami in lemami, za analizo pa smo se ga odločili uporabiti, ker ga v nadaljevanju raziskav nameravamo označiti tudi na pomenski ravni. Ker wordnet vsebuje predvsem samostalnike, smo preverili, do katere mere pokriva te, ki se pojavljajo v korpusu. Čeprav smo z opisanimi avtomatskimi postopki v slovenski wordnet dodali tudi večbesedne literale, smo se pri korpusni analizi omejili na enobesedne, saj je večbesedne literale zaradi variacij, besednega reda in pregibanja v korpusu težko identificirati.

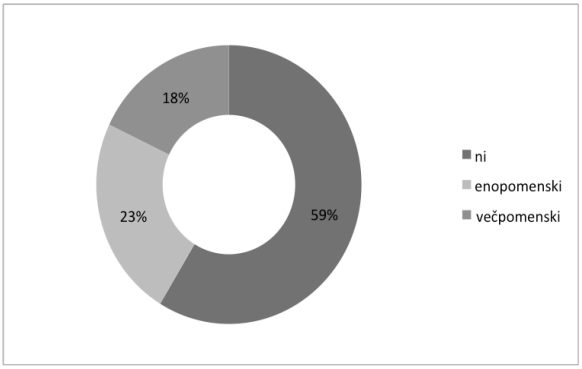
		Frekvenca			
Samostalniki	Št. pomenov	<3	<30	≥30	Skupaj
lastni	ni v wn	2.256	314	1	2.571
	enopomenski	104	43	2	149
	večpomenski	8	1	0	9
	lastni skupaj	2.368	358	3	2.729
občni	ni v wn	2.632	625	10	3.267
	enopomenski	761	530	12	1.303
	večpomenski	266	642	90	998
	občni skupaj	3.659	1.797	112	5.568
Samostalniki skupaj		6.027	2.155	115	8.297

Tabela 5. Pokritje enobesednih samostalnikov v korpusu jos100k.

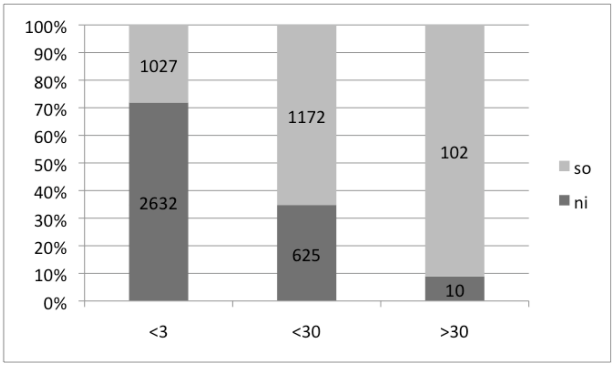
Kot prikazuje tabela 5, je v korpusu jos100k 8.297 različnih samostalniških lem; 33 % je lastnih imen, preostalo so občni samostalniki. Kot je za distribucijo besed v korpusih običajno, je 73 % samostalnikov v jos100k redkih (se pojavijo le enkrat ali dvakrat). 26 % se jih pojavi do tridesetkrat, zelo pogostih samostalnikov, ki se pojavijo več kot tridesetkrat, je malo (1 %). Izdelan wordnet vsebuje 30 % enobesednih samostalnikov iz korpusa JOS. Med njimi je le 158 lastnih imen, preostalo so občni samostalniki (2.301). Lastna imena so večinoma enopomenska, občni samostalniki pa tako eno- kot večpomenski (glej sliko 2). Jasnejšo sliko o večpomenskih literalih daje tabela 6, iz katere je razvidno, da ima 75 % večpomenskih samostalnikov po dva ali tri pomene, literali z več kot petimi pomeni so redki (6 %). Največ pomenov ima literal *položaj*, ki se pojavlja v 15 sinsetih. Pogosti samostalniki iz korpusa so v wordnetu zelo dobro zastopani. Od samostalnikov s frekvenco, višjo od 30, jih v je v wordnetu več kot 90 %, tistih s frekvenco do 30 pa dve tretjini. Redki samostalniki so najslabše zastopani, teh je v wordnetu zgolj slaba tretjina (glej sliko 3).

	Število pomenov v wordnetu						
Samostalnik	1	2	3	4	5	>5	Skupaj
lastni	149	4	2	1	2	0	158
občni	1.303	515	233	122	63	65	2.301
Skupaj	1.452	519	235	123	65	65	2.459

Tabela 6. Večpomenskost enobesednih samostalnikov v korpusu jos100k.



Slika 2. Pokritje občnih samostalnikov v korpusu jos100k glede na njihovo število pomenov v wordnetu.



Slika 3. Pokritje enobesednih samostalnikov v korpusu jos100k glede na frekvenco pojavitev.

Na koncu nas je še zanimalo, ali wordnet za tiste samostalnike iz korpusa, ki jih pokriva, vsebuje ustrezne pomene, ki bi jim jih lahko pripisali ročno ali avtomatsko. Ker so lastna imena večinoma enopomenska, smo izbrali pet večpomenskih občnih samostalnikov, ki se v korpusu pojavijo do desetkrat, in pet takšnih, ki se pojavijo več kot desetkrat in jim s pomočjo konkordanc na roke poskušali določiti pomen iz wordneta.

Tabela 7 prikazuje rezultate analize. V prvem stolpcu je seznam besed, ki smo jih analizirali, v drugem njihova frekvenca v korpusu JOS, tretji stolpec pa vsebuje število pomenov, ki jih imajo te besede v slovenskem wordnetu. Kot vidimo, imajo redkejše besede v korpusu tudi manj pomenov v wordnetu. Največ pomenov ima beseda *pot* (9). V četrtem stolpcu je število sinsetov, v katerih se iskana beseda pojavi zaradi napake v wordnetu. Pri avtomatskem generiranju wordneta je prišlo do napak pri razdvoumljanju besed *pot* in *znamka*, ki sta poleg pravilnih sinsetov pristali tudi v enem oz. dveh napačnih. Te napake bi lahko povzročale težave pri avtomatskem razdvoumljanju korpusa, zato jih je potrebno čim prej odpraviti.

Peti stolpec vsebuje število pomenov, ki se v korpusu pojavijo, v wordnetu pa jih ni, kar pomeni, da v teh primerih besedam niti ročno, kaj šele avtomatsko ne bi mogli določiti pravega pomena. Glede na to, da se ti pomeni v korpusu pojavljajo, so za slovenščino relevantni in jih je potrebno čim prej dodati v wordnet. Največ manjkajočih pomenov je za besedi *glas* in *rak*, ki jima manjkata po dva pomena, v obeh primerih gre za pomena, ki sta v korpusu pogosta, zato je luknja v wordnetu še toliko resnejša.

beseda	# JOS	# wn	# napak v wn	# lukenj v wn	# dodatnih v wn	najpog. pomen
bitje	7	2	0	1	1	ok
čelo	5	1	0	1	0	ni
prst	7	4	0	0	2	ok
rak	8	1	0	2	0	ok
zmaj	6	1	0	1	0	ni
glas	31	5	0	2	2	ni
jezik	33	6	0	1	4	ok
pot	52	9	1	0	2	ok
zemlja	11	5	0	1	2	ni
znamka	14	4	2	0	0	ok

Tabela 7. Določanje pomena izbranim večpomenskim besedam v korpusu jos100k.

Predzadnji stolpec vsebuje informacije o tem, koliko pomenov iz wordneta se v korpusu za iskane besede sploh ne pojavi, zadnji pa prikazuje, za katere besede wordnet vsebuje najpogostejši pomen iz korpusa. Kot vidimo, v wordnetu manjka kar nekaj najpogostejših pomenov besed glede na podatke, pridobljene iz korpusa: kar štirim od desetih analiziranih besed pomena glede na wordnet največkrat ni bilo mogoče določiti.

5. Sklep

V prispevku smo predstavili drugo različico slovenskega semantičnega leksikona tipa wordnet, ki je bila avtomatsko izdelana na podlagi prosto dostopnih večjezičnih virov, kot so večjezični korpusi in leksikoni. SloWNet je pod licenco Creative Commons prosto dostopen v raziskovalne namene na naslovu: http://nl.ijs.si/slownet.

Slovenski wordnet trenutno vsebuje skoraj 20.000 različnih literalov, ki pokrivajo večino osnovnih konceptov ter kar precej specifičnih, ki so bili večinoma pridobljeni iz Wikivirov in so s področja biologije. V wordnetu so zaenkrat predvsem samostalniki, poleg enobesednih literalov je precej tudi večbesednih. Najpogostejša relacija med sinseti je hipernimija, ki predstavlja 46 % vseh relacij v wordnetu. Skoraj polovica hipernimskih verig ne vsebuje vrzeli. Primerjava wordneta s korpusom jos100k je pokazala, da wordnet pokriva tretjino besedišča iz korpusa, še posebej dobro so zastopane najpogostejše besede, ki jih je v wordnetu več kot 90 %. Nekoliko slabše se wordnet odreže pri analizi pomenov večpomenskih besed v korpusu, saj jim v 40 % primerov na podlagi wordneta ni bilo mogoče določiti pravega najpogostejšega pomena. To pomanjkljivost wordneta bi bilo potrebno čim prej odpraviti.

6. Literatura

Bentivogli Luisa, Pianta Emanuele and Pianesi Fabio (2000): Coping with lexical gaps when building aligned multilingual wordnets. V: *Proceedings of the 2nd International Conference on Language Resources and Evaluation, LREC 2000*, str. 993-997. Atene, Grčija.

Bentivogli, Lisa; Pamela Forner; Bernardo Magnini in Emanuele Pianta (2004): Revising theWordNet Domains Hierarchy: Semantics, Coverage, and Balancing. V: *Proceedings of the COLING Workshop on Multilingual Linguistic Resources*, 2004.

Casado, R. M., E. Alfonseca, in P. Castells (2005): Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. V: *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, NLDB 2005*, Alicante, Španija, 15.-17. junij 2005.

Declerck, Thierry, Asunción Gómez Pérez, Ovidiu Vela, Zeno Gantner in David Manzano-Macho (2006): Multilingual Lexical Semantic Resources for Ontology Translation. V: *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*. Genova, Italija, 24.-26. maj 2006.

Diab, Mona (2004): The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. V: *Proceedings of the Arabic Language Technologies and Resources, NEMLAR 2004*, Kairo, Egipt.

Dyvik, Helge (1998): Translations as semantic mirrors. V: *Proceedings of Workshop W13: Multilinguality in the lexicon II of the 13th biennial European Conference on Artyificial Intelligence, ECAI 1998*, str. 24-44, Brighton, Velika Britanija.

Erjavec, Tomaž in Simon Krek (2008): The JOS morphosyntactically tagged corpus of Slovene. V: *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*. Marakeš, Maroko, 28.-30. maj 2008.

Erjavec, Tomaž; Fišer, Darja (2006): Building Slovene WordNet, V: *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*. Genova, Italija, 24.-26. maj 2006.

Fellbaum, Christine (ur.) (1998): WordNet. *An Electronic Lexical Database*. MIT Press, Cambridge, Massachusetts.

Fišer, Darja, in Benoît Sagot (2008): Combining multiple resources to build reliable wordnets. V: *Proceedings of the 11th Conference on Text, Speech and Dialog, TSD 2008*. Brno, Češka, 8.-12. September 2008.

Horak, Aleš; Pavel Smrž (2000): New Features of Wordnet Editor VisDic, V: Dascalu, Dan (ur.): *Romanian Journal of Information Science and Technology Special Issue* 7/1–2.

Ide, Nancy; Erjavec, T. in Tufis, D. (2002): Sense Discrimination with Parallel Corpora, V: *Proceedings of ACL'02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, str. 54-60, Philadelphia, ZDA.

Knight, Kevin in S.K. Luk (1994): Building a Large-Scale Knowledge Base for Machine Translation. V: *Proceedings of the 12th National Conference on Artificial Intelligence, AAAI*, zvezek 1, str. 773-778. AAAI Press, Seattle, ZDA.

Orav, Heili in Kadri Vider (2004): Concerning the Difference Between a Conception and its Application in the Case of the Estonian WordNet. V: *Proceedings of the 2nd Global WordNet Conference*, str. 285-290. Brno, Češka.

Resnik, Philip in David Yarowsky (1997): A perspective on word sense disambiguation methods and their evaluation. V: *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?*, Washington D.C., ZDA.

Rigau, German; Rodríguez H. in Agirre E. (1998): Building Accurate Semantic Taxonomies from Monolingual MRDs. V: *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, COLING-ACL 1998*. Montreal, Kanada.

Steinberger, Ralf; Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis in Daniel Varga (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. V: *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006,* str. 2142-2147, Genova, Italija, 24.-26. maj 2006.

Tiedemann, Jörg (2003): *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. Doktorska disertacija, Studia Linguistica Upsaliensia 1.

Tufis in dr. (2008): Language models from SEE-ERA.NET corpus. V: *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages, FASSBL 2008*. Dubrovnik, Hrvaška 25.-28. september 2008.

Tufis, Dan; Dan Cristea in Sofia Stamou (2000): BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. V: Dascalu, Dan (ur.): *Romanian Journal of Information Science and Technology Special Issue*. 7/1-2, 9-43.

van der Plas, Lonneke in Jörg Tiedemann (2006): Finding Synonyms Using Automatic Word Alignment and Measures of Distributional Similarity. V: *Proceedings of ACL/COLING 2006*, Syndey, Avstralija.

Vossen, Piek (ur.) (1998). *EuroWordNet : A multilingual database with lexical semantic networks.* Kluwer Academic Press, Dordrecht.

Wong, Shun Ha Sylvia (2004): Fighting arbitrariness in WordNet-like lexical databases - A natural language motivated remedy. V: *Proceedings of the Second Global WordNet Conference 2004*, str. 234-241. Brno, Češka.

Yao, Qing; QingXian Wang; Yan Liu; Qiang Wang in JunYong Luo (2007): A New Methodology for Building Ontology Based on Reusing the Heterogeneous Ontologies. V: *Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007*, zvezek 1, str. 717-721. Haikou, Kitajska, 24. – 27. avgust 2007.

Samodejno luščenje slovarja iz vzporednega korpusa s pomočjo vmesnega jezika in pomenskega razdvoumljanja

Peter Holozan

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

Povzetek

Težava pri nestatističnem strojnem prevajanju je, da je veliko dela z ročnim vnosom slovarja. Zato sem poskusil iz vzporednega korpusa samodejno izluščiti slovar, in sicer z uporabo analizatorja za prevedbo v vmesni jezik, pri čemer se uporabi tudi pomensko razdvoumljanje analizatorja. Opisane so nekatere težave, ki so se pri tem pojavile, in načini njihovega reševanja. Predstavljen je tudi primer izluščenega slovarja.

Automatic lexicon extraction from parallel corpus using Interlingua and sense disambiguation

The problem of non-statistical machine translation is huge amount of manual work needed to build a dictionary. That was the reason to try automatically to extract the lexicon from parallel corpus with help of an analyser for translation to Interlingua using sense disambiguation in a process. Some encountered problems are described together with possible solutions. A sample of extracted lexicon is provided.

1 Uvod

Velika prednost statističnega strojnega prevajanja pred prevajanjem s pravili je, da ga je mogoče mnogo hitreje prilagoditi na nove jezikovne pare, če le imamo ustrezne vzporedne korpuse. Pri nestatističnem strojnem prevajanju je glavna ovira izdelava slovarja, kar tipično zahteva veliko ročnega dela (Rosner, 2004). Po drugi strani pa nestatističnem strojnem prevajanju laže dodajati nova slovnična pravila in besede, ki so preveč redke, da bi se dovoljkrat pojavljale v korpusih. Vmesni jezik pa je mogoče uporabiti tudi za pomensko analizo besedilo, kar je lahko prvi korak pri izdelavi povzetkov, odgovarjanju na vprašanja ipd.

Opisan bo poskus, kako čim hitreje razširiti strojni prevajalnik Presis[1], ki za zdaj prevaja med slovenščino in angleščino, še na druge jezikovne pare, in to s čim manj ročnega dela.

Najprej bodo v razdelku 2 na kratko predstavljeni nekateri dosedanji poskusi luščenja leksikona iz vzporednih korpusov, čemur bo v razdelku 3 sledil kratek opis delovanja prevajalnika Presis in načina zapisa slovarja, ki ga potrebuje za delovanje.

V razdelku 4 bo na primerih opisana predlagana metoda luščenja, v 4.2 pa bodo opisane nekatere težave, ki so se pojavile pri preizkušanju v praksi in uporabljene rešitve.

Na koncu bosta v razdelku 5 predstavljena rezultata na primeru nemško-slovenskega in angleško-francoskega vzporednega korpusa.

Dodano bo nekaj idej za dodatno izboljšanje postopka.

2 Samodejno luščenje leksikonov

Poskusi luščenja slovarjev iz vzporednih korpusov so bili že v devetdesetih letih prejšnjega stoletja (Tiedemann, 1997). Luščenje leksikonov je sestavni del pri metodah statističnega strojnega prevajanja. Pri dovolj velikih korpusih dosegajo te metode zelo dobre rezultate.

Največkrat omejujejo na statistiko samih besednih oblik (čeprav so tudi poskusi z lematiziranim vhodom (Vintar, 2003)), kar povzroča težave pri pregibnih jezikih. Druga težava je primerjava jezikov z zelo drugačnim besednim redom (npr. nemščine z angleščino (Langlais et al., 2005)).

3 Prevajalnik sistem Presis

Strojni prevajalnik Presis temelji na jezikovni zbirki Ases. V njej so shranjene besede (skupaj z vsemi oblikami), zveze (dveh ali več besed), skupine (po pomenu tesno povezane skupine besed, na primer v slovenščini pri osebah obliki za moški in ženski spol ter njuna svojilna pridevnika) in predloge (s podatki o vezljivosti glagolov), vse pa povezujejo pomeni, prek katerih potem poteka prevajanje. Za razdvoumljanje so dodane tudi povezave med pomeni (katere kombinacije pomenov se tipično pojavljajo in katere so tipično prepovedane). Podatki so v Ases vneseni ročno, pri vnosu pa se pomaga tudi z raznimi spiski in statistikami, pridobljenimi iz korpusov.

Presis vhodne stavke najprej s pomočjo analizatorja prevede v vmesni jezik, ki ga potem generator prevede v izhodni jezik. Če analizator ne uspe analizirati stavka, se uporabijo pomožne metode prevajanja (na koncu tudi dobesedno prevajanje besedo za besedo).

Analizator je napisan splošno za vse jezike, posebnosti posameznih jezikov rešuje z nastavitvami o značilnostih posameznih jezikov. Zato ga je relativno preprosto prilagoditi na nov jezik, še posebej v primeru, da je podoben kateremu od že obstoječih jezikov (v nasprotnem primeru je treba uvesti nove značilnosti, kar zahteva več dela).

Ker prevajanje poteka preko jezikovno neodvisnih pomenov, je za katero koli kombinacijo jezikov dovolj, da sta oba jezika ustrezno povezana na pomene in niso potrebne povezave med vsemi kombinacijami jezikov.

Tudi pomenske povezave so jezikovno neodvisne, vendar so v različnih jezikih dvoumnosti na različnih mestih, zato jih je treba za dodani jezik dopolniti.

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

3.1 Presisov vmesni jezik

Poved *Peter je kosilo.* se v vmesni jezik prevede takole (za večjo preglednost so dodane vrstice in zamiki pri elementih):

```
(-POV:
 (-STAg-n-psdvt-------:
  (1OSB:
   (-SFR:
    (-DSF:
     (-JED:
      (-SAMe:{3d0f9b;fbb21a}[0]<26c>)))))),
  (0PVD:
   (-GGL:{210074;8a3abf}[1]<618>)),
  (*PR4:
   (-SFR:
    (-DSF:
     (-JED:
      (-SAMe:{9e7d;a520c0}[2]<28>)))))),
  (-LOCk:[3])))
```

Primer 1. Primer vmesnega jezika

Elementi so urejeni hierarhično, in sicer z oklepaji. Imena elementov so označena s tremi velikimi črkami, ki jim lahko sledijo še parametri (male črke). V zavitih oklepajih je najprej oznaka (oznake se zapisane šestnajstiško) leme, ki ji sledi seznam oznak možnih pomenov. V oglatih oklepajih je napisana zaporedna številka besede v povedi, na katero se element nanaša. V lomljenih oklepajih je oznaka kode MSD (uporabljene so kode, ki so uporabljene v korpusu Fida+ (Arhar, 2007), dodane pa so nekatere dodatne informacije).

Zgornji primer je tako stavek, sestavljen iz osebka, povedka in predmeta v tožilniku. Rezultat analizator je sicer seznam možnih analiz, tukaj je le analiza, ki jo je analizator izbral za najbolj verjetno.

oznaka	pomen
POV	povedek
STA	stavek
OSB	osebek
PR4	predmet v tožilniku
PR3	predmet v dajalniku
PRD	neposredni predmet
PRI	posredni predmet
PVD	povedek
PDO	prislovno določilo
LOC	ločilo
SFR	samostalniška fraza
DSF	del samostalniške fraze
PFR	pridevniška fraza
DPF	del pridevniške fraze
SVZ	svojilni zaimek
JED	jedro samostalniške fraze
SAM	samostalnik
PRV	pridevnik
CLN	člen
OSZ	osebni zaimek
NOZ	navidezni osebni zaimek
GGL	glavni glagol
GPO	pomožni glagol
PRF	prislovna fraza
PRS	prislov

Tabela 1: Nekateri elementi vmesnega jezika

Pred oznakami so lahko številke, ki pomenijo oznake delov glagolske predloge. Te oznake omogočajo, da se npr. osebek lahko prevede v predmet ali nasprotno.

V vmesnem jeziku so ob pomenih zapisane tudi oznake lem, tako da lahko v primeru, če besede še nimajo pripisanih pomenov, prevaja tudi s pomočjo pomožnega slovarja, kjer so vhodnim lemam pripisane prevodne leme (ta pomožni slovar se uporablja tudi pri uporabniškem dodajanju besed). In ta podatek uporablja tudi postopek luščenja.

4 Luščenje s pomočjo vmesnega jezika

Za svoje delovanje Presis torej potrebuje povezave besed na pomene in ne le seznam možnih prevodov besede. Dodatna težava samodejnega luščenja so jeziki s pregibanjem in z relativno prostim besednim redom (npr. slovenščina ali nemščina).

4.1 Ideja postopka

Ideja je postopka je, da se pri povezovanju besed na pomene izrabi vmesni jezik, ki ga naredi analizator. Ker analizator hkrati tudi razdvoumlja, je s tem avtomatsko rešeno razdvoumljanje. Hkrati je s tem rešeno tudi pregibanje besed in besedni red v stavku.

Ročno so bili določeni deli stavkov, ki jih je med seboj možno izenačiti, sama primerjava pa potem poteka avtomatsko.

```
Allumez votre ordinateur portable.
Switch on your laptop.
```

Primer 2. Primer usklajenega para francoskega in angleškega stavka

Presisov vmesni jezik za ta primer je naslednji:

```
(-POV:(=STAg-n-v--vt-------:(0PVD:(-
GGL:{5df4;}[0]<14a0>)),(2PRD:(-SFR:(-DSF:(-PFR:(-
DPF:(-SVZ-m-d:{d0c3c;49fad}[1]<c1ac>))),(-JED:(-
SAMe:{88ffd;}[2]<7d0>)),(-PFR:(-DPF:(-
PRVo:{95750;}[3]<758>)))))),(-LOCk:[4]),(1OSB:(-
SFR:(-DSF:(-JED:(-NOZnm-t:[0])))))))

(-POV:(-STAg-n-v--vt-------:(0PVD:(-
GGL:{41a978;2674bc}[0]<14>)),(*PVD:(-
GPT:{411ea;6ddb96}[1]<f0>)),(2PRD:(-SFR:(-DSF:(-
PFR:(-DPF:(-SVZpm-d:{4a0e8f;9c86f0}[2]<ac8>))),(-
JED:(-SAMe:{26d4f9;589f94}[3]<a0>))))),(-
LOCk:[4]),(1OSB:(-SFR:(-DSF:(-JED:(-NOZnm-
d:[0])))))))
```

Primer 3. Prevod v vmesni jezik za usklajeni par v francoščini in angleščini

V tem primeru je možno izenačiti povedka (*0PVD*) in neposredna predmeta (*2PRD*). Pri predmetu najprej izluščimo leva prilastka, ki sta svojilna zaimka, na koncu pa je treba izenačiti frazo *ordinateur portable* (*88ffd + 95750*) in **pomen** besede *laptop* (ki je *prenosni računalnik* (*589f94*)).

Den Hund jagt der Hase. Zajec lovi psa.

Primer 4. Nemško-slovenski primer

Tukaj je v nemškem stavku najprej predmet in na koncu osebek, v slovenščini pa nasprotno (bolj dosleden prevod bi sicer bil *Psa lovi zajec.*).

(-POV:(=STAg-n-ps-vt-------:(2PR4:(-SFR:(-DSF:(-CLNd:[0]),(-JED:(-SAMe:{1007d0;}[1]<50>))))),(0PVD:(-GGL:{bd5f8;}[2]<49c>)),(1OSB:(-SFR:(-DSF:(-CLNd:[3]),(-JED:(-SAMe:{e9ecc;}[4]<3c>))))),(-LOCk:[5])))

(-POV:(-STAg-n-psdvt-------:(1OSB:(-SFR:(-DSF:(-JED:(-SAMe:{ed58;e84d35,30a40d,e84e4f,e84f23}[0]<3c>))))),(0PVD:(-GGL:{257873;783b95}[1]<618>)),(2PR4:(-SFR:(-DSF:(-JED:(-SAMe:{22deb;9f219a,b59}[2]<230>))))),(-LOCk:[3])))

Primer 5. Vmesni jezik za nemško-slovenski primer

Lahko izenačimo osebka (*1OSB*) in predmeta v tožilniku (*2PR4*) ter povedka (*0PVD*). Prost besedni red torej ne pomeni nobene težave pri luščenju.

Težava je lahko le pri večjem številu prislovih določil, kjer v primeru, da analizator še ne zna ugotoviti njihove vrste, ne moremo vedeti, kako jih povezati med sabo.

4.2 Opažene težave

Pri preizkušanju postopka so se pokazale nekatere težave.

4.2.1 Neujemanje strukture

Marsikateri stavek se da analizirati na različne načine. Ta problem načeloma odpravi razdvoumljanje, vendar se razdvoumljanje v veliki meri zanaša na odnose med pomeni in ker pri besedah, katerih pomene iščemo, ti še niso znani, se lahko zgodi, da analizator izbere napačno možnost. Druga možnost so napake v vhodnih podatkih, na primer to, da povedi v vzporednih korpusih niso pravilno poravnane ali pa so stavki napačno prevedeni.

Zato je smiselno, da v primeru, da sta analizi preveč različni, luščenje preskočimo. Druga možnost bi bila, da med vsemi najdenimi analizami pri ciljnem jeziku izberemo prvo, ki ustreza referenčni analizi.

Neujemanje struktur lahko kaže tudi na različnost glagolskih predlog, kar bo smiselno upoštevati pri luščenju le-teh.

4.2.2 Zanikanje izvornega stavka

Občasno se zgodi, da je ciljni stavek nikalni, referenčni pa trdilni (ali pa ravno obratno).

tako, da podatki niso zakriti so daß die Angaben lesbar sind Localized heat treatment is not permitted. Un traitement thermique local est interdit.

Primer 6. Para trdilnega in nikalnega stavka

Ena možnost je, da se v takih primerih stavki izločijo iz luščenja.

Če gre za povedkovo določilo, bi si bilo mogoče pomagati tudi s podatki o protipomenkah, ki jih vsebuje Ases. V takih primerih bi se pridevnik namesto na osnovni pomen vezal na protipomenko (če je ta seveda vpisana v Asesu).

V nekaterih primerih nikalnost odraža na morfološkoglagolski ravni (zanikanost je morfološko del glagola).

ki ne more prisostvovati das verhindert ist This category does not include freezer trawlers. Cette catégorie exclut tout chalutier congélateur.

Primer 7. Para z glagolom, ki vsebuje zanikanje

Take primere bi bilo treba upoštevati pri luščenju glagolskih predlog.

4.2.3 Napake pri razdvoumljanju

Do te težave pride, če analizator napačno razdvoumi referenčno besedilo in da na prvo mesto napačen pomen, na katerega se potem poveže beseda.

Problem je zelo pereč, ker je popolno razdvoumljanje zelo zapleten problem. Kljub temu pa ni popolnoma nerešljiv.

Precej ga omili dejstvo, da je predlagani postopek predviden predvsem kot zagonsko polnjenje podatkovne zbirke z novim jezikom, ki se bo potem predvidoma dopolnjevala naprej ročno, pri čemer se bodo take napake lahko opazile in odpravile.

Druga možna rešitev pa je, da se namesto enega referenčnega jezika uporabita dva ali celo več (Cohn in Lapata, 2007). Tako bi bili za luščenje francoskega slovarja uporabljeni ob francoščini hkrati slovenščina in angleščina, postopek pa bi preverjal, ali se rezultata slovenske in angleške analize ujemata. V primeru, da se pomena ne ujemata, bi preskočil luščenje iz stavka, hkrati pa bi bil seznam takih parov stavkov zelo koristen pripomoček za dopolnjevanje razdvoumljanja v slovenščini in angleščini.

4.2.4 Manjkajoče fraze

Predvsem pri luščenju nemškega slovarja se lahko pogosto zgodi, da se ena nemška beseda poveže na zvezo pridevnika in samostalnika, ki pa sta vnesena le kot samostojna pomena, ne pa tudi kot zveza. Razlog za to je, da sta sklapljanje in zlaganje kot besedotvorna postopka v nemščini izredno pogosti.

Pokazalo se je, da je spisek manjkajočih fraz lahko zelo uporaben (glej 6.1).

4.2.5 Mešanje pridevniških in samostalniških prilastkov

Zgodi se, da se pridevniški prilastek v drugem jeziku spremeni v samostalniški prilastek.

Pierrov klobuk le chapeau de Pierre

Primer 8. Par z različnima prilastkoma

Če gre za sistemski pojav (tipičen primer so svojilni pridevniki), je smiselno, da to rešuje že analizator in zgornji primer že analizira tako, da se vidi, da gre za svojilni pridevnik.

V nasprotnem primeru je začasna rešitev, da se taki pari izločijo iz luščenja.

4.2.6 Mešanje trpnika in tvornika

V nekaterih jezikih se trpnik bolj uporablja kot v drugih, zato v korpusu lahko naletimo na pare, kjer je v enem od jezikov trpnik, v drugem pa je tvornik.

Organ predstavlja izvršilni direktor. L'Autorité est représentée par son directeur exécutif.

Primer 9. Par tvornika in trpnika

V takih primerih se ne sme primerjati osebka z osebkom, ampak osebek trpnika s predmetom tvornika, osebek tvornika pa se izraža v prislovnem določilu.

Tudi Presis uporablja tako pretvorbo pri prevajanju iz angleščine v slovenščino (stavek *A book was read by Peter.* tako prevede v *Peter je bral knjigo.* in ne dobesedno v *Knjiga je bila brana od Petra.*), zato je vmesni jezik zgrajen tako, da je ta pretvorba dovolj preprosta.

4.2.7 Dodatni zaimki

V referenčnem ali pa tudi ciljnem besedilu se občasno pojavijo dodatne besede, in sicer predvsem zaimki.

Ta direktiva je naslovljena na države članice. Način sklicevanja določijo države članice. Diese Richtlinie ist an alle Mitgliedstaaten gerichtet. Die Einzelheiten dieser Bezugnahme regeln die Mitgliedstaaten.

Primer 10. Dodatni zaimki v nemškem prevodu

Zgornja primera tako vplivata na rezultat v Tabeli 3. Smiselno bi bilo, da bi luščenje avtomatsko izločalo kazalne in svojilne zaimke iz luščenja. Celostni zaimki so že problematični in bi bilo nanje najbrž bolje le opozoriti, da se zveze preverijo ročno.

4.2.8 Spremenjen vrstni red stavkov

Pojavilo se je kar nekaj primerov, ko je bil spremenjen vrstni red pogojnega odvisnika in glavnega stavka v povedi.

Če ob izteku roka takšnega ugovora ni, nosilec v kraju stalnega prebivališča nudi storitve.
DER TRÄGER DES WOHNORTS GEWÄHRT DIE SACHLEISTUNGEN, WENN ER BIS ZUM ABLAUF DIESER FRIST KEINEN ABLEHNENDEN BESCHEID ERHALTEN HAT.

Primer 11. Spremenjen vrstni red stavkov

Iz tega primera sledi, da je smiselno delati luščenje na ravni povedi in ne le stavkov, in sicer tako, da se najprej uparijo ustrezni odvisniki oz. glavni stavki.

4.3 Pomanjkljivosti metode

Problematično je dejstvo, da je potrebno pred uporabo narediti slovar za želeni jezik, v katerem imajo besede pripisane oznake MSD in leme. Ta del za zdaj poteka bolj ali manj ročno, zato bi bilo smiselno razmisliti o njegovi avtomatizaciji (npr. v smeri (Vičič, 2008)).

Drugi del je izdelava slovničnih pravil za analizator za želeni jezik, vendar je ta del tako potreben tudi za sam prevajalnik, analizator pa je tudi že v osnovi zastavljen večjezično. Možno je tudi postopno dodajanje pravil, da analizator najprej analizira preprostejše povedi, s čimer se lahko izlušči osnovni slovar in zgradi preizkusni prevajalnik, s katerim se v nadaljevanju lahko preizkuša, kaj je treba še dopolniti.

4.4 Kako naprej

Ob že prej omenjenih možnih izboljšavah je ena najpomembnejših možnih razširitev luščenje glagolskih predlog. To luščenje bo moralo potekati fazno, ker bo treba imeti pomene osebkov in predmetov določene na obeh straneh, da se bo dalo zanesljivo povedati, kaj se s čim povezuje v primerih, ko se ne povežeta preprosto osebek z osebkom in predmet s predmetom. Tako bo v prvi fazi smiselno izluščiti pomene samostalnikov, pridevnikov, prislovov in predlogov; potem pa v drugi fazi uporabiti te znane pomene za določanje predlog, ki bodo v naslednji fazi spet pripomogle k boljšemu luščenju pomenov drugih besednih vrst.

5 Rezultati

Za preizkušanje sem uporabil *The DGT Multilingual Translation Memory of the Acquis Communautaire: DGT-TM*. Zbirka vsebuje evropsko zakonodajo v 22 jezikih v formatu za pomnilnik prevodov, torej že poravnano po stavkih.

Barve ozadja vrstic pomenijo naslednje:

pomen
pravilno izluščena povezava na pomen
ni najpogostejši pomen
napačen pomen
odvečen zaimek pri zvezi
napačno razdvoumljen pomen

Tabela 2: Pomen ozadij

frk.	pomen	nemški prevod
4337	komisija	Kommission
2808	ta	dieser
2530	država članica	Mitgliedstaat

1539	za (kraj)%j	für
1456	član	Mitglied
1384	za (splošno)%n	in (d)
1127	predsednik (zborovanja, ustanove ...)	Präsident
902	uredba	Verordnung
859	smernica	Richtlinie
762	ob upoštevanju	auf (t)
621	odločba	Entscheidung
577	mnenje	Stellungnahme
472	parlament	Parlament
456	dan (čas štiriindvajsetih ur)%c	Tag
438	ukrep	Maßnahme
436	ob upoštevanju	nach
410	način	Einzelheit
401	splošen	allgemein
387	naslednji%c	folgend
374	za (splošno)%n	für
343	določba	Bestimmung
324	skupnost	Gemeinschaft
298	postopek (oblika načrtnega, premišljenega dela)	Verfahren
291	sklicevanje	Bezugnahme
259	ime	Name
259	na (časovna opredelitev)%c	an (d)
257	priloga (časopis)	Anhang
248	odbor	Ausschuss
246	pri (bližina)%j	bei
245	vsak%c	jede (p)
225	po (kraj)%j	nach
220	z (način, orodnik)%n	mit
218	država članica	alle (p) + Mitgliedstaat
217	ves	alle (p)
209	z (čas)%c	an (d)
196	informacija	Information
187	primer (konkreten)	Fall
177	tak	dieser
174	podjetje	Unternehmen
168	pristojni organ	zuständig + Behörde
167	cilj (kar se hoče doseči s prizadevanjem)	Ziel
154	zahteva	Anforderung
153	načelo (kar kdo sprejme, določi za usmerjanje svojega ravnanja, mišljenja)	Grundsatz
149	končna določba	Schlussbestimmung
148	odbor	Ausschuß
147	proizvajalec	Hersteller
147	podatek	Angabe
146	sodelovanje	Zusammenarbeit
143	sprememba	Änderung
143	rezultat	Ergebnis
142	drug	andere
141	drug	sonstig
140	opis	Beschreibung
138	industrija	Wirtschaftszweig
137	vrsta (zvrst)	Art
136	proizvod	Erzeugnis
134	poseben	besondere
134	sklicevanje	dieser + Bezugnahme
131	svet	Rat
130	pogoj	Bedingung
129	seznam	Liste
129	poročilo	Bericht
129	člen (oštevilčen odstavek zakona ali določbe)	Artikel
127	tehničen	technisch
...	...	...
1	carinski	zuständig
1	carinski	periodisch
1	carinski	dieser
1	carinski	betreffend
1	carinski	öffentlich
1	brezpogojen	unentgeltlich
1	brezobresten	zinslos
1	brezalkoholen	alkoholfrei
1	brezžičen	schnurlos
1	bolnišničen	stationär
1	bolan	krank
1	biološki	genetisch
1	bazičen (kemija)	basisch
1	barven	farbig
1	balističen	ballistisch

Tabela 3: Primer leksikona, izluščenega iz nemško-slovenskega vzporednega korpusa

V rezultat je vključenih nekaj zadetkov z začetka in nekaj s konca seznama. Ideja je, da so pogostejši zadetki bolj verjetno pravilni kot redkejši, ker prihaja do več nenavadnih rezultatov. Vendar se je pokazalo, da je tudi pri zadetkih, ki se pojavljajo samo po enkrat, zelo veliko pravilnih (tudi na splošno se v tipičnem korpusu velik del besed pojavi samo po enkrat), zato bi jih bilo škoda izgubiti s postavljanjem fiksne pogostosti, do katere je rezultate smiselno uporabiti. Najbrž bi bilo smiselno razmišljati v smeri, da se pri vsaki besedi uporabi prvih nekaj najdenih pomenov (morda dokler skupna pogostost najdenih pomenov ne doseže 80 % vse pogostosti neke besede; to idejo bi bilo treba še preizkusiti v praksi).

Poskus je pokazal nekaj zelo pogostih težav pri razdvoumljanju v slovenščini. Urediti je treba tudi izločanje odvečnih zaimkov pri zvezah.

Za primere, kjer niso izbrani najpogostejši pomeni, je vzrok predvsem vhodni korpus, v katerem zelo prevladujejo pravna besedila. S korpusi z različnih področij bi se to dalo rešiti, vendar velikih vzporednih korpusov ni lahko dobiti. Prednost predlagane metode pa je to, da se besede vežejo na pomene, kar pomeni, da se npr. za izdelavo slovensko-francoskega slovarja da uporabiti tudi angleško-francoski vzporedni korpus, ki ga je bistveno laže najti kot slovensko-francoskega. Rezultat luščenja iz angleško-francoskega leksikona je tako naslednji:

frk.	pomen	francoski prevod
6138	za (splošno)%n	par
3375	komisija	commission
2972	ob upoštevanju	vu
2966	predsednik (države, republike)	président
2835	svet (organ)	conseil
2265	predlog	proposition
860	aneks	annexe
707	ta	ce
626	določba	disposition
584	splošen	général
577	definicija	définition
571	za (splošno)%n	pour
560	predpis	règlement
478	drug	autre
431	za (splošno)%n	à
357	primer (konkreten)	cas
330	končna določba	disposition + final
250	sodelovanje	coopération
246	vstop	entrée
234	v (čas)%c	en
227	parlament	parlement
211	tisti	ce
199	načelo (kar kdo sprejme, določi za usmerjanje svojega ravnanja, mišljenja)	principe
191	postopek (oblika načrtnega, premišljenega dela)	procédure
190	k (smer)%a	à
187	Francija (+P)	France
185	izdelek	produit
184	dodajanje	outre
183	naslednji%c	suivant

Tabela 4: Primer leksikona, izluščenega iz francosko-angleškega vzporednega korpusa

5.1 Stranski rezultat

Kot stranski rezultat luščenja med nemščino in slovenščino je nastal seznam slovenskih večbesednih fraz, ki so se prevajale v eno nemško besedo (in še niso bile vnesene kot zveze v Asesu).

fraza, ki je v nemščino prevedena z eno besedo
uradni list
normalna vrednost
ta sklep
carinski organi
izvozna cena
ta predpis
zadevni izdelek
vsaka država članica
v celoti
neto teža
tretje države

Tabela 5: Začetni del potencialnih zvez

Napake (označene s sivim ozadjem) so predvsem pri dodanih zaimkih, kar je hitro rešljivo, drugače pa je seznam uporaben pripomoček, ki pove, katere slovenske zveze (in potem ustrezne pomene) je smiselni vnesti v Ases.

Ob tem, da takih zvez pogosto ni smiselno dobesedno prevajati (npr. *uradni list* v angleški *official leaf*) in se zato z vnosom izboljša tudi prevajanje med slovenščino in angleščino (pri čemer se lahko angleški prevodi poiščejo tudi z luščenjem iz angleško-slovenskega vzporednega korpusa, dovolj je torej, da se vnesejo slovenske zveze in pomeni), vnesene zveze tudi olajšajo delo analizatorju oz. razdvoumljanju, saj avtomatsko dobijo prednost pred kombinacijami brez takih zvez.

6 Literatura

Arhar, Špela, 2007. *Kaj početi z referenčnim korpusom Fidaplus*, Priloga A. Filozofska fakulteta, UL.

Cohn, Tevor and Lapata, Mirella, 2007. Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.* 348-355.

Giguet, Emmanuel and Luquet, Pierre-Sylvain, 2006. Multilingual Lexical Database Generation From Parallel Texts In 20 European Languages With Endogenous Resources. In *Proceedings of the COLING-ACL 2006 Main Conference Poster Sessions.* 271-278.

Langlais, Philippe, Cao, Guihong, Gotti, Fabrizio, 2005. RALI: SMT shared task system description. In *Proceedings of the ACL Workshop on Building and Using Parellel Texts.* 137-140.

Rosner, Michael, 2004. Corpus-Driven Bilingual Lexicon Extraction. In *CSAW '04 Proceedings.* 78-85.

Tiedemann, Jörg, 1997. *Automatical Lexicon Extraction from Aligned Bilingual Corpora.* Diplomska naloga, University of Magdeburg.

Vičič, Jernej, 2008. Samodejno luščenje leksikona za visoko pregibne jezike. *4. mednarodni sestanek Sodobne jezikovne tehnologije v medkulturni komunikaciji.* Koper.

Vintar, Špela, 2003. Extracting terms and terminological collocations from the ELAN Slovene-English parallel corpus. *SLLT, Volume 12, December 2003*., 48-58.

Oblikoskladenjske specifikacije in označeni korpusi JOS

Tomaž Erjavec, Simon Krek

Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si, simon.krek@ijs.si

Povzetek

Jezikovne vire JOS trenutno sestavljajo oblikoslovne specifikacije in dva korpusa. Prvi korpus je "jos100k", enojezični vzorčeni in uravnoteženi korpus slovenskega jezika s 100.000 besedami in z ročno označenimi oz. pregledanimi lemami ter oblikoskladenjskimi oznakami. Drugi je "jos1M", enomilijonski delno ročno pregledani korpus. Oba korpusa sta bila vzorčena iz 620-milijonskega korpusa FidaPLUS. Jezikovni viri JOS so označeni v skladu s označevalnimi standardi, oblikoskladenjske specifikacije skladno s sistemom MULTEXT-East, tako specifikacije kot korpusa pa skladno z navodili združenja Text Encoding Initative (Guidelines P5). Vsi viri so na voljo kot zbirka podatkov za raziskovalne namene po licenci Creative Commons in so namenjeni razvoju jezikovnih tehnologij za slovenski jezik.

The JOS Language Resources: Morphosyntactic Specifications and Annotated Corpora

The JOS morphosyntactic resources for Slovene consist of the specifications and two corpora: jos100k, a 100,000 word balanced monolingual sampled corpus annotated with hand validated morphosyntactic descriptions (MSDs) and lemmas, and jos1M, the 1 million word partially hand validated corpus. The two corpora have been sampled from the 620 million word Slovene reference corpus FidaPLUS. The JOS resources have a standardised encoding, with the MULTEXT-East-type morphosyntactic specifications and the corpora encoded according to the Text Encoding Initiative Guidelines P5. JOS resources are available as a dataset for research under the Creative Commons licence and are meant to facilitate developments of HLT for Slovene.

1. Uvod

Jezikoslovno označeni korpusi so osnovni vir za jezikovne tehnologije, vendar za mnoge jezike še niso na voljo, predvsem kot zaključene podatkovne zbirke. Eden od pomembnih virov so ročno oblikoskladenjsko označeni korpusi, ki so potrebni za učenje oblikoskladenjskih označevalnikov (part-of-speech taggers), ki so sami del osnovne jezikovnotehnološke infrastrukture za nek jezik.

Jezikovni viri za slovenščino, razviti v okviru projekta MULTEXT-East[1] (1995-1997), so vsebovali doslej edini prosto dostopni ročno označeni korpus – slovenski prevod romana "1984" Georgea Orwella. Nabor oznak ter način označevanja korpusov po sistemu MULTEXT-East je bil kasneje uporabljen pri številnih slovenskih korpusih, med drugim pri referenčnem korpusu FIDA (Erjavec in dr., 1998) in FidaPLUS (Arhar in Gorjanc, 2007). Slabost korpusa "1984" pa je njegova velikost (100.000 besed) in predvsem dejstvo, da vsebuje le en preveden roman, kar pomeni, da predstavlja razmeroma skromen vir za učenje oblikoskladenjskih označevalnikov. Poleg tega je večletna raba slovenskega nabora oznak MULTEXT-East pokazala, da bi bile pri naboru potrebne določene spremembe.

Projekt JOS skuša zapolniti vrzel pri jezikovnih virih za slovenščino z izdelavo standardiziranih prosto dostopnih označenih korpusov, skupaj z revidiranim naborom oblikoskladenjskih specifikacij. V članku poročamo o prvih rezultatih: korpusu "jos100k", ki predstavlja zlati standard za označevanje, in korpusu "jos1M", na katerem trenutno poteka delo. Oba korpusa vsebujeta vzorčene odstavke iz korpusa FidaPLUS in sta označena z razdvoumljenimi in ročno preverjenimi lemami in oblikoskladenjskimi oznakami.

[1] http://nl.ijs.si/ME/

2. Vzorčenje korpusa FidaPLUS

Korpus FIDA[2] (Erjavec et al., 1998) je 100-milijonski referenčni korpus sodobne slovenščine in vsebuje besedila, nastala med leti 1990-2000. Označen je v formatu SGML in v skladu s priporočili združenja Text Encoding Initiative TEI P3 (Sperberg-McQueen in Burnard, 1999). Posameznim pojavnicam je pripisan podatek o lemi ter oblikoskladenjski oznaki po sistemu MULTEXT-East (Erjavec, 2004), vendar v primeru večih možnih lem ali oznak te niso razdvoumljene.

Korpus FidaPLUS[3] (Arhar in Gorjanc, 2007) je bistveno večji (620 milijonov besed) ter vsebuje besedila, nastala med leti 1990-2006. Za razliko od korpusa FIDA so leme in oblikoskladenjske oznake v korpusu FidaPLUS razdvoumljene, celoten postopek procesiranja – pretvorba iz izvornega formata, tokenizacija, oblikoskladenjsko označevanje in razdvoumljanje – je bilo izvedeno z programskimi orodji v lasti jezikovnotehnološkega podjetja Amebis.[4] Korpus je prosto dostopen za raziskovalne namene, kot podatkovna zbirka pa je dostopen le članom konzorcija (Amebis, DZS, Univerza v Ljubljani, Univerza v Mariboru, Institut Jožef Stefan), za dostop je namreč potreben podpis konzorcijske pogodbe. Druga omejitev pri uporabi korpusa FidaPLUS za namen učnega korpusa je dejstvo, da je bil v celoti avtomatično označen. Označevalnik podjetja Amebis označuje s približno 85-odstotno natančnostjo, ostalih 15 % poleg napak pri označevanju vključuje tudi neoznačene neprepoznane besede, ki jih je približno 2 %. Kljub temu korpus predstavlja dobro podlago za razvoj prosto dostopnega ročno označenega korpusa za namene jezikovnotehnoloških raziskav.

[2] http://www.fida.net/
[3] http://www.fidaplus.net/
[4] http://www.amebis.si/

Prvi korak na poti od korpusa FidaPLUS do korpusa JOS je bila pretvorba v format XML po priporočilih TEI P4 (Sperberg-McQueen in Burnard, 2002), da bi s tem ohranili standardni format in omogočili uporabo orodij za delo s formatom XML, predvsem XSLT. Format TEI P4 je sicer povratno združljiv s formatom TEI P3 korpusa FIDA in XML je podmnožica formata SGML, vendar končni korpus FidaPLUS kot podatkovna zbirka ni v celoti skladen niti s formatom SGML niti s specifikacijami MULTEXT-East, zato je bil proces pretvorbe zahtevnejši, kot je bilo pričakovati. Procesiranje je bilo izvedeno s pomočjo niza skript v programskem jeziku Perl, končni pretvorjeni korpus FidaPLUS pa imenujemo Fida+X. Ta je za malenkost manjši kot izvorni korpus, ker smo izpustili besedila, pri katerih hevristični postopki s pomočjo Perl skript niso zadostovali za njihovo kompatibilnost s standardom TEI P4. Korpus Fida+X je bil uporabljen kot vir za izdelavo korpusov JOS.

2.1. Postopek vzorčenja

Korpusa jos100k in jos1M sta nastala iz korpusa Fida+X z dvostopenjskim filtriranjem in procesom vzorčenja z namenom, da pri končnem rezultatu dosežemo naslednje cilje:

- Uravnoteženost in reprezentativnost: slednja lastnost izhaja iz zasnove korpusa FidaPLUS, pri čemer pa ta vsebuje velik delež časopisnih besedil in relativno majhen delež literarnih, predvsem pa strokovnih besedil. Ob preprostem prevzemanju razmerij iz korpusa FidaPLUS bi bila uravnoteženost korpus JOS vprašljiva.
- Kvaliteta besedil: ker bo končni korpus jezikoslovno označen, je bilo pomembno, da vsebuje le smiselne in zaključene odstavke in pojavnice: FidaPLUS vsebuje dele besedil, ki se ponavljajo, in primere, kjer so ob pretvorbi iz izvornih besedil nastali zelo kratki stavki ali odstavki, ki včasih vsebujejo tudi ostanke podatkov o formatiranju.
- Avtorske pravice: čeprav bi bilo za določene tipe analiz bolje vključiti celotna besedila, bi bilo to vprašljivo s stališča kršenja avtorskih pravic besedilodajalcev, vključitev krajših delov besedil pa je manj problematična – postopek vzorčenja poleg tega prispeva k večji raznolikosti korpusa.

Pri prvem koraku vzorčenja so naključna izbrana celotna besedila iz korpusa Fida+X, vendar so izključena besedila z nepravilnimi deli, določeni tipi besedil pa so v vzorcu bolj zaželeni. Filter najprej izključi besedila, ki so prekratka ali predolga, vsebujejo podatke o formatiranju ali so slabo oblikovana glede na različne hevristične kriterije. Sledi faza obteževanja po tipih besedil in drugih metapodatkih z namenom, da se uravnoteži delež časopisnih besedil v primerjavi z drugimi tipi besedil. V drugem koraku postopka so izbrani naključni odstavki, ki so ponovno podvrženi preverjanju: minimalni ali maksimalni dolžini ter enkratnosti pojavljanja. Oba koraka sta bila izvedena dvakrat, vsakič z drugimi nastavitvami. Za korpus jos100k je bil za prvi korak izdelan 10-milijonski korpus, za jos1M pa 100-milijonski. Za oba korpusa je bil potem v drugem koraku izbran en odstotek odstavkov.

Tabela 1 vsebuje podatke o številu besedil, odstavkov, stavkov, besed ter pojavnic (besed skupaj z ločili).

Korpus	jos100k	jos1M
<text>	249	2.565
<p>	1.599	15.758
<s>	6.151	60.291
<w>	100.003	1.000.019
<w> + <c>	118.394	1.182.945

Tabela 1: Število oznak v obeh korpusih JOS.

V korpusu FidaPLUS je vsako besedilo označeno glede na tip in zvrst besedila. Pri prvem koraku vzorčenja korpusov JOS smo te kategorije različno obtežili in razmerja se tako razlikujejo od korpusa FidaPLUS. V tabeli 2 in 3 navajamo razmerja med vrhnjimi kategorijami v obeh korpusih JOS. Razmerja smo med prvim in drugim vzorčenjem rahlo spremenili, zato se pri obeh razlikujejo. JOS1M vsebuje manj literarnih besedil in časopisnega gradiva in več naravoslovnih in tehničnih besedil ter mesečnih publikacij.

Zvrst	jos100k	jos1M
umetnostna besedila (proza)	10,1 %	6,7 %
neumetnostna besedila (nestrokovna)	67,6 %	66,6 %
družboslovje in humanistika	9,6 %	9,9 %
naravoslovje in tehnologija	6,5 %	13,3 %
skupaj	93,8 %	96,6 %

Tabela 2: Razmerja med besedili po zvrsteh

Tip	jos100k	jos1M
monografija	28,1 %	22,9 %
mesečna revija	9,6 %	16,2 %
štirinajstdnevna revija	2,3 %	2,0 %
tedenska revija	7,3 %	9,1 %
tedenski časopis	9,0 %	10,4 %
dnevni časopis	37,7 %	24,3 %
skupaj	94,1 %	95,0 %

Tabela 3: Razmerja med besedili po tipu

3. Oblikoskladenjske specifikacije in nabor oznak JOS

Oblikoskladenjske specifikacije MULTEXT(-East) temeljijo na delu skupine EAGLES (Calzolari in Monachini, 1996) in določajo strukturo in vsebino veljavnih oblikoskladenjskih oznak ali MSD-jev (morpho-syntactic descriptions). Specifikacije za vsak posamezen jezik opredeljujejo, katere so veljavne oznake in kaj pomenijo. Tako na primer določajo, da je MSD s črkovnim nizom `Sometn` veljaven za označevanje slovenščine in je ekvivalenten naboru naslednjih lastnosti: `samostalnik, vrsta = občni, spol = moški, število = ednina, sklon = tožilnik, živost = ne`. Ker je slovenščina oblikoslovno izjemno bogat jezik z velikim številom lastnosti pri pregibnih besednih

vrstah, je število veljavnih oznak precej večje kot pri večini zahodnoevropskih jezikov – okrog 2.000.

Izhodišče pri odločanju o prenovi nabora oznak MULTEXT-East je bila ocena, da osnovni način formalnega zapisa in struktura oznak dobro služi svojemu namenu, da pa prihaja do težav pri določenih lastnostih in njihovih vrednostih, predvsem pri nekaterih dovoljenih kombinacijah lastnost-vrednost, ter pri pripisovanju nekaterih MSD-jev določenim lemam in oblikam. Nadaljnja težava, tokrat celotnega nabora specifikacij MULTEXT-East, je razvrstitev lastnosti v črkovni niz MSD-jev. Ker specifikacije veljajo za celo vrsto različnih jezikov, tiste lastnosti, ki so značilne samo za določen jezik, končajo na koncu črkovnega niza s praznimi mesti pri lastnostih, ki jih jezik ne izkazuje. Tako lahko pride do izjemno dolgih črkovnih nizov, kot je npr. glagolski `Gppspe--n-----d`. Smiselno je torej dovoliti prerazporeditev mesta lastnosti glede na posamezen jezik, kar omogoči, da so nizi krajši, hkrati pa s pomočjo preslikave lastnosti in vrednosti ohranimo kompatibilnost označevanja z drugimi jeziki.

Namen prenove oblikoskladenjskih specifikacij je bil med drugim tudi standardizacija nabora oznak za slovenščino. Zato je bila opravljena analiza označevanja korpusa FidaPLUS ter razmeroma obsežen pregled drugih naborov oznak za slovenščino ter za druge jezike. Pri slovenskem jeziku sta bila upoštevana nabora oznak, uporabljena pri označevanju korpusa LC-STAR (Verdonik et al., 2004) ter korpusa Nova beseda (Jakopin in Bizjak, 1997). Zadnji se od prvih dveh temeljno razlikuje, saj ne uporablja pozicijskega načela pri pripisovanju lastnosti in se močno opira na tradicionalni slovnični opis jezika (Lönneker, 2005). Od drugih jezikov so bili podrobneje analizirani nabori CLAWS za angleški jezik ter češki nabor AJKA ter nabor oznak, uporabljen pri označevanju Češkega nacionalnega korpusa[5] ter Praške odvisnostne drevesnice.[6]

Končni nabor oznak JOS v osnovi ohranja načela nabora MULTEXT-East, vendar postopek preslikave med obstoječimi oznakami v korpusu FidaPLUS po sistemu MULTEXT-East in novimi po sistemu JOS ni trivialen, kajti pri pretvorbi je potrebno upoštevati tudi besedno obliko in/ali lemo.

Tabela 4 kaže povzetek kategorij in lastnosti nabora oznak JOS. Specifikacije vsebujejo tudi listo veljavnih MSD-jev, ki jih je 1.908, pri vsakem pa so dodani tudi konkretni primeri besednih oblik iz leksikona. Nabor oznak je torej kljub spremembam precej obsežen in nekatere kombinacije lastnosti so še vedno precej redke. Tako denimo se v korpusu jos100k pojavlja 1.064 različnih MSD-jev, kar je le 55,7 % celotnega nabora.

Slika 1 kaže primer iz specifikacij nabora JOS v formatu XML. Specifikacije so na voljo v angleškem in slovenskem jeziku, to velja za kategorije, lastnosti in vrednosti. S pomočjo specifikacij in ustrezne datoteke XSLT je tako denimo mogoče prevesti slovensko oznako `Sometn` v angleško `Ncmsan` (`Noun, Type = common, Gender = masculine, Number = singular, Case = accusative, Animate = no`), kar omogoča enostavno prehajanje med oznakami v angleškem in slovenskem jeziku.

besedna vrsta	lastnosti s številom vrednosti
samostalnik	vrsta(2), spol(3), število(3), sklon(6), živost (2)
glagol	vrsta(2), vid(3), oblika(7), oseba(3), število(3), spol(3), nikalnost(2)
pridevnik	vrsta(3), stopnja(3), spol(3), število(3), sklon(6), določnost(2)
prislov	stopnja(3), deležje(2)
zaimek	vrsta(9), oseba(3), spol(3), število(3), sklon(6), število_svojine(3), spol_svojine(3), oblika(2)
številnik	zapis(3), vrsta(4), spol(3), število(3), sklon(6), določnost(2)
predlog	sklon(6)
veznik	vrsta(2)
členek	brez lastnosti
medmet	brez lastnosti
okrajšava	brez lastnosti
neuvrščeno	vrsta(3)

Tabela 4: Kategorije nabora oznak JOS z lastnostmi in številom vrednosti

```
<div type="section" xml:id="msd.N">
<head>SAMOSTALNIK</head>
<table n="msd.cat" xml:id="msd.cat.N">
<head xml:lang="sl">Tabela atributov in
vrednosti za samostalnik</head>
<head xml:lang="en">Attribute-value table for
Noun</head>
<row role="type">
<cell role="position">0</cell>
<cell role="name"
xml:lang="sl">samostalnik</cell>
<cell role="code" xml:lang="sl">S</cell>
<cell role="name" xml:lang="en">Noun</cell>
<cell role="code" xml:lang="en">N</cell>
</row>
<row role="attribute">
<cell role="position">1</cell>
<cell role="name" xml:lang="sl">vrsta</cell>
<cell role="name" xml:lang="en">Type</cell>
<cell role="values">
<table>
<row role="value">
<cell role="name" xml:lang="sl">občno
ime</cell>
<cell role="code" xml:lang="sl">o</cell>
<cell role="name" xml:lang="en">common</cell>
<cell role="code" xml:lang="en">c</cell>
</row>
<row role="value">
<cell role="name" xml:lang="sl">lastno
ime</cell>
<cell role="code" xml:lang="sl">l</cell>
<cell role="name" xml:lang="en">proper</cell>
<cell role="code" xml:lang="en">p</cell>
</row>
</table>
</cell>
</row>
...
```

Slika 1: oblikoskladenjske specifikacije JOS – začetek tabele za samostalnik

[5] http://ucnk.ff.cuni.cz/

[6] http://ufal.mff.cuni.cz/pdt/

4. Jezikoslovno označevanje korpusa

Ročno označevanje korpusov je izvedla ekipa študentov pod strokovnim nadzorom. Študenti so preverjali in popravljali oznake, ki so bile iz izvornih oznak po naboru MULTEXT-East iz korpusa FidaPLUS preslikane v nabor JOS. Ročno označevanje je potekalo s pomočjo spletnega vmesnika, ki na podlagi danih parametrov (npr. regularnih izrazov, ki lahko upoštevajo besedno obliko, lemo ali MSD) generira preglednice v formatu MS Excel. Preglednica vsebuje list z osnovnimi podatki o vsebini, list z besedilom in oznakami, ki jih je potrebno pregledati, ter list z navodili. Po ročnem preverjanju preglednico na isti spletni strani naložimo nazaj v korpus, pri čemer je korpus avtomatsko obnovljen z novimi ročno pregledanimi oznakami. Sistem je bil izvorno razvit za popravljanje in označevanje zgodovinskih besedil (Erjavec, 2007) in je bil uspešno uporabljen tudi v projektu JOS. Proces označevanja je cikličen, z mešanimi ročnimi in avtomatskimi postopki.

4.1. Označevanje korpusa jos100k

Označevanje korpusa jos100k je potekalo vzporedno z nastajanjem oblikoskladenjskih specifikacij JOS, novega nabora oznak in preslikave MSD-jev. Proces je bil zato bolj zapleten, vendar so specifikacije, nabor oznak in oznake v korpusu zato konsistentnejši. Celoten korpus jos100k sta preverjala po dva različna označevalca, pojavnice, pri katerih je med njima prihajalo do razlik, pa so bile preverjene še s strani tretjega. V primerih, kjer so bile odkrite nekonsistentnosti pri označevanju določenih kategorij oz. lastnosti, so bile te v kasnejših korakih odpravljene, korpus pa poenoten, rezultat pa je bil v celoti ročno označeni zlati standard JOS.

4.2. Označevanje korpusa jos1M

Ker finančna shema projekta ne dopušča ročnega preverjanja celotnega milijonskega korpusa JOS, se pri milijonskem korpusu jos1M poleg prvega koraka avtomatske pretvorbe iz nabora oznak MULTEXT-East v nabor JOS ročno preverjajo le "sumljive" oznake, v nadaljevanju pa opisujemo uporabljeno metodologijo.

Na ročno označenem korpusu jos100k smo izvedli učni proces z označevalnikom TnT (Brants, 2000) in mu pri postopku kot oporo dodali leksikon besednih oblik, izdelan iz celotnega korpusa Fida+X, skupaj s konvertiranim naborom oznak JOS. V korpusu jos1M ročno preverjamo le pojavnice, kjer se oznake, ki jih je pripisal označevalnik TnT, ter že obstoječe (pretvorjene) oznake iz korpusa Fida+X, ki jih je pripisal Amebisov označevalnik, razlikujejo.

Da bi preverili točnost obeh označevalnikov in prekrivanje napak, ki jih delata, smo izvedli poskus z desetkratnim navskrižnim preverjanjem na korpusu jos100k z označevalnikom TnT, kar pomeni, da smo učili označevalnik na 90 % besedila in označili preostalih 10 %, postopek pa smo ponovili desetkrat na vseh desetinah korpusa. Rezultat je bil korpus, kjer je vsaka pojavnica označena s tremi oznakami: ročno preverjena oznaka, oznaka, ki jo je pripisal Amebisov označevalnik in oznaka, ki jo je pripisal označevalnik TnT. V tabeli 5 je prikazan rezultat poskusa.

V prvi vrstici podamo celoten obseg korpusa glede na število besed. Druga vrstica kaže oceno natančnosti označevalnika TnT (86,6%), tretja pa oceno natančnosti Amebisovega označevalnika (85,7%). Amebisov označevalnik sicer ne pripisuje oznak neznanim besedam, ki obsegajo približno dva odstotka pojavnic, pri ročnem označevanju pa so bile označene vse pojavnice (tudi npr. tujejezične citatno pisane pojavnice, ki spadajo v kategorijo "neuvrščeno"), enako tudi z označevalnikom TnT. Četrta vrstica zajema primere, ko sta oba označevalnika pripisala pravilni MSD (78%), v naslednjih štirih vrsticah pa podajamo rezultat za primere, ko sta se bodisi en ali drugi ali oba označevalnika zmotila. Predvsem je pomemben rezultat v zadnji vrstici, kjer sta oba označevalnika pripisala enako oznako, vendar oba napačno. Pri ročnem preverjanju korpusa jos1M namreč zajemamo le tiste pojavnice, kjer se označevalnika pri pripisu oznake ne strinjata. Pri milijonskem korpusu izračun 7,7% + 8,6% + 2,4% pokaže, da je potrebno ročno validirati okoli 190.000 pojavnic, končna pravilnost označevanja pa je ocenjena na 96,8 %, saj moramo odšteti primere (3,2%), kjer s križanjem rezultatov obeh označevalnikov ne moremo zaznati napak.

Ročno označevanje milijonskega korpusa je v teku, hkrati pa raziskujemo načine, kako izboljšati rezultat pri označevanju izvornih besedil iz korpusa Fida+X s kombiniranjem rezultatov označevanja obeh označevalnikov.

	Št. besed	Ročno	Amebis	TnT	Razlaga
1	100,003	MSD1			Besed v korpusu jos100k
2	86,617	MSD1		MSD1	TnT pravilno označenih
3	85,719	MSD1	MSD1		Amebis pravilno označenih
4	78,011	MSD1	MSD1	MSD1	Oba pravilno označila
5	7,708	MSD1	MSD1	MSD2	Amebis pravilno, TnT narobe
6	8,606	MSD1	MSD2	MSD1	Amebis narobe, Tnt pravilno
7	3,238	MSD1	MSD2	MSD2	Oba narobe, in enako
8	2,440	MSD1	MSD2	MSD3	Oba narobe, in različno

Tabela 5: natančnost označevanja korpusa jos100k – označevalnik Amebis / TnT

5. Format korpusov

Tako korpusa kot oblikoslovne specifikacije so kodirani v formatu XML v skladu s priporočili združenja Text Encoding Initative. Čeprav je bila pri razvoju korpusov zaradi kompatibilnosti s korpusom FidaPLUS uporabljena inačica priporočil P4, je za javno dostopno inačico uporabljena zadnja izdaja priporočil, TEI P5 (TEI Consortium, 2007). Rezultati so dostopni skupaj s pripadajočo shemo XML, ki omogoča formalno validiranje specifikacij in korpusov.

Korpusa sta sestavljena iz dveh delov. Kolofon (TEI header) vsebuje metapodatke, kjer se med drugim nahajajo podatki o velikosti korpusa, o uporabi oznak, naštete so odgovorne osebe, bibliografski podatki o vsebovanih besedilih, vsebuje pa tudi celoten nabor

oblikoslovnih oznak nabora JOS in njihovo dekompozicijo na pare lastnost-vrednost. Drugi del korpusa sestavljajo besedila.

Vsako besedilo v korpusu vsebuje povezavo na opis v kolofonu in je sestavljeno iz niza vzorčenih odstavkov, te sestavljajo stavki in stavke posamezne pojavnice in ločila. Poseben element označuje presledke, vsaka pojavnica pa ima v obliki atributa podatek o lemi in oblikoslovni oznaki.

6. Dostopnost

Jezikovni viri JOS so dostopni na spletni strani projekta,[7] korpuse pa je dovoljeno prenesti na svoj računalnik in uporabljati v skladu z licenco Creative Commons: Priznanje avtorstva-Nekomercialno.[8] Komercialna uporaba korpusov ni dovoljena, ker tega ne dopušča pogodba med besedilodajalci in konzorcijem partnerjev, ki so izdelali korpusa FIDA in FidaPLUS.

Korpusa JOS sta na voljo v izvornem formatu XML TEI P5 ter v več formatih, ki so primerni za predprocesiranje in druge predelave. Med njimi je omembe vreden predvsem format za Corpus Workbench[9] (Christ, 1994), pri katerem vsaka vrstica vsebuje bodisi formalno strukturno oznako ali s tabulatorjem ločeno informacijo o besedni obliki, lemi in MSD-ju, skupaj z razčlenjenimi lastnostmi. To omogoča iskanje po posameznih lastnostih pojavnice ne glede na najvišjo kategorijo, besedno vrsto. Tako bi pri iskanju vseh pojavnic, ki označene z vrednostjo "ženski spol" in "rodilnik", v iskalniku Corpus Workbench uporabili iskalni pogoj [spol="ženski" & sklon="rodilnik"]. Korpusi so dostopni tudi preko spletnega vmesnika, ki uporablja Corpus Workbench, in je ravno tako dosegljiv na domači strani projekta.

7. Zaključek

V prispevku smo predstavili prve rezultate projekta JOS, predvsem dokončani korpus jos100k in korpus jos1M, ki je v zaključni fazi izdelave. Prispevek je opisal korpus FidaPLUS kot vir za oba korpusa JOS, postopke čiščenja in vzorčenja besedil, prenovljen nabor oznak JOS ter oblikoslovnih specifikacij. Komentiral je tudi postopek jezikoslovnega označevanja korpusov, njun format in dostopnost. Predstavljena korpusa sta prva kvalitetno označena in brezplačno dostopna jezikovna vira za slovenščino in bosta znatno olajšala raziskave pri avtomatskem oblikoskladenjskem označevanju in lematizaciji slovenskih besedil.

Nadaljnje delo pri projektu JOS zajema naslednji dve ravni jezikoslovnega označevanja besedil, predvsem skladenjsko označevanje korpusa jos100k in pomensko označevanje in razdvoumljanje s pomočjo slovenskega semantičnega leksikona, narejenega po vzoru WordNet (Erjavec in Fišer, 2006).

Zahvala

Avtorja se zahvaljujeta recenzentom za koristne pripombe. 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”.

[7] http://nl.ijs.si/jos/

[8] http://creativecommons.org/licenses/by-nc/2.5/si

[9] http://cwb.sf.net

Literatura

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

Brants T. (2000). TnT - A Statistical Part-of-Speech Tagger. V Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000 (str. 224–231). ACL.

Calzolari, N. in Monachini, M. (ur.) (1996). Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora: A common proposal and applications to European languages. EAGLES Report EAG—CLWG—MORPHSYN/R. Pisa: ILC.

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

Erjavec, T. in Fišer, D. (2006). Building Slovene WordNet. V Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006 (str. 1678–1683). Pariz: ELRA.

Erjavec, T., Gorjanc, V. in Stabej, M. (1998). Korpus FIDA. V Proceedings of the Conference ’Language Technologies for the Slovene Language’ (str. 124–127). Ljubljana: IJS.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. V Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (str. 1535–1538). Pariz: ELRA.

Erjavec, T. (2007). An architecture for editing complex digital documents. V Proceedings of INFuture2007: “Digital Information and Heritage” (str. 105-114) Zagreb: Fakulteta za humanistiko in socialne vede**.**

Jakopin, P. in Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija 45 (3–4), 513–532.

Lönneker, B. (2005). Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo? Slavistična revija 53 (2), 193–210.

Sharoff, S., Kopotev, M., Erjavec, T., Feldman A. in Divjak, D. (2008). Designing and evaluating a Russian tagset. V Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2008. Pariz: ELRA.

Sperberg-McQueen, C. M. in Burnard, L. (ur.) (1999). Guidelines for Electronic Text Encoding and Interchange Revised Reprint. The TEI Consortium.

Sperberg-McQueen, C. M. in Burnard, L. (ur.) (2002). Guidelines for Electronic Text Encoding and Interchange, The XML Version of the TEI Guidelines. The TEI Consortium.

TEI Consortium (2007). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Verdonik, D., Rojc, M. in Kačič, Z. (2004). Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004. Pariz: ELRA.

Oblikoskladenjske oznake JOS: revizija in nadgradnja nabora oznak za avtomatsko oblikoskladenjsko označevanje slovenščine

Špela Arhar,* Nina Ledinek†

* Amebis, d. o. o., Kamnik
Bakovnik 3, 1341 Kamnik
spela.arhar@amebis.si
† Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU
Novi trg 4, 1000 Ljubljana
nledinek@zrc-sazu.si

Povzetek

Prispevek predstavlja revizijo ter nadgradnjo nabora oznak za oblikoskladenjsko označevanje slovenščine, ki sta v letu 2007 potekali v okviru projekta Jezikovno označevanje slovenščine. Končni rezultat nadgradnje – nabor oblikoskladenjskih oznak JOS – je zasnovan s ciljem vzpostavitve enotnega označevalnega standarda za slovenščino in je kot tak ponujen v oceno zainteresirani strokovni javnosti. V članku so navedeni razlogi za revizijo, utemeljena je izbira kodne tabele Multext-East kot izhodiščne za nadgradnjo, opisan je potek dela, skupaj z izpostavitvijo najbolj problematičnih mest oblikoskladenjskega označevanja. Revizija ter nadgradnja sta podrobneje predstavljeni na primeru sprememb nabora oznak za glagolsko besedno vrsto. Članek zaključujejo podatki o dostopnosti nove kodne tabele ter povezane dokumentacije.

The JOS morphosyntactic tags: the revision and upgrade of the tagset for automatic morphosyntactic annotation of Slovene

The paper presents the revised and upgraded tagset for morphosyntactic annotation of Slovene which is one of the first results of the JOS project – "Linguistic annotation of the Slovene language". The JOS tagset was designed with the aim to become the standard tagset for morphosyntactic tagging of Slovene and it is now available to be examined and evaluated by the interested professional public. The paper starts with the discussion of the underlying reasons for the revision and with the arguments for choosing the Multext-East tagset as the basis of the upgrade. We describe the revision process and underline the most problematic issues in morphosyntactic annotation. The revision and upgrade are exemplified by the changes in the morphosyntactic features and values for the verb. The paper concludes with the information on the availability of the new tagset and its documentation.

1. Uvod

Pripisovanje oblikoskladenjskih oznak besedam je ena najosnovnejših oblik vključevanja interpretativnih informacij jezikoslovnega tipa v besedilne korpuse oz. podobne jezikovne vire. S tem, ko posamezni pojavnici pripišemo, v kateri osnovni slovnični razred spada v specifičnem jezikovnem sistemu ter nabor lastnosti, ki jih izkazuje znotraj tega razreda, omogočimo nadaljnjo izrabo jezikovnega vira na višjem nivoju: posledica označenosti je možnost raziskovanja besedil z uporabo abstraktnejših jezikoslovnih kategorij, ne le golih črkovnih nizov ali lem, kar odpira nove možnosti za jezikoslovne raziskave označenih virov, izrabo le-teh za razvoj metod obdelave naravnega jezika, jezikovnih tehnologij ipd. (Gorjanc, 2005; McEnery, Xiao in Tono, 2006).

Iz navedenega je jasno, da je uspešnost nadaljnje izrabe korpusnih virov pogojena s premišljeno zasnovanostjo označevalnih kategorij in znotraj njih naborov vrednosti, med katerimi pri označevanju izbiramo. Želja po kvaliteti označevanja utemeljuje revizijo obstoječih naborov oznak ter poskus njihove optimizacije. Pozornost pa je naboru oznak potrebno nameniti še z zornega kota standardizacije: za jezikovna področja z manjšim številom govorcev se še posebej kaže za smiselno vzpostavitev enotnega standarda za oblikoskladenjsko označevanje besedil na privzeti ravni, tj. standarda, pri snovanju katerega so upoštevane potrebe širše raziskovalne skupnosti in ki obenem v končni fazi ponuja stabilno izhodišče za pretvorbo v robustnejši ali finejši sistem označevanja, kadar se to izkaže za potrebno.

2. Razpoložljivi nabori oblikoskladenjskih oznak za slovenščino

Potrebe po oblikoskladenjski označenosti besedil so se do nedavnega reševale v okvirih posameznih institucij oz. projektov. V zvezi s tem je smiselno omeniti vsaj tri trenutno aktualne sisteme oblikoskladenjskih oznak za slovenščino:

- jezikovnospecifičen nabor oznak Inštituta za slovenski jezik Frana Ramovša ZRC SAZU je nastal za potrebe označevanja oblikoslovno označenega korpusa Beseda (Jakopin in Bizjak, 1997; Lönneker in Jakopin, 2004),
- za slovenščino prilagojeni nabor oznak LC-STAR se uporablja pri gradnji jezikovnih virov za simultano prevajanje govora (Verdonik, Rojc in Kačič, 2004)
- tako trenutno aktualni referenčni korpus FidaPLUS (http://www.fidaplus.net/) kot predhodni korpus FIDA (http://www.fida.net/) sta označena z naborom oznak, ki so nastale v sklopu projekta Multext-East.

Pri snovanju označevalnega sistema JOS, o katerem bo govora v pričujočem prispevku, so bili upoštevani vsi trije označevalni sistemi, poleg tega pa še nekateri tujejezični (glej 3.3.2).

Večji poudarek je bil namenjen zadnjemu od naštetih, primarno zato, ker od naštetih v največji meri sledi priporočilom iniciative EAGLES/ISLE. Ostali argumenti za izbor nabora Multext-East kot izhodiščnega za nadgradnjo so bolj ali manj vezani na z njim označeni referenčni korpus, ki na osnovi besedilne sestave, velikega obsega ter kvantitativnim analizam prijaznega

korpusnobesedilnega formata (Arhar in Gorjanc 2007) prinaša za revizijo nepogrešljive informacije[1] (glej 3.3.2).

2.1. Nabor oznak Multext-East

Prva verzija nabora oznak Multext-East je nastala med letoma 1995 in 1997, trenutno je na voljo v svoji tretji različici (Erjavec, 2004; http://nl.ijs.si/ME/V3/). Slednja prinaša tabele oblikoskladenjskih oznak in pojasnila o sistemu njihove rabe za deset evropskih jezikov, med njimi tudi slovenskega. Na osnovnem nivoju sistem predvideva klasično besednovrstno uvrstitev obravnavane besede, na drugem nivoju pa se označuje njene nadaljnje lastnosti glede na slovarske ter slovnične kategorije, ki jih posamezni besedni vrsti lahko pripišemo (pridevniku po tem sistemu npr. vrsto, stopnjo, spol, število, sklon, določnost ter živost).

Pri pripravi označevalne tabele za več jezikov hkrati je bila potrebna harmonizacija le deloma prekrivnih jezikoslovnih kategorij različnih jezikov (Erjavec, 2004). V označevalni praksi se zaradi tega izkazuje problem prostih mest v kodni tabeli in posledično posameznih kodah (kodna tabela za glagol denimo prinaša šest prostih mest, kot je razvidno iz Tabele 1) in problem opredeljevanja kategorij, ki jih jezik sicer izkazuje, je pa stopnja njihove avtomatske pripisljivosti včasih premajhna, da bi bili vneseni podatki relevantni.

V času priprave kodne tabele referenčni korpus za slovenski jezik še ni obstajal, kar je za snovanje kategorij in njihovih vrednosti pomenilo možnost naslonitve na predvsem predkorpusno jezikoslovno vedenje. Prav tako še niso bile na voljo informacije o oblikoskladenjskem označevanju slovenščine v smislu empiričnih podatkov o tem, na kakšen način se jezikovnosistemske kategorije, kakor jih opisuje slovensko jezikoslovje, v praksi lahko avtomatsko označijo. Šele po izkušnjah z označevanjem različnih jezikovnih virov se je torej lahko pokazalo, da kodna tabela prinaša kategorije oz. vrednosti, ki ne omogočajo zadovoljivo natančne avtomatske pripisljivosti. Problemi se kažejo na mestih, kjer je za določitev kategorije oz. vrednosti potrebno:

- upoštevanje besedne okolice na skladenjski ravni, ker je določitev vezana na niz pojavnic in ne le na posamezno pojavnico,
- razdvoumljanje na osnovi semantičnih informacij, ki je mogoče le na ravni ročnega označevanja ali pa še to ne (za primere glej 4.2.2).

Težave se pojavljajo tudi na ravni neuravnoteženosti v razmerju pogostnosti oznake v korpusu v primerjavi s številom lem, ki jih oznaka pokriva, in sicer na ravni:

- neuravnoteženosti med vrednostmi znotraj kategorije (npr. vrednost *elativ* nasproti drugim vrednostim kategorije *stopnja* pri pridevniku in prislovu),
- neuravnoteženi razpršenosti posameznih oznak (npr. vrednost *svojilni* kategorije *nanašanje*, ki je pripisana samo zaimkovni lemi *svoj*).

Poleg navedenih pomanjkljivosti je tudi že omenjena širša raba oblikoskladenjsko označenih korpusnih virov botrovala želji po odpravi določenih šibkih mest v sistemu oznak, kot so identificirana na podlagi

- nabora povratnih informacij uporabnikov referenčnega korpusa,
- rezultata analize kvalitete označenosti referenčnega korpusa FidaPLUS,
- rezultata analize statistik o pogostnosti ter razpršenosti posamezne oznake v referenčnem korpusu (glej poglavje 3.3).

3. Revizija in nadgradnja nabora oznak

3.1. Projekt JOS

Revizija ter nadgradnja nabora oblikoskladenjskih oznak je potekala prvi polovici leta 2007 v sklopu projekta JOS (Jezikoslovno označevanje slovenščine), katerega glavni cilj je zagotoviti slovenski raziskovalni javnosti prosto dostopen večnivojsko (oblikoskladenjsko, skladenjsko, semantično) označen milijonski korpus besedil vsakdanje rabe – korpus JOS (več o projektu Erjavec in Krek, 2008; http://nl.ijs.si/jos/).

3.2. Načela nadgradnje označevalnega sistema

Revizijo ter nadgradnjo so vodila načela, v večji meri povzeta po priporočilih za oblikoskladenjsko označevanje korpusov *EAGLES* (http://www.ilc.cnr.it/EAGLES/annotate/annotate.html), pri čemer pa je bilo glavno vodilo pri izdelavi prenovljenega označevalnega sistema izhajanje izključno iz slovenskega jezika, z upoštevanjem potrebe po primerljivosti oznak z obstoječimi standardi.[2] V pomoč slednjemu je bila ohranjena struktura kodnih tabel Multext-East, ki za vsako od besednih vrst predvidevajo na besedno vrsto vezane kategorije ter znotraj le-teh nabor vrednosti z enočrkovno reprezentacijo (prim. Tabeli 1 in 2). Kjer je bilo mogoče, se je že ob snovanju sprememb kategorij oz. vrednosti upoštevalo potencialne probleme pretvorbe med označevalnim sistemom Multext-East ter novopredlaganim sistemom JOS.

Želja po strnjenosti ter berljivosti oznak je vodila k odpravi prostih mest v kodnih tabelah. V primeru, da so bile v kodno tabelo dodane nove kategorije, so bile le-te razvrščene na čim bolj ustrezno mesto (glej 4.2.3).

Načelo pripisljivosti je vodilo k odstranitvi kategorij oz. vrednosti, za katere se je izkazalo, da avtomatsko niso pripisljive na zadovoljivi ravni (zaradi potrebe po upoštevanju besednega konteksta, zaradi dvoumnosti itd.). Načelo uravnoteženosti pa je vodilo k preoblikovanju kodne tabele na mestih, kjer so se oznake kazale za preveč ali premalo razpršene oz. preveč ali premalo specifične.

3.3. Potek dela

3.3.1. Analiza kvalitete označenosti korpusa FidaPLUS

Tekom večletne rabe označevalnega sistema Multext-East so se v jezikovnotehnoloških krogih pojavljale pobude za revizijo sistema. Prva empirična analiza problemov označevanja pa je bila analiza kvalitete oblikoskladenjske označenosti korpusa FidaPLUS, ki je bila po označitvi referenčnega korpusa izvedena na

[1] Korpus FidaPLUS je bil obenem tudi vir za izdelavo podkorpusa JOS (glej 3.1).

[2] Priporočila so se kot abstrakcija zgledov dobrih praks zaradi pragmatičnih potreb izoblikovala v nekakšen neformalni standard oblikoskladenjskih oznak, zato je njihovo upoštevanje (ob hkratnem upoštevanju namembnosti korpusa, ki ga oblikoskladenjsko označujemo) pomembno zlasti z vidika primerljivosti in izmenljivosti korpusnih podatkov.

podjetju Amebis, d. o. o., Kamnik.[3] V raziskavi je bilo ročno pregledanih 78.000 vrstic označenega korpusnega besedila.

Na tej ravni analize so se za najbolj problematične izkazale kategorije oz. vrednosti, ki za ustrezno določitev zahtevajo upoštevanje okolice obravnavane besedne oblike – tj. označevanja na višjem nivoju (skladenjskega, pomenskega), vendar mora biti njihovo označevanje podprto z ustrezno programsko infrastrukturo, ki za slovenščino trenutno še ni na voljo. Na ravni oblikoskladenjskega označevanja je torej smiselno vztrajati pri kategorijah, ki so pripisljive na osnovi informacij, ki jih izkazuje besedna oblika kot taka.

3.3.2. Sistematičen pregled potencialnih kategorij in vrednosti

Na prvi ravni označevanja je bila ohranjena kategorizacija po besednih vrstah.[4] Za vsako od besednih vrst je bila izvedena natančnejša analiza potencialnih označevalnih kategorij, pri kateri so bili upoštevani nabori kategorij ter vrednosti, kot jih prinašajo za slovenščino relevantni sistemi označevanja – poleg že omenjenih slovenskih ISJ ZRC SAZU (Jakopin in Bizjak, 1997) in LC-STAR (Verdonik, Rojc, Kačič, 2004) so bili pregledani še označevalni sistemi Češkega nacionalnega korpusa (http://ucnk.ff.cuni.cz/bonito/) ter sistema Claws (http://ucrel.lancs.ac.uk/claws/) in Ajka (http://nlp.fi.muni.cz/projekty/ajka/tags.pdf).

Pri pregledu potencialnih kategorij so bili upoštevani tudi razpoložljivi jezikoslovni priročniki (Slovenska slovnica, SSKJ, Slovenski pravopis, Besedišče slovenskega jezika itd.) – predvsem v zvezi z vprašanji določanja mej med posameznimi vrednostmi znotraj kategorij – vedno pa v kombinaciji z analizo primerov dejanske jezikovne rabe, kakor se kaže v referenčnem korpusu. V primerih suma, da so meje težko določljive, je bil za problematično kategorijo ali nabor vrednosti izdelan testni nabor korpusnih primerov, ki so bili na ravni obravnavanega problema ročno označeni s strani treh označevalcev, oznake pa v nadaljevanju pregledane ter primerjane.

Vzporedno s pregledom kategorij je potekala analiza že aktualiziranih oznak Multext-East v referenčnem korpusu s stališča podatkov o:

- pogostnosti zastopanosti obravnavane oznake v korpusu,
- številu različnih lem, ki so v korpusu z obravnavano oznako označene – v večini primerov je sledila tudi analiza nabora teh lem.

Na osnovi teh informacij je bila z upoštevanjem zgoraj navedenih načel za nadgradnjo pripravljena nova kodna tabela za oblikoskladenjsko označevanje.

3.3.3. Izdelava leksikonov

Analiza kvalitete označevanja (glej 3.3.1) in netrivialnost postopka prekodiranja oblikoskladenjskih oznak Multext-East v oznake JOS, ki je poleg samih oznak zajemal tudi spremembe v pripisovanju lem, sta med drugim pokazali tudi probleme označevanja na ravni zlasti funkcijskih besednih vrst, zato je bila sprejeta odločitev, da se za označevanje v projektu JOS izdelajo leksikoni za izbrane zaprte besedne vrste (zaimek, predlog, veznik, členek). Pri izdelavi so bili upoštevani tako podatki iz korpusa kot tudi obstoječih jezikovih priročnikov, uvrstitev posameznih besed v izbrani leksikon pa je bila v nadaljevanju preverjena še na osnovi ročnega označevanja korpusnega vzorca, kot bo razvidno iz sledečega poglavja[5].

3.3.4. Ročno popravljanje avtomatsko označenega testnega podkorpusa

V sklopu projekta JOS je bil razvit postopek pretvorbe izvornih oznak Multext-East v oznake JOS (Erjavec in Krek, 2008). Z uporabo le-tega je bil z novimi oznakami preoznačen podkorpus korpusa FidaPLUS, obsegajoč 100.000 besed. Avtomatsko pripisane oznake so bile nato ročno pregledane ter popravljene. Vsak segment besedila sta pregledala najmanj dva označevalca, v večini primerov po trije. Spremembe oznak so bile nato nadalje analizirane s stališča:

- razlik med avtomatsko ter ročno pripisanimi oznakami[6] ter
- razlik med oznakami človeških označevalcev[7].

Na osnovi v tej fazi pridobljenih informacij so bili nova kodna tabela ter leksikoni še enkrat evalvirani. V samem testnem podkorpusu so bile razlike v označevanju naknadno poenotene – slednji je bil kot tak na voljo kot izhodišče za naslednje stopnje projekta JOS, tj. za označevanje korpusa JOS velikosti milijon besed.

4. Primer – spremembe nabora oznak za besedno vrsto glagol[8]

4.1. Tabelarni prikaz sprememb

Glede na sistem Multext-East in revidirani sistem JOS slovenskim glagolskim pojavnicam pripisujemo naslednje kategorije in vrednosti:[9]

[3] Marca 2006. Raziskava je internega tipa, primarni cilj analize je bila izboljšava jezikovnega analizatorja, s pomočjo katerega je označevanje referenčnega korpusa potekalo, vendar se je ob analizi izkazalo, da številni primeri neustrezne označenosti izvirajo pravzaprav iz zasnovanosti oblikoskladenjskih kodnih tabel.

[4] Slednja ima v jezikoslovju neizpodbitno tradicijo in je kot taka med uporabniki korpusnih virov močno zasidrana (po tipologiji oznak EAGLES gre npr. za t. i. obvezne oznake). Na ravni pregibnih besednih vrst je vsekakor primerna tudi za avtomatsko obdelavo jezika, problemi se pojavljajo na ravni nepregibnih besednih vrst oz. besednih oblik, ki so glede besednovrstne uvrstitve tudi na ravni leksikona dvoumne (npr. razmerje med členki in prislovi).

[5] Pri čemer je potrebno izpostaviti dejstvo, da so na tak način preverjene seveda le tiste oblike, ki se v korpusnem vzorcu dejansko pojavljajo, ne pa ves leksikonski potencial.

[6] Slednje so brez dvoma pomemben vir podatkov za izboljšavo postopka avtomatskega pripisovanja, žal pa na tem mestu za to temo ni dovolj prostora (nekaj več informacij je na voljo v Erjavec in Krek, 2008).

[7] Potrebna konsistentnost pri pregledovanju označevanja je bila omogočena na osnovi pripravljenih navodil za označevalce (ki so na voljo na internetni strani projekta JOS). Razlike, o katerih je govora na tem mestu, so večinoma posledica možnosti različnih interpretacij besedila.

[8] Pri glagolu kot najkompleksnejši besedni vrsti so se med primerjanimi sistemi oblikoskladenjskega označevanja kazale tudi največje razlike.

P	Kategorija	Vrednost	Koda
0	besedna vrsta	glagol	G
1	vrsta	polnopomenski naklonski vezni	p n v
2	glagolska oblika	povednik velelnik pogojnik nedoločnik deležnik namenilnik	p v g n d m
3	čas	sedanjik prihodnjik nesedanjik	s p r
4	oseba	prva druga tretja	1 2 3
5	število	ednina množina dvojina	e m d
6	spol	moški ženski srednji	m z s
7	način	tvornik trpni deležnik	t r
8	nikalnost	nezanikani zanikani	n d
…	…	…	…
14	vid	nedovršni dovršni	n d

Tabela 1: Nabor kategorij in vrednosti zanje za glagol v označevalnem sistemu Multext-East.

P	Kategorija	Vrednost	Koda
0	besedna vrsta	glagol	G
1	vrsta	glavni pomožni	g p
2	vid	dovršni nedovršni dvovidski	d n v
3	glagolska oblika	nedoločnik namenilnik deležnik sedanjik prihodnjik pogojnik velelnik	n m d s p g v
4	oseba	prva druga tretja	p d t
5	število	ednina dvojina množina	e d m
6	spol	moški ženski srednji	m z s
7	nikalnost	nezanikani zanikani	n d

[9] Slovenski glagoli kategorij *določnost, klitičnost, sklon, živost, klitičnost s -s* ter *vljudnost*, ki jih sistem Multext-East predvideva na pozicijah 9–13 ter 15, ne izkazujejo.

Tabela 2: Nabor kategorij in vrednosti zanje za glagol v označevalnem sistemu JOS.

4.2. Diskusija

4.2.1. Vrsta glagola

Do razhajanja v vrednostih kategorij obeh sistemov prihaja že na ravni označevanja vrst glagolov. Multext-East predvideva vrednosti *polnopomenski*, *naklonski* in *vezni* glagol, sistem JOS pa vrednosti *glavni* in *pomožni* glagol. Razlog za spremembo nabora vrednosti je dejstvo, da vrednost *naklonski* glagoli izstopa glede na hierarhijo pomembnosti informacij, ki jih označevalni sistem nudi, pri čemer je potrebno upoštevati dva argumenta.

Če je bil pri vzpostavitvi sistema Multext-East za slovenščino razlog za umestitev te vrste glagolov v posebno skupino na prvem hierarhičnem nivoju formalnoskladenjski (tj. tvorjenje zvez z nedoločnikom), bi pričakovali, da bodo v to skupino vključeni vsi glagoli, ki takšne zveze lahko tvorijo (npr. tudi fazni glagoli, ki po tem sistemu ostanejo nezajeti). Hkrati je treba opozoriti, da zveze teh glagolov z nedoločniki skladenjsko oz. pomensko-skladenjsko niso enakovredne (prim. npr. *morati delati* : *dovoliti (komu) delati*) in da jih vseh ne moremo obravnavati kot nosilce predvsem fleksijskih kategorij, saj precejšnje število naklonskih glagolov ne ohranja samo slovnične pomenskosti[10].

Če je bil razlog semantičen (izražanje modalnosti), potem je kategorija v hierarhijo informacij postavljena previsoko, saj v slovenščini modalnost izražamo na različne načine, ne le z naklonskimi glagoli.

4.2.2. Glagolski čas in način

Večina slovenskih glagolskih oblik je zloženih, zato bi bila raven razdvoumljanja pojavnic – če bi njihove kategorije in vrednosti zanje sledile predvsem uveljavljenim opisom slovenskega oblikoslovja – skladenjska. Zaradi odločitve, da oblikoskladenjsko označevanje v sistemu JOS temelji v največji možni meri na enobesedni enoti, sta bili kategoriji *čas* in *način* iz nabora izločeni.[11] Kadar so vezane na eno pojavnico, so vrednosti kategorij *čas* in *način* po novem obravnavane v okviru kategorije *glagolska oblika*, skladno z obravnavanjem kategorije glagolskega naklona.

Do obstoja kvalitetnega skladenjskega označevanja slovenščine bo v korpusnih virih podatke o času (in naklonu) glagolske oblike mogoče razmeroma enostavno poiskati z upoštevanjem značilne sintagmatike glagolskih pojavnic, pri čemer je relevantno tudi dejstvo, da informacije o glagolskem času in načinu nekaterih besednih oblik prinašata vrednosti *sedanjik* in *prihodnjik*,

[10] Zdi se, da je bila kategorija *naklonskih* glagolov v sistemu označevanja Multext-East izločena kot posebna vrednost na podlagi primerjave vloge tovrstnih glagolov v slovenščini in angleščini, pri čemer je treba poudariti, da je kategorija modalnih pomožnikov v angleščini za razliko od slovenščine jasno prepoznavna in v opisih izločena na podlagi istih slovničnih lastnosti, ki kategorijo opredeljujejo.

[11] "Zloženih glagolskih oblik (tj. kategorij, ki se izražajo z zloženo glagolsko obliko, op. N. L.) na oblikoskladenjski ravni označevanja ne obravnavamo, ker v večji strukturi združujejo kombinacijo več glagolov.« (Eagles 1996: 8; prev. N. L.)

ki ju označevalni sistem JOS prinaša v sklopu kategorije *glagolska oblika*.

Kategorija *glagolski način* je bila iz sistema oznak JOS izpuščena tudi zato, ker se je pri analizi skladnosti ročnega označevanja pri označevalcih izkazalo, da je diskrepanca med pripisanimi vrednostmi za potencialno kategorijo glagolskega načina prevelika, da torej označevalci v realnih besedilih med trpnimi strukturami in strukturami s pomenom stanjskosti[12] pogosto ne razločujejo – kar je seveda indikacija za to, da z metodami avtomatskega razdvoumljanja relevantnega podatka o tej kategoriji najbrž ne bo mogoče dobiti.

Poleg tega je bilo na podlagi analize frekvence pojavljanja deležniških oblik na -n in -t v različnih strukturah vzorca besedil iz korpusa FidaPLUS ugotovljeno, da gre večinoma za deležnike stanja in da jih je kot take bolj smiselno uvrstiti v besedno vrsto pridevnik, in sicer na prvi hierarhični nivo, kot posebno vrsto pridevnika. V sistemu JOS so torej kot deležniki v okviru besedne vrste glagol obravnavani samo opisni deležniki na -l, ostali deležniki so umeščeni med pridevnike.

4.2.3. Glagolski vid

Da bi zadostili kriterijema strnjenosti in berljivosti oznak, je bila v označevalnem sistemu JOS kategorija *vid* prestavljena na drugo pozicijo v kodni tabeli. Obravnava te kategorije v korpusu je leksikonskega tipa.[13] Vrednostma *dovršni* in *nedovršni* je dodana še vrednost *dvovidski*, in sicer za označevanje glagolov, pri katerih njihovega vidskega statusa iz same oblike pojavnice enoznačno ni mogoče določiti, pri čemer lahko gre za znotraj- in »medleksemsko« dvovidskost. Do znotrajleksemske dvovidskosti pride, kadar je vidski status (več sememov večpomenskega) glagolskega leksema v različnih besedilnih okoljih različen (*proizvodnjo so mehanizirali : z vadenjem klavirja mehaniziram gibe prstov*). Gre za glagole, ki so v jeziko(slo)vnih priročnikih označeni kot dvovidski. Z izrazom »medleksemska« dvovidskost pa označujemo pojav homonimije glagolskih oblik najmanj dveh različnih glagolskih leksemov z različnima vrednostma za kategorijo vida (*kokoš leže jajca : otrok leže v posteljo*).

Določanje vrednosti kategorije *vid* je torej odvisno od razpoznave semantičnih informacij o pojavnicah, vendar pa so slednje v ustrezni meri nadomestljive s podatki, ki jih prinaša leksikalnopodatkovna zbirka ustreznega tipa. Gradnja tovrstne zbirke je za slovenščino predvidena (in za kvalitetno avtomatsko označevanje naravnega jezika na vseh nivojih zaželena), za potrebe označevanja korpusa JOS pa je bila oblikovana manj obsežna leksikonska baza podatkov o glagolskem vidu glagolskih lem.

Za določanje vidskega statusa vseh ustreznih glagolskih pojavnic v ročno pregledanem 100.000-besednem podkorpusu so bili analizirani vzorčni segmenti konkordančnih nizov iz korpusa FidaPLUS z ustrezno glagolsko pojavnico v jedru niza in upoštevani podatki v leksikonskih virih (SSKJ, SP, Besedišče slovenskega jezika, Veliki slovar tujk ipd.), opravljena pa je bila tudi analiza razlik v pripisovanju vidskih vrednosti pojavnicam. Vidski status pojavnic, pri katerih je pri pripisovanju vidskih vrednosti prišlo do razlik, je bil ponovno še natančneje analiziran in napake pri označevanju odpravljene, vse omenjene raziskave pa so bile osnova za oblikovanje omenjene leksikonske baze.

V nastajajočem avtomatsko označenem in delno ročno korigiranem milijonskem korpusu JOS bo kategorija glagolskega vida ohranjena, pri čemer bo zaenkrat za pojavnice privzet vid, kot jim je pripisan v korpusu FidaPLUS. Z oznakami, za glagol predvidenimi v kodni tabeli JOS, bodo označene le pojavnice lem, za katere je leksikonska podpora JOS že oblikovana.

5. Ostale spremembe kodne tabele

Na podlagi zgoraj opisanih kriterijev so bile zasnovane tudi spremembe nabora kodne tabele pri drugih besednih vrstah (in njim podobnih uporabnostnih kategorijah). Okviren kvantitativen pregled kategorij in njihovih vrednosti ter sprememb med naboroma Multext-East (MTE) ter JOS prikazuje Tabela 3.

Besedna vrsta	Kategorije in št. vrednosti (MTE)	Kategorije in št. vrednosti (JOS)	Razlike v številu (kategorij; vrednosti)
Samostalnik	vrsta (2) spol (3) število (3) sklon (6) živost (2)	vrsta (2) spol (3) število (3) sklon (6) živost (2)	
Glagol	vrsta (3) vid (2) oblika (6) čas (3) oseba (3) število (3) spol (3) način (2) nikalnost (2)	vrsta (3) vid (3) oblika (7) oseba (3) število (3) spol (3) nikalnost (2)	2; 7
Pridevnik	vrsta (3) stopnja (4) spol (3) število (3) sklon (6) določnost (2) živost (2)	vrsta (3) stopnja (3) spol (3) število (3) sklon (6) določnost (2)	1; 3
Prislov	stopnja (4)	stopnja (3) deležje (2)	1; 3
Zaimek	vrsta (9) oseba (3) spol (3) število (3) sklon (6) število svojine (3)	vrsta (9) oseba (3) spol (3) število (3) sklon (6) število svojine (3)	3; 8

[12] Upoštevan je trpnik, tvorjen z deležnikom na -n/-t. Trpnik s *se* v oblikoskladenjskih označevalnih sistemih navadno ni predpostavljen zaradi kompleksnosti avtomatske določitve statusa besede/morfema *se* (prim. Verdonik, Rojc in Kačič, 2003).

[13] Potreben je poudarek, da oblikoskladenjsko označevanje ni vezano na leksikonske informacije samo v primerih, ki so izpostavljeni v članku. Korpus FidaPLUS, katerega oznake so izhodišče za pretvorbo večine oznak, je bil v celoti označen na osnovi informacij iz leksikalne baze ASES podjetja Amebis.

	spol svojine (3) oblika (3) nanašanje (2) skladenjska vloga (3) živost (2)	spol svojine (3) oblika (2)	
Števnik	zapis (3) vrsta (4) spol (3) število (3) sklon (6) določnost (2) živost (2)	zapis (3) vrsta (4) spol (3) število (3) sklon (6) določnost (2)	1; 2
Predlog	sestavljenost (2) sklon (6)	sklon (6)	1; 2
Veznik	vrsta (2) oblika (2)	vrsta (2)	1; 2
Členek	Ni kategorij in vrednosti.	Ni kategorij in vrednosti.	
Medmet	Ni kategorij in vrednosti.	Ni kategorij in vrednosti.	
Okrajšava	Ni kategorij in vrednosti.	Ni kategorij in vrednosti.	
Neuvrščeno	Ni kategorij in vrednosti.	vrsta (3)	1; 3

Tabela 3: Pregled razlik med naboroma oblikoskladenjskih oznak Multext-East in JOS.

Sam kvantitativen pregled sicer ne izkazuje dejanskih razsežnosti sprememb označevalnega sistema, saj je – kot kaže zgled za glagol – bistvena predvsem preureditev kategorij ter vrednosti oz. njihovih mej in sistema njihovega pripisovanja pojavnicam. Vsa dokumentacija o spremembah kodnih tabel je zato na voljo na internetni strani JOS (http://nl.ijs.si/jos/), skupaj z navodili za označevanje ter označenimi korpusi.

6. Sklep

Za nadaljevanje dela so predvideni še nekateri koraki, ki so v večji meri vezani na odziv zainteresirane raziskovalne javnosti na pripravljeni označevalni sistem. V prvi vrsti je zaželena priprava načinov za avtomatsko pretvorbo med obstoječimi označevalnimi sistemi ter sistemom JOS. Predvidena je tudi robustizacija sistema oznak JOS, katere rezultat bi bil možnost označevanja z različno granuliranimi nabori oznak, kot je za označevanje angleščine na voljo denimo v sistemu Claws (http://ucrel.lancs.ac.uk/claws/).

Razen v okviru projekta Jezikoslovno označevanje slovenščine (http://nl.ijs.si/jos/) je predvideno nadaljnje delo tudi v okviru projekta Sporazumevanje v slovenskem jeziku (http://www.slovenscina.eu), s ciljem razvoja označevalnega sistema za skladenjsko označevanje slovenščine[14] oz. prosto dostopnega avtomatskega skladenjskega razčlenjevalnika (ang. *parser*) za slovenski jezik.

7. Literatura

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

EAGLES (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R, ILC, Pisa: <http://www.ilc.cnr.it/EAGLES/annotate/annotate.html > (dostopno 19. 6. 2008).

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (pp. 1535--1538). Pariz: ELRA.

Erjavec, T. in Krek, S. (2008). The JOS morphosyntactically tagged corpus of Slovene. Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2008. Pariz: ELRA.

Gorjanc, V. (2005). Uvod v korpusno jezikoslovje. Domžale: Izolit.

Jakopin, P. in Bizjak, A. (1997). O strojno podprtem oblikoslovnem označevanju slovenskega besedila. Slavistična revija 45 (3–4), 513--532.

Jezikoslovno označevanje slovenščine: <http://nl.ijs.si/jos/> (dostopno 19. 6. 2008).

Lönneker, B. in Jakopin, P. (2004). Checking POSBeseda, a Part-of-Speech Tagged Slovenian Corpus. Jezikovne tehnologije: zbornik konference Informacijska družba IS 2004 (pp. 48--55). Ljubljana: IJS.

McEnery, T., Xiao, R. in Tono Y. (2006). Corpus-Based Language Studies. An Advanced Resource Book. London: Routledge.

Označevalni sistem Ajka: <http://nlp.fi.muni.cz/projekty/ajka/tags.pdf> (dostopno 19. 6. 2008).

Označevalni sistem Claws: <http://ucrel.lancs.ac.uk/claws/> (dostopno 19. 6. 2008).

Označevalni sistem Češkega nacionalnega korpusa: <http://ucnk.ff.cuni.cz/bonito/> (dostopno 19. 6. 2008).

Verdonik, D., Rojc, M. in Kačič, Z. (2003). Analiza jezikovnih vprašanj, nastalih pri gradnji SImflexa – oblikoslovnega in glasoslovnega slovarja za slovenski knjižni jezik. Raziskovalno delo podiplomskih študentov v Sloveniji – ena znanost: e-zbornik (pp. 434—443). Ljubljana, Društvo mladih raziskovalcev Slovenije.

Verdonik, D., Rojc, M. in Kačič, Z. (2004). Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004 (pp. 1399--1402). Pariz: ELRA.

[14] Ta del je predviden in deloma že izpeljan v sklopu projekta Jezikoslovno označevanje slovenščine. Razvoj skladenjskega razčlenjevalnika pa je eden od ciljev projekta Sporazumevanje v slovenskem jeziku.

Vpliv namembnosti korpusa na označevanje besedilnega gradiva za »Večjezični korpus turističnih besedil«

Vesna Mikolič, Ana Beguš, Davorin Dukič, Miha Koderman
Inštitut za jezikoslovne študije, Znanstveno-raziskovalno središče Koper, Univerza na Primorskem
Garibaldijeva 1, 6000 Koper
vesna.mikolic@zrs.upr.si, ana.begus@zrs.upr.si; davorin.dukic@zrs.upr.si; miha.koderman@zrs.upr.si

Povzetek

Prispevek najprej predstavi jezikovnotehnološki projekt Večjezični korpus turističnih besedil: informacijski vir in analitična baza slovenske naravne in kulturne dediščine. Cilj projekta je zgraditi primerljivi in delno vzporedni korpus besedil s področja turizma v slovenskem, italijanskem in angleškem jeziku, ki bo uporaben kot prevajalski vir, za jezikoslovne raziskave in raziskave turizma. Prispevek nato opiše in utemelji metabesedilno in oblikoslovno označevanje Večjezičnega korpusa turističnih besedil glede na načrtovane namene njegove uporabe.

The Use of the Multilingual Corpus of Tourist Texts and Its Influence on the Annotation of the Corpus

The article first presents the project 'Multilingual corpus of tourist texts: information source and analytical database of Slovene natural and cultural heritage'. The aim of the project is to build a comparable and partly parallel corpus of tourist texts in Slovene, Italian and English. The corpus will be used as a translation resource and for research in linguistics and tourism. The article then describes the procedure of metatextual and morphological annotation of the Multilingual corpus of tourist texts on the basis of the planned uses of the corpus.

1. Uvod

Večjezični korpus turističnih besedil je jezikovnotehnološki projekt, ki nastaja na Znanstveno-raziskovalnem središču Univerze na Primorskem pod vodstvom dr. Vesne Mikolič, financira pa ga Agencija za raziskovalno dejavnost Republike Slovenije (ARRS) kot tematsko usmerjeni temeljni raziskovalni projekt. Cilj projekta je zgraditi primerljivi in delno vzporedni korpus turističnih besedil v slovenskem, italijanskem in angleškem jeziku. Odločitev za takšen korpus izhaja iz dejstva, da je potrebno turizem pojmovati kot sestavljeno dejavnost, zaradi česar bi moralo biti preučevanje turizma izrazito interdisciplinarno. Jezikovni vidiki turizma se tako vežejo na jezikovno pogojenost konstrukcije identitete tistih elementov naravne in kulturne dediščine ali turistične infrastrukture, ki jih želimo predstaviti kot primarno ali sekundarno turistično ponudbo.

Korpus bo uporaben predvsem kot prevajalski vir, za raziskave turizma ter naravne in kulturne dediščine ter za analizo turističnega diskurza ter različne jezikoslovne raziskave. Na osnovi projekta večjezičnega korpusa v sodelovanju in ob sofinanciranju Slovenske turistične organizacije poteka tudi projekt izgradnje korpusno-podprtega turističnega terminološkega slovarja.[1] Slovar bo dvojezičen, angleško-slovenski ter razlagalen. Pri projektu bo kot pomemben terminološki vir uporabljen tudi *Tezaver turizma in prostočasnih dejavnosti* Svetovne turistične organizacije (World Tourism Organization, 2001).

2. Opis in dejavnosti projekta

Projekt izgradnje korpusa se je pričel leta 2006, ko so bile postavljeni osnovni vsebinski in metodološki temelji projekta, izbrana projektna skupina in zadolžitve posameznih članov ter opravljene priprave na fazo zajemanja besedil. Fazo vzpostavljanja stikov s potencialnimi besedilodajalci in predstavitev projekta lahko ocenimo za uspešno: zavrnitev sodelovanja skorajda ni bilo, z izjemo nekaterih institucij, kjer so nastale težave zaradi avtorskih pravic.[2]

Vprašanje načrtovanja reprezentativnosti in uravnoteženosti večjezičnega in specializiranega korpusa je bilo izpeljano skozi trud, da bi zajeli vse ključne institucije, ki se s turizmom ukvarjajo. Ključni kriterij pri izboru besedil je bila prisotnost besedila v slovenskem jeziku, zato smo želeli v korpus vključiti čim večji obseg institucij, pri katerih lahko pričakujemo, da se bo ob tujejezičnih besedilih pojavilo tudi besedilo v slovenščini. To so institucije, ki delujejo v t.i. slovenskem kulturnem prostoru (torej tako v okviru slovenskega državnega prostora kot v slovenskem zamejstvu) ali širše, npr. v slovenskem izseljenstvu, v evropskem prostoru in širšem mednarodnem kontekstu. Turizem je namreč eno izmed tistih družbenih področij, kjer prihaja do intenzivnih medjezikovnih in medkulturnih stikov, zato se tudi slovenski jezik na tem področju sooča s tujimi jeziki. Korpus tako vključuje pretežno besedila v slovenskem, angleškem in italijanskem jeziku, v približnem obsegu 30

[1] Namen projekta Turistični terminološki slovar je na osnovi rezultatov projekta Večjezični korpus, tj. urejenega korpusa turističnih besedil ter opravljenih jezikovnih analiz v njem, zbrati terminologijo turizma in jo urediti v slovarski (knjižni in elektronski) format, ki je uporabnikom bolj poznan in zato bližji, hkrati pa preko povezave slovarja s korpusom pokazati na pomen in prednosti sodobnih elektronskih jezikovnih virov, tudi za področje strokovnega jezika s področja turizma.

[2] Težave z avtorskimi pravicami pri zajemu besedil v korpus, ki lahko močno vplivajo tudi na reprezentativnost korpusa, so v strokovni literaturi dobro izpričane, gl. med drugim Gorjanc (2005), Stabej (1998).

milijonov besed, kar ga uvršča med večje večjezične korpuse, ki vključujejo slovenski jezik. [3]

Besedilodajalce lahko razdelimo v tri glavne skupine. Prvo skupino sestavljajo strokovne organizacije, ki se ukvarjajo s turizmom kot dejavnostjo, skrbijo za turistično promocijo države ali regije in predstavljajo ter tržijo turistično ponudbo, ter ponudniki turističnih storitev, ki neposredno tržijo določene turistične destinacije ali znamenitosti. V to skupino lahko uvrstimo npr. Slovensko turistično organizacijo, Turistično zvezo Slovenije, Skupnost slovenskih naravnih zdravilišč, Urad za turizem Furlanije Julijske krajine, Agencijo Kompas. V drugo skupino lahko umestimo strokovne ali raziskovalne organizacije, ki proučujejo turizem ali skrbijo za razvoj turistične stroke in vede, ne tržijo pa neposredno turistične ponudbe. Sem lahko med drugim uvrstimo univerzitetne zavode, kot so UP Turistica – Fakulteta za turistične študije, Znanstveno-raziskovalno središče Univerze na Primorskem ter Visoko šolo za gostinstvo in turizem Bled. V tretji skupini najdemo medije in založbe, ki pišejo o turizmu ali predstavljajo turistično ponudbo, npr. nekatere večje slovenske časopise, kot sta Delo in Dnevnik, ter specializirane revije za turizem, kot je revija Horizont.

Po končani fazi zbiranja in označevanja gradiva je bil za korpus izdelan konkordančnik, ki je dostopen na www.jt.upr.si/turisticnikorpus.

3. Označevanje korpusa

3.1. Metabesedilno označevanje

Z metabesedilnim označevanjem mislimo na označevanje besedil glede na nekatere njihove lastnosti; gre za podatke, ki bodo vidni nad vsakim izpisanim zadetkom. Besedila so se razvrščala po za ta namen pripravljenih kategorijah. Ker je moral označevalec vsako besedilo videti, opredeliti njegove temeljne lastnosti in ga šele nato ustrezno označiti, je to označevanje potekalo ročno.

3.1.1. Tipologija in stopenjskost oznak

Vsakemu besedilu v korpusu smo skušali določiti pet lastnosti: 1) mesto, kjer se pojavlja oz. prenosnik; 2) temeljno funkcijo, ki jo opravlja, oz. namen, ki ga skuša uresničiti, torej zvrstnost; 3) ali je lektorirano ali ne; 4) jezik, v katerem je napisano; 5) s katerim področjem turistične dejavnosti je povezano.

Poudariti je potrebno, da je tipologija oznak pri prvih dveh lastnostih omogočala zelo natančno opredelitev besedila, saj so bile oznake razdeljene na več nivojev oz. stopenj; na splošnejši stopnji sta natančnost in informativnost manjši, na podrobneje razvrščeni pa večji. Pri vsakem besedilu smo skušali biti čimbolj informativni, torej besedilo čim natančneje/podrobneje označiti. Stopenjskost oznak se je izkazala za zelo koristno pri »problematičnih« besedilih, kjer je bilo posamezno lastnost težko podrobno opredeliti, zato je bilo varneje ostati pri višji oznaki.

3.1.1.1. Oznake glede na prenosnik, lektoriranost in jezik

Kriterij prenosnika je bil povzet po korpusu FidaPLUS. Osnovni nivo razlikovanja predstavljajo oznake 'govorni', 'elektronski' in 'pisni'. Ker korpus govorjenih besedil ne zajema, sta v poštev prišla le elektronski in pisni prenosnik. Pri elektronskem prenosniku smo bili pozorni na ločevanje med besedili, ki so bila oblikovana za splet in torej nosijo s seboj značilnosti spletnega besedila (npr. morebitni avdio in video dodatki, elektronske povezave, besedila, ki nastajajo preko elektronske pošte), ter tistimi, ki so primarno nastala kot tradicionalna (za)pisana besedila, pa so bila zgolj prenesena oz. postavljena na splet (npr. monografije in članki v elektronski obliki, sporočila, ki se pojavijo tako v periodiki kot tudi na spletu); slednja smo uvrstili pod pisni prenosnik. Drugi nivo razlikovanja izhaja iz pisnega prenosnika, loči pa med 'objavljenim' in 'neobjavljenim'. Neobjavljeno gradivo se nadalje deli na 'javno' (npr. sporočilo s spletnega foruma), 'interno' (npr. zapisnik s seje) in 'zasebno' (npr. elektronska pošta), objavljeno pa loči med 'časopisnimi' in 'revialnimi' publikacijami. Časopisne publikacije so glede na pogostost izhajanja razdeljene na 'dnevno', 'večkrat tedensko' in 'tedensko', revije pa glede na isti kriterij na 'tedensko', 'štirinajstdnevno', 'mesečno', 'redkeje kot na mesec' in 'občasno'.

Kriterij lektoriranosti, ki je prav tako povzet po korpusu FidaPLUS, loči le med oznakama 'da' za besedila, ki so opredeljena kot lektorirana oz. kot tista, pri katerih lahko na lektoriranost nedvoumno sklepamo (npr. pri časopisnih člankih, diplomskih nalogah, monografijah), ter 'ne' za besedila, ki so nelektorirana (npr. takšna, ki nastajajo v interni ali zasebni komunikaciji). Kadar besedilu (ne)lektoriranosti nismo mogli z gotovostjo pripisati, smo pustili kriterij neoznačen.

Ker korpus zajema besedila v slovenskem, angleškem in italijanskem jeziku, je tudi nabor oznak glede na jezik razdeljen le na tri: 'slovenščino', 'angleščino' in 'italijanščino'. Primerov, ko se v okviru enega besedila pojavljata dva ali več različnih jezikov (npr. nekatera besedila spletnih strani), v korpus nismo vključevali. Najprej smo želeli označiti tudi, ali je besedilo izvirnik ali prevod, vendar se je po premisleku izkazalo, da je pri večini besedil (npr. pri reviji Slovenia Times ali pri delu turističnopromocijskega gradiva) to pravzaprav težko nedvoumno določiti, zato smo to oznako izpustili. Delno indikacijo o tem, ali je besedilo prevod ali izvirnik, lahko dobimo iz informacije o besedilodajalcu.

[3] Med večjezičnimi korpusi v slovenskem prostoru je potrebno omeniti najprej korpus prevodov zakonodaje Evropske unije Evrokorpus (www.gov.si/evrokorpus), ki obsega skupaj približno 113 milijonov besed, novi 22-večjezični vzporedni korpus (http://evrokorpus.gov.si/k2/) ter slovensko-angleški vzporedni korpus zakonodaje IJS SVEZ ACQUIS (http://np.ijs.si/svez). Omeniti je potrebno še korpusa Elan (http://nl.ijs.si/elan/) in TRANS, ki oba obsegata po približno en milijon besed. V teku so tudi novi projekti večjezičnih korpusov, med drugim npr. slovensko-francoskega korpusa, ki nastaja na Oddelku za prevajalstvo Filozofske Fakultete v Ljubljani. Poleg tega naj omenimo še večjezične korpuse manjšega obsega, ki so bili narejeni z namenom specifične jezikoslovne raziskave; več o tem v Beguš, Dukič (ur.), 2008.

3.1.1.2. Oznake glede na funkcijsko zvrstnost

Na osnovnem nivoju razlikovanja besedil glede na zvrstnost korpus ločuje med 'umetnostnimi' in 'neumetnostnimi' besedili; prvih korpus turističnih besedil ne vključuje. Neumetnostna besedila se nadalje delijo na šest velikih skupin: 'znanstvena', 'strokovna', 'poslovna', 'oglaševalska', 'publicistična' in 'pravna' besedila.[4] Znanstvena besedila ločimo na 'znanstvene monografije', 'znanstvene članke' ter 'magistrska in doktorska dela'. Strokovna besedila se členijo na 'naravoslovna in tehnična' ter na 'humanistična in družboslovna', slednja pa še naprej na 'strokovne monografije', 'strokovne članke', 'poljudnoznanstvene monografije', 'poljudnoznanstvene članke', 'diplomske naloge', 'učbenike' in 'drugo'. Poslovna besedila delimo na 'korespondenco' (tu je mišljena običajna pošta) ter na 'elektronsko pošto'. Oglasna besedila se delijo na 'oglasna sporočila', 'letake, prospekte, brošure, kataloge', 'prodajne oglase', 'oglase na prostem', 'vodnike' in 'drugo'. Pri publicističnih besedilih razlikujemo med 'novinarskimi članki' (npr. vest, poročilo, članek, komentar), 'strokovnimi novinarskimi članki', 'reportažami', 'članki v turističnih prilogah' in 'drugim'. Pravna besedila so edina skupina, ki ostaja le pri tej prvostopenjski oznaki.[5]

Navkljub dokaj razdelanemu naboru oznak smo naleteli na veliko besedil, ki jih je bilo težko natančno in nedvoumno označiti. Razlog za to gre najverjetneje iskati v veliki hibridnosti funkcijske zvrstnosti, ki se odraža v tem, da se v posameznih besedilih prepletajo lastnosti, značilne za dve ali celo za več posameznih zvrsti.[6] V takih primerih se je moral označevalec odločiti glede na prevladujoče značilnosti besedila; drugo rešitev je ponujala sama stopenjskost oznak, ki je omogočala ustavitev označevanja pri višji ravni, torej pri tisti, kjer o ustrezanju oznake konkretnemu besedilu ni bilo dvoma.

Gotovo je, da bi manj kompleksna klasifikacija v tem pogledu dajala zanesljivejše, predvsem pa konsistentenjše rezultate, vendar se je projektni skupini zdelo pomembno, da se posamezna besedila, kjer je to nedvoumno, za kasnejše raziskovalne namene vseeno definira do najgloblje stopnje, pri čemer se zanesljivost na višji ravni ne zmanjša. Po drugi strani pa se je treba zavedati, da takšno označevanje vedno vključuje tudi določeno mero interpretativnosti, saj je od posameznega označevalca odvisno, ali bo šel pri posameznih oznakah globlje ali se bo ustavil že na višji ravni in tako zadostil večji nedvoumnosti.

3.1.1.3 Oznake glede na zvrsti turistične dejavnosti

Definiranje turističnih zvrsti oziroma dejavnosti v procesu priprave korpusa in tudi samega označevanja besedil je zaradi kompleksnosti in širine področij, ki jih turistična dejavnost obsega, predstavljalo avtorjem in označevalcem precejšen izziv. Ob pregledu obstoječih klasifikacij in delitev posameznih zvrsti znotraj turistične dejavnosti je bil za potrebe označevanja besedil v sklopu korpusa v veliki meri upoštevan *Tezaver turizma in prostočasnih dejavnosti* Svetovne turistične organizacije (World Tourism Organization, 2001), ki s svojim podrobnim vsebinskim razločevanjem turističnih dejavnosti (24 podzvrsti) jasno definira poglavitne zvrsti znotraj te ekonomske dejavnosti. Preostale obravnavane klasifikacije so turistične aktivnosti delile večinoma z ekonomskega in le delno tudi tematskega vidika; nekatere so razločevale med prostočasnimi in poslovnimi dejavnostmi (Swarbrooke in Horner, 1999), druge med izvornimi državami turistov.

Zaradi izrazite usmerjenosti posameznih zvrsti turizma na nekatera »nišna« področja znotraj turistične dejavnosti in hkrati smiselne vsebinske dopolnitve same klasifikacije za označevanje besedil, smo izbrano zvrstno delitev Svetovne turistične organizacije dopolnili še s tremi dodatnimi zvrstmi turističnih aktivnosti (glej Prilogo 1). Dodane zvrsti so »igralniški« in »kulinarični turizem,« ki se navezujeta na relevantne turistične aktivnosti znotraj teh podzvrsti, ter splošna kategorija »turizem«, v katero so bila uvrščena besedila, ki jih bodisi zaradi pomanjkanja ključnih pojmov bodisi zaradi prepletanja več podzvrsti ni bilo mogoče umestiti v posamezno podzvrst.

Najpogostejše dileme, ki so se v procesu označevanja besedil občasno pojavljale pri razvrščanju le-teh po predstavljenih turističnih podzvrsteh, so bile vezane na prisotnost in prepletanje več posameznih podzvrsti oziroma panog znotraj istega besedila. V primeru takšnih zadreg je bil potreben temeljitejši pregled besedila, posvet s kolegi in šele nato dokončna uvrstitev, pri kateri se je moral označevalec odločiti za tisto podzvrst, ki je v besedilu prevladovala. Označevalcu je v primeru enakovredne prepletenosti turističnih podzvrsti ostala možnost, da besedilo uvrsti v splošno kategorijo »turizem«. Enako je označevalec ravnal tudi v (sicer redkih) primerih, ko besedilo ni vsebovalo nobene značilnosti posamezne turistične podzvrsti.

3.2 Oblikoslovno označevanje

Večjezičnost in namembnost korpusa – uporaben naj bo kot prevajalski vir, podatkovna baza za jezikoslovne raziskave in analize turizma – je porodila razmišljanja, kako oz. koliko oblikoslovno označevati. Kratek pregled literature o oblikoslovnem označevanju korpusov v slovenskem raziskovalnem področju, ki bi v tem pogledu ponudil uporabne informacije, je vezan predvsem na razvoj in aplikacijo označevalnih naborov za enojezične korpuse: korpus Fida in nadgradnjo FidaPLUS (Erjavec 1998a, 1998b, 2003) ter korpus Beseda in Nova beseda (Jakopin 1997, 1999). Podatkov o označevanju večjezičnih korpusov, ki vsebujejo slovenščino, je manj ali niso javno dostopni. Miran Željko, urednik Evrokorpusa in povezane terminološke baze Evroterm, ki dnevno beleži tudi do 50.000 obiskov, meni, da je za velike korpuse namesto oblikoslovnega označevanja bolj smiselno uporabljati iskanje s krnjenjem (ang. stemming),

[4] Osnovna razdelitev je bila povzeta po V. Mikolič (2007, 109–11), ki tipologijo besedilnih zvrsti oblikuje glede na namen besedila oz. njegovo vplivanjsko vlogo.

[5] Gre za sorazmerno lahko prepoznavna besedila z visoko stopnjo konvencionaliziranosti oblike, kot so zakoni, pogodbe, odločbe, akti, sklepi ipd.

[6] Največ prepletanja je bilo med publicističnimi in oglaševalskimi besedili, med strokovnimi in znanstvenimi, pa tudi poslovna besedila so se izkazala za visoko hibridno zvrst.

na isti način, kot deluje spletni iskalnik najdi.si.[7] Razvoj interneta in dejavnosti semantičnega spleta gotovo sili v razmislek, kakšne naj bodo uporabne in hkrati kakovostne informacijske baze, in s tem tudi jezikovni korpusi, v prihodnosti. Oblikoslovna neoznačenost nedvomno nudi določene prednosti, saj omogoča enostavnejše in hitrejše posodabljanje in večanje baze jezikovnih podatkov, kot je to v primeru Evrokorpusa, ki pa je specifičen v tem, da nastaja sočasno s prevajanjem dokumentov in se terminološki vnosi v bazo vnašajo sproti. Pri korpusu turističnih besedil pa gre za tradicionalni korpusni postopek – zbiranje že dostopnega gradiva in njegova obdelava. Zato smo se odločili, da bomo v ta namen uporabili programe avtomatskega luščenja terminologije (Vintar 1999, 2001, 2002), ki dajejo znatno boljše rezultate pri oblikoslovno označenih korpusih. Korpus je bil oblikoslovno označen po naboru JOS, razvitem na Inštitutu Jožefa Štefana v Ljubljani.

4. Sklepne misli ob označevanju korpusa

Vprašanje označevanja je v korpusnem jezikoslovju že samo po sebi pereče vprašanje za vse snovalce jezikovnih korpusov, saj je vsako označevanje korpusa vedno nujno tudi že interpretacija samih podatkov, pri čemer »ne gre za pripisovanje lastnosti, ki bi bile za vselej veljavne jezikovne resnice« (Gorjanc, 2002: 262).

Pri Večjezičnem korpusu turističnih besedil se je za težavnejše izkazalo metabesedilno označevanje, saj so orodja za oblikoslovno označevanje korpusov tudi za slovenščino že zelo razvita in dostopna. Pri poteku samega metabesedilnega označevanja smo ugotovili, da smo navkljub dokaj razdelanemu naboru oznak naleteli na veliko besedil, ki jih je bilo težko natančno in nedvoumno označiti. Najpomembnejši razlog za to sta gotovo kompleksnost in razvejanost same turistične stroke, ki pa prav zato potrebuje sistematično analizo tipologije turističnih besedil glede na različne kriterije. Poleg tega pa takšno metabesedilno razlikovanje in označevanje narekujejo tudi različni nameni gradnje korpusa, ki bo uporaben kot prevajalski vir, za raziskave turizma in različne jezikoslovne raziskave. Metabesedilno označevanje korpusa je zato zelo pomembna faza v gradnji večjezičnega korpusa turističnih besedil v najširšem smislu.

5. Literatura

Beguš, Ana (ur.) in Dukič, Davorin (ur.), 2008. Mednarodni znanstveni sestanek Jezikovne tehnologije v medkulturni komunikaciji. Program in povzetki. Glasnik UP ZRS, let. 13, št. 3.

Erjavec, Tomaž, 1998a. Oznake korpusa FIDA. *Uporabno jezikoslovje*, št. 6, , str. 85-95.

Erjavec, Tomaž, 1998b. Standardizacija zapisa jezikovnih podatkov. V: Erjavec, Tomaž in Gros, Jerneja (ur.), *Jezikovne tehnologije za slovenski jezik. Zbornik konference / Language Technologies for the Slovene Language. Proceedings of the Conference*. Ljubljana, Institut Jožef Štefan, str. 119-123.

Erjavec, Tomaž, 2003. Označevanje korpusov. *Jezik in slovstvo*, let. 48 (2003), št.3-4. Ljubljana: Filozofska fakulteta, str. 61-77.

Gorjanc, Vojko, 2002. Jezikovna infrastruktura: kje je tu slovenščina. V: Krakar-Vogel, Boža (ur.). *Ustvarjalnost Slovencev po svetu. Zbornik predavanj*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete Univerze v Ljubljani, str. 257-270.

Gorjanc, Vojko, 2005. Uvod v korpusno jezikoslovje. Domžale: Izolit.

Jafari, Jafar, 2000. Encyclopedia of Tourism. New York, London: Routledge.

Jakopin, Primož in Bizjak Aleksandra, 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija* 45/3-4, str- 513-532.

Jakopin, Primož, 1999. EVA - an internet tool for textual and lexical resources. V: *Linguistics and language studies / 32nd Annual Meeting,* Ljubljana, 8-11 July 1999. Ljubljana: Faculty of Arts :Societas Linguistica Europaea, 1999, str. 98.

Mikolič, Vesna, 2007. Tipologija turističnih besedil s poudarkom na turističnooglaševalskih besedilih. *Jezik in slovstvo*, let. 52, št. 3-4, maj/avg. 2007, str. 107-116.

Stabej, Marko, 1998. Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje,* št. 6, str. 96-106.

Swarbrooke, John in Horner, Susan, 1999. Consumer behaviour in tourism. Oxford: Butterworth-Heinemann.

Vintar, Špela, 1999. Računalniško podprto iskanje terminologije v slovensko-angleškem vzporednem korpusu. *Uporabno jezikoslovje*, št. 7/8, str. 156-169.

Vintar, Špela, 2001. Using parallel corpora for translation-oriented term extraction. *Babel*, let 47, št. 2, str. 121-132.

Vintar, Špela, 2002. Avtomatsko luščenje izrazja iz slovensko-angleških vzporednih besedil. V: *Jezikovne tehnologije. Zbornik konference / Language technologies: proceedings of the conference.* Ljubljana: Institut Jožef Štefan, str. 78-85.

World Tourism Organization, 2001. Thesaurus on tourism and leisure activities: a structured list of descriptors for indexing and retrieving information on tourism and leisure activities. Madrid: World Tourism Organization.

[7] Podatki so bili pridobljeni preko elektronske pošte z g. Željkom; za prijazno pomoč se mu zahvaljujemo.

Priloga 1: Razvrstitev turističnih in prostočasnih dejavnosti, uporabljenih za označitev besedil v *Večjezičnem korpusu turističnih besedil*, s kratkim pojasnilom posameznih zvrsti

Št.	Zvrsti	Pojasnilo pojma
1.	Religiozni turizem	Vsa religiozno motivirana potovanja, romanja, prodaja religioznih spominkov ter s prenočevanjem povezana dejavnost v romarskih krajih.
2.	Kulturni turizem	Široko pojmovanje vseh dejavnosti, ki so povezane z obiskovanjem kulturnih prireditev, razstav, muzejev in etnografskih posebnosti dežel.
3.	Poslovni turizem	Potovanje, prenočevanje in druge aktivnosti, ki so povezane s poslovno dejavnostjo. Plačnik tovrstnih dejavnosti je podjetje oziroma organizacija, ki ima interes, da njeni predstavniki postanejo turistični »potrošniki«.
4.	Pustolovski turizem	Turizem posvečen aktivnemu preživljanju prostega časa v naravnem okolju in doživljanju novih izkušenj, ki pogosto vključujejo določeno stopnjo tveganja in udeležencem prestavljajo nekakšen osebnostni izziv.
5.	Industrijski turizem	S tem pojmom označujemo obiske industrijskih okolij, ki pogosto vključujejo ogled proizvodnega procesa.
6.	Festivalski turizem	Obiskovanje tradicionalnih letnih kulturnih ali glasbenih prireditev, festivalov.
7.	Izobraževalni turizem	Potovanje z namenom izobraževanja, pogosto organizirano s strani podjetja oziroma organizacije z namenom izpopolnjevanja njenih sodelavcev.
8.	Vrtni turizem	Turistične dejavnosti povezane z agrikulturo, s poudarkom na gostinski in namestitveni ponudbi, ponudbi lokalnih pridelovalcev, ki svojo kmetijsko dejavnost dopolnjujejo tudi s turistično ponudbo.
9.	Luksuzni turizem	Vsa petična potovanja in turistične aktivnosti, ki imajo prestižen pomen in vključujejo eksotične destinacije ter namestitvene objekte najvišjega razreda.
10.	Gorniški turizem	Turizem obiskovanja gora in s tem povezanih aktivnosti: prenočevanja, gostinske ponudbe in prodaje spominkov ter najema gorskih vodnikov in opreme.
11.	Naravni turizem	Naravni turizem se označuje tudi s pojmom »eko« turizem in temelji na okolju prijaznih dejavnostih, ki se bistveno razlikujejo od »klasičnih«, »masovnih« turističnih dejavnosti. Aktivnosti turistov pri tovrstnih dejavnostih se odvija pretežno v naravi, kjer lahko poteka tudi njihovo bivanje.
12.	Turizem bližjih destinacij	S tem pojmom se v veliki meri označuje turistični obisk sosednjih regij, dežel oziroma držav. Tovrsten turizem je še posebno pogost v obmejnih področjih ter v področjih, ki imajo izrazito tranzitno lego.
13.	Zdraviliški turizem	Označuje nastanitev in aktivnosti turistov v zdraviliščih. Zelo pogosto so te aktivnosti povezane s preventivno skrbjo za zdravje, pa tudi z rehabilitacijo po poškodbah, boleznih. Večina aktivnosti je vezanih na ugodne učinke termalne vode.
14.	Turizem za osebe s posebnimi potrebami	Ta zvrst turizma je vezana na individualne turiste ali skupine, ki imajo za svoje turistično udejstvovanje posebne zahteve. Lahko gre za gibalno ovirane osebe, osebe z motnjami v razvoju ali druge, ki pogosto potrebujejo tudi posebno infrastrukturo ali opremo.
15.	Mladinski turizem	Mladinski turizem obravnava potovanja mladih, ki so pogosto vezana na nižje stroške, prenočevanje v (cenejših) mladinskih hotelih ter ostalim aktivnostim, ki zanimajo mlade (zabava, ogled znamenitosti...). Posamezna zvrst znotraj turizma mladih je tudi t.i. turizem »nahrbtnikarjev.«
16.	Turizem za starejše	S tem pojmom označujemo turizem, ki obravnava starejšo populacijo. Specifičnost te dejavnosti je, da jo v nekaterih primerih delno financira inštitucija za zdravstveno zavarovanje, pri kateri je oseba zavarovana.
17.	Avtodomni turizem	Ta zvrst turizma je vezana na potovanje in bivanje v »avtodomih«, torej vozilih, posebej prirejenih za daljša cestna potovanja. Deloma lahko ta pojem zajema tudi bivanje v počitniških prikolicah.
18.	Rečni turizem	Rečni turizem je podzvrst športnega turizma, ki se osredotoča na turistične aktivnosti na in ob rekah.
19.	Obmorski turizem	S pojmom obmorski turizem je zajeta paleta turistični aktivnosti, ki so vezane na prostor v neposrednem vplivnem območju morja.
20.	Morski turizem, navtični turizem	Morski turizem zajema aktivnosti, vezane na morje. Specifičen morski turizem je navtični, ki poleg različnih plovil (čolnov, bark, jadrnic...) zajema tudi z njimi povezano infrastrukturo.
21.	Podeželski turizem	Podeželjski turizem je prostorsko vezan na podeželje, na ruralno okolje. Zajema lahko tudi ogled posameznih kulturnih, etnografskih posebnosti, značilnih za neko pokrajinsko področje.
22.	Podzemni turizem	Turizem vezan na turistični obisk kraških jam ali v turistični namen urejenih rudnikov.
23.	Športni turizem	S precej širokim pojmom »športnega« turizma zajemamo vse s turizmom povezane dejavnosti, katerih glavni namen je športna aktivnost turista. Glede na številna področja športa je možna podrobnejša delitev na posamezne podzvrsti.
24.	Urbani turizem	Turizem osredotočen na obisk velikih mest, evropskih ali svetovnih metropol. Pri tem se pojem navezuje tudi na aktivnosti v urbanih mestih (obiskovanje znamenitosti, galerij, klubov...).
25.	Igralniški turizem	S pojmom označujemo vse turistične aktivnosti gostov, katerih glavni namen je igranje na srečo v igralnih salonih, casinojih in podobnih obratih ter z njimi povezane spremljevalne aktivnosti.
26.	Kulinarični turizem	Kulinarični turizem zajema v prvi vrsti turistične dejavnosti, katerih glavni namen je vezan na kulturo prehrane in uživanja tradicionalnih lokalnih specialitet (jedi in pijač), ki so značilne za neko pokrajino.
27.	Turizem	Oznaka za besedilo, ki ga ni mogoče zvrstno določiti s turističnega vidika.

Vir: *Thesaurus on tourism and leisure activities*, World Trade Organisation, 2001; *Encyclopedia of Tourism*, Jafari, 2000; dopolnitev in pojasnitev zvrsti po predlogu raziskovalne skupine.

iKorpus in luščenje izrazja za Islovar

Špela Vintar[♦], Tomaž Erjavec[♠]

[♦]Univerza v Ljubljani, Filozofska fakulteta
Aškerčeva 2, 1000 Ljubljana
spela.vintar@guest.arnes.si
[♠]Odsek za tehnologije znanja, Institut Jožef Stefan
Jamova cesta 39, 1000 Ljubljana
tomaz.erjavec@ijs.si

Povzetek

Slovensko računalniško izrazje se zbira v dvojezičnem spletnem terminološkem slovarju Islovar. Kot podpora slovarju je bil že pred leti zasnovan tudi korpus računalniških besedil, a so bili v njem doslej le zborniki konferenc DSI, poleg tega pa korpus nikdar ni bil uporabljen za sistematično terminografsko analizo ali samodejno luščenje izrazja. Prispevek predstavlja nadgradnjo in označevanje novega iKorpusa, nato pa opisuje rezultate samodejnega luščenja računalniškega izrazja iz njega. Podana je primerjava med dosedanjim besediščem Islovarja in izluščenimi enotami, v zaključku pa nakažemo tudi smernice za nadaljnje delo.

iKorpus and terminology extraction for Islovar

The Slovene specialized vocabulary of Computer Science is represented in the bilingual online dictionary project Islovar. To support the editorial and terminographical efforts involved in the making of Islovar, a corpus of Computer Science texts has been compiled out of several years' DSI conference proceedings. Since the corpus contained only one text type, the scientific article, and from a single source, it was never used in a thorough and systematic term extraction experiment. Firstly, this paper describes an upgraded, enlarged and tagged version of the corpus, now called iKorpus, and secondly we present the results of automatic term extraction performed on this corpus. We give a comparison between the terms currently included in the Islovar dictionary and the extracted term candidates, and the paper concludes with a discussion of the results and future perspectives.

1. Uvod

Računalništvo je eno najhitreje razvijajočih se področij, na katerem se nove tehnologije in z njimi nova poimenovanja pojavljajo že kar vsakodnevno. Slovensko izrazje s tega področja beleži Islovar,[1] spletni terminološki slovar informatike, ki že od leta 2001 nastaja pod okriljem Slovenskega društva informatika (Puc in Turk 2007). Kmalu po prvih zametkih Islovarja se je pojavila tudi zamisel, da bi za podporo slovaropisnemu delu oblikovali tudi specializirani korpus računalništva in informatike. Ker Slovensko društvo informatika organizira tudi vsakoletno strokovno posvetovanje Dnevi slovenske informatike (DSI), je prva različica korpusa vsebovala besedila iz zbornika tega posvetovanja iz leta 2003, vsako leto pa se je korpus povečal še za en zbornik (Erjavec in Vintar 2004).

Korpus DSI je urednikom slovarja služil predvsem kot vir informacij o pogostosti izrazov, za številne temeljne izraze pa je bilo iz njega mogoče pridobiti podpomenke in terminološke kolokacije (Puc in Erjavec 2006). Ker pa je korpus DSI obsegal le besedila ene besedilne vrste, se pravi konferenčne prispevke, in so bila vsa besedila iz istega vira, je bila njegova uporabnost za slovaropisne namene omejena (prim. Gorjanc in Logar 2007). Pričujoči prispevek predstavlja dopolnjeni in precej povečani korpus, ki je tudi na novo oblikoskladenjsko označen in lematiziran ter javno dostopen preko posodobljenega spletnega iskalnika.

Obenem smo želeli preveriti, ali je novi iKorpus primeren za podporo terminografskemu delu v okviru iSlovarja. Najprej nas je zanimalo, ali je izrazje iSlovarja zastopano v iKorpusu, potem pa še, ali lahko z metodami samodejnega luščenja izrazja pridobivamo terminološke kandidate za dopolnitev iSlovarja. Rezultati eksperimenta kažejo nekatere presenetljive zaključke, denimo da pogostost v korpusu ni dober indikator terminološkosti, saj se številni izrazi iz iSlovarja v iKorpusu pojavijo le enkrat.

2. iKorpus

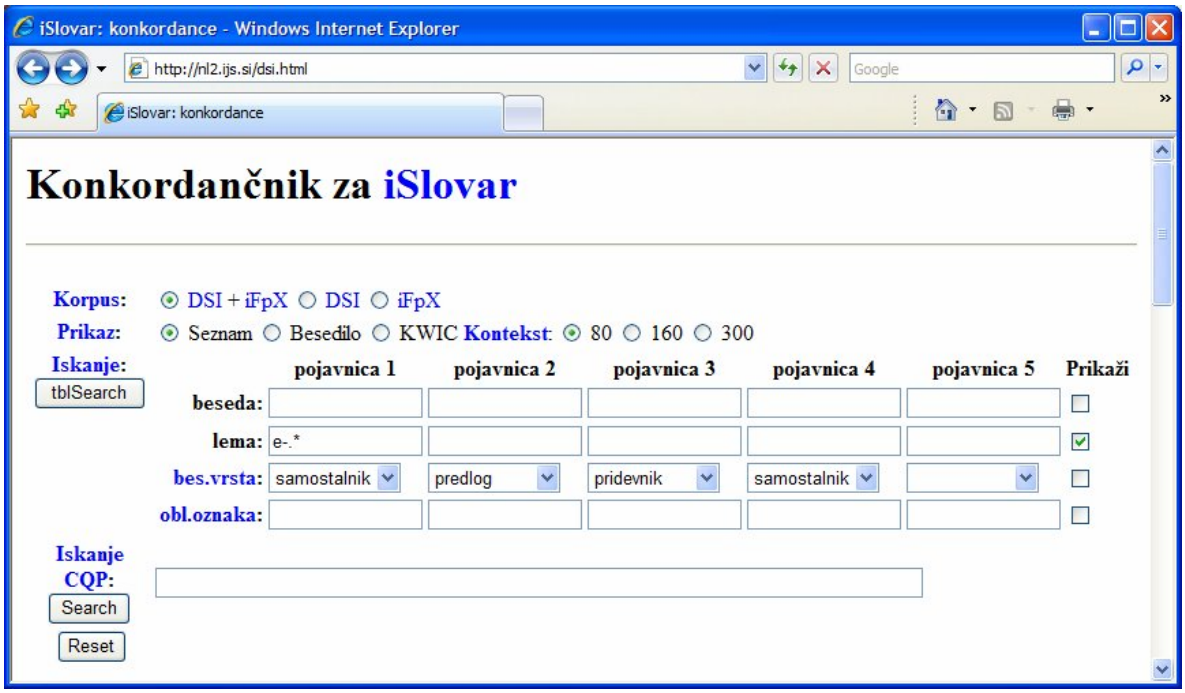
iKorpus je sestavljen iz dveh podkorpusov, od katerih eden (DSI) pokriva področje informatike, drugi, večji, (iFpX) pa računalništva.

Korpus DSI v različici (Puc in Erjavec 2006) zajema pet zbornikov posvetovanj Dnevi slovenske informatike iz let 2003 do 2007. Dopolnili smo ga z računalniškimi besedili iz korpusa FidaPLUS[2], in sicer iz revij Monitor, Moj mikro, Connect, PC Mediji, Joker in Računalniške novice. Velikost obeh podkorpusov, starega DSI,[3] dodanega podkorpusa FidaPLUS, poimenovan iFpX, in skupnega iKorpusa, kaže Tabela 1.

[1] http://www.islovar.org

[2] http://www.fidaplus.net

[3] Odkar je bila narejena pričujoča raziskava, je bil korpus DSI posodobljen še z zbornikom konference 2008, tako da sedaj vsebuje 721 besedil oz. 1.404.372 besed. Temu primerno se je povečal tudi iKorpus.



Slika 1. Vmesnik spletnega konkordančnika za iKorpus

	DSI	iFpX	iKorpus
Št. besedil	606	175	781
Št. odstavkov	18.630	463.571	482.201
Št. stavkov	60.110	994.183	1.054.293
Št. besed	1.192.580	12.678.086	13.870.666

Tabela 1: iKorpus v številkah

Tako stari DSI kot FidaPLUS sta sicer že bila avtomatsko označena z oblikoslovnimi oznakami ter lematizirana, vendar pa je bilo v oznakah korpusa DSI razmeroma veliko napak, korpus FidaPLUS pa nima pripisanih oznak neznanim besedam, torej tistim, ki niso zajete v leksikonu Amebisovega označevalnika, poleg tega pa oblikoslovne oznake obeh korpusov med seboj niso bile povsem skladne.

V okviru projekta JOS (Erjavec in Krek, 2008) je bil izdelan nov nabor oblikoslovnih oznak ter z njimi ročno označen korpus jos100k, ki vsebuje 100.000 besed, vzorčenih iz korpusa FidaPLUS. Ta korpus je služil za učno množico označevalniku in lematizatorju, s katerima sta bila na novo označena korpusa DSI in iFpX.

iKorpus tako vsebuje bolj kvaliteten nabor oznak, ki so dostopne tako v slovenskem kot v angleškem jeziku. Čeprav formalna primerjava med starim in novim označevalnim modelom še ni bila narejena, pregled posameznih primerov nakazuje, da je v novih oznakah tudi manj napak. Nekatere od preostalih napak gotovo vplivajo tudi na kvaliteto luščenih terminov, vendar pa večina napak zadeva bodisi pregibne lastnosti besed (npr. sklon), ali pa nekatere funkcijske besedne vrste, kar pa ne vpliva na kvaliteto luščenja.

Celoten iKorpus oz. njegova sestavna dela so javno dostopni za spletno iskanje[4] skozi posodobljeni konkordančnik, posebej prirejen za terminološko delo. Vmesnik (Slika 1) omogoča enostavno iskanje in prikaz po besednih oblikah, lemah, besednih vrstah ali celotnih oblikoslovnih oznakah. Iskalne kriterije lahko podamo kot regularne izraze in jih poljubno kombiniramo, izpis pa je lahko v obliki seznama zadetkov ali pa v formatu KWIC (keyword in context).

3. Islovar

Islovar je bil prvi spletni terminografski projekt, ki je že leta 2001 na način Wikipedije pričel zbirati izrazje in znanje s področja informatike in računalništva in se tako odzval na problematiko neažurnosti klasičnih (tiskanih) terminoloških priročnikov, še posebej na področjih naglega tehnološkega razvoja. Islovar zajema informacijsko izrazje, to je temeljno izrazje informatike, informacijske tehnologije in telekomunikacij, pa tudi izrazje posebnih področij, kot so baze podatkov, uporabniški vmesniki, poslovna informatika, objektna tehnologija, umetno zaznavanje in sociološki vidiki. Besede splošnega pomena Islovar vsebuje le, kadar imajo te specializiran pomen v računalništvu ali informatiki.

Islovar nastaja sproti, neposredno na spletu, vanj pa lahko prispeva vsakdo po predhodni prijavi. Vnešene izraze pregleduje in dopolnjuje uredniški

[4] http://nl2.ijs.si/dsi.html

odbor, zato so izrazi opremljeni z oznakami ureditve (predlog, pregledano, strokovno pregledano in urejeno). Urejeni sestavki vsebujejo razlage in kvalifikatorje. Urednica Islovarja je Katarina Puc, uredniški odbor pa sestavljajo še Vladimir Batagelj, Katja Benevol Gabrijelčič, Jurij Jaklič, Jože Kranjc, Niko Schlamberger in Tomaž Turk. Čeprav se področje informatike razvija tako hitro, da novemu izrazju tudi Islovar ni sposoben slediti, gre nedvomno za vzorni primer spletnega slovaropisja s soudeležbo široke strokovne javnosti (Turk in Puc 2007).

V pričujoči raziskavi smo primerjali besedišče Islovarja in iKorpusa. Za namene eksperimenta je bil uporabljen geslovnik Islovarja v takšnem obsegu in obliki, kot ga je imel marca 2008. Skupno število slovenskih gesel je bilo 4.742, Tabela 2 pa kaže dolžino gesel v besedah. Med daljšimi gesli so tudi poslovno-ekonomski izrazi, kot je *pogodba s povračilom stroškov in stimulativnim plačilom za delo,* a tudi opisni računalniški izrazi, npr. *koda za shranjevanje in prenašanje črkovnih in numeričnih znakov.*

Št. besed	Št. terminov
1	1.744
2	2.074
3	599
4	225
5	74
6	15
7	7
8	1
9	3

Tabela 2: Termini v Islovarju

Za potrebe primerjave gesel Islovarja in samodejno izluščenih terminoloških kandidatov iz iKorpusa smo geslovnik Islovarja lematizirali s spletnim lematizatorjem CLOG.[5]

4. Samodejno luščenje izrazja

Za samodejno luščenje izrazja je v literaturi predstavljenih več statističnih (Dunning 1993, Frantzi in Ananiadou 1999), skladenjskih (Heid 1999), semantičnih in hibridnih (Dias 2000, Vintar 2004) pristopov. Statistični sistemi se razvijajo predvsem za jezike in aplikacije, kjer predhodno oblikoskladenjsko označevanje korpusa ni mogoče, bodisi ker programi za samodejno razčlembo niso na voljo bodisi gre za preveliko količino besedil in/ali jezikov, kar velja za spletne aplikacije. Sistem, ki ga uporabljamo za pričujoči eksperiment, temelji na oblikoskladenjskih vzorcih za pridobivanje terminoloških kandidatov iz korpusa, nato pa se pridobljeni kandidati razvrščajo s pomočjo statistike terminološkosti. S tem izrabimo kar največ jezikoslovnih informacij o jeziku, ki jih ponujajo oblikoskladenjske oznake na eni strani in primerjava med specializiranim in referenčnim korpusom na drugi strani.

Osnovna predpostavka tega pristopa je, da so specializirane enote v besedilu večinoma večbesedne in tipično sestavljene po določenih oblikoskladenjskih pravilih, denimo pridevnik in samostalnik (operacijski sistem), samostalnik in samostalnik v rodilniku (kraja identitete), dva pridevnika in samostalnik (omrežno priklopljeni pomnilnik) itd. Ker je sestava terminoloških enot do neke mere lastna jeziku nasploh, v določenem segmentu pa je odvisna od posameznega področja, smo za eksperiment z iKorpusom tipične vzorce pridobili iz Islovarja. Uporabljeni so bili vzorci za enote dolžine od 2 do 4 besed, in sicer skupno 10 vzorcev, ki poleg zaporedja besednih vrst določajo tudi jedrno besedo.

Za boljše merjenje pogostosti enot luščenje poteka na lematiziranem korpusu, hkrati pa za vsako izluščeno lematizirano enoto poskušamo poiskati še njeno kanonično obliko (spleten stran → spletna stran). Če se izluščena lematizirana enota v korpusu niti enkrat ne pojavi z jedrno besedo v imenovalniku, kanonične oblike ne generiramo.

Izluščene enote se nato uredijo po terminološki relevantnosti. Za merjenje le-te je bilo predlaganih že precej metod, od katerih nekatere temeljijo na primerjavi relativnih pogostosti specializiranega in splošnega korpusa (Heid 1999), nekatere opazujejo stabilnost izluščene enote v korpusu in absolutno pogostost (Frantzi in Ananiadou 1999), spet drugi pristopi pa uporabljajo kontekstualne vektorje (Maynard in Ananiadou 2001). Naša metoda je prirejena in normalizirana primerjava relativnih pogostosti posameznih besed v izluščeni enoti, ki jo pomnožimo s kvadratom pogostosti cele enote v korpusu. Za primerjavo relativne pogostosti uporabljamo referenčni besedni seznam iz korpusa FidaPLUS.

Vzorec	Jedro
PRID+SAM	2
SAM+SAM	1
SAM+SAM+SAM	1
PRID+PRID+SAM	3
PRID+SAM+SAM	2
SAM+PRID+SAM	1
SAM+PRED+SAM	1
SAM+PRED+SAM+SAM	1
SAM+PRED+PRID+SAM	1
PRID+SAM+PRED+SAM	2

Tabela 3: Skladenjski vzorci za luščenje

Za izluščeno enoto a, ki vsebuje n besed, se terminološka utež izračuna po formuli

$$W(a) = \frac{\sum \log \frac{f_{n,D}}{N_D} \Big/ \frac{f_{n,R}}{N_R}}{n} * f_a^2$$

[5] http://nl2.ijs.si/analyze/

kjer sta N_D in N_R velikosti specializiranega korpusa D in referenčnega korpusa R v besedah. Sistem za luščenje terminologije tako razvrsti terminološke kandidate od najbolj do najmanj področno-specifičnih.

Kot je razvidno iz Tabele 2, je bilo iz korpusa izluščenih prek 300.000 samostalniških enot, od tega pa je bila le za dobro tretjino ugotovljena tudi kanonična oblika. Že prvi pregled kandidatov je pokazal, da je med njimi ogromno lastnih imen, in sicer tako osebnih imen kot imen proizvajalcev računalniške in fotografske opreme (Vladimir Djurdjič; Sony, Fujitsu, Canon, Microsoft). Poleg tega je računalniško izrazje pogosto akronimsko; med izluščenimi enotami jih kar 12.725 vsebuje akronim.

Tip enote	Št. izluščenih
Lastna imena	40.433
Akronimi	12.725
Vsebuje število	848
Kanonična oblika	125.528
Vsi	309.418

Tabela 4: Tipi izluščenih enot

Tabela 5 podaja prvih dvajset izluščenih enot v lematizirani in kanonični obliki.

	Lematizirana oblika	Kanonična oblika
1	igrovja pc	igrovje PC
2	operacijski sistem	operacijski sistem
3	programski oprema	programska oprema
4	plošča cd	plošča CD
5	enota cd	enota CD
6	vmesnik usb	vmesnik USB
7	spleten stran	spletna stran
8	zaslon lcd	zaslon LCD
9	vsebina po idje	
10	procesor pentium	procesor Pentium
11	digitalen fotoaparat	digitalni fotoaparat
12	revija joker	revija Joker
13	vladimir djurdjič	Vladimir Djurdjič
14	intelov procesor	Intelov procesor
15	računalnik pc	računalnik PC
16	nosilec cd	nosilec CD
17	jaka mele	Jaka Mele
18	uporabniški vmesnik	uporabniški vmesnik
19	sony ericsson	Sony Ericsson
20	pogon cd	pogon CD

Tabela 5: Prvih 20 izluščenih enot

Kot je razvidno iz Tabele 4, izluščene enote vsebujejo tudi osebna imena, ki razkrivajo sestavo korpusa. Vladimir Djurdjič je strokovni urednik Monitorja, Jaka Mele pa sodelavec revije Moj mikro, zato se njuni imeni znajdeta med prvimi dvajsetimi terminološkimi kandidati. Čeprav lahko pri aplikacijah za luščenje terminologije vse imenske sestavine navadno samodejno izločimo, je na področju računalništva opaziti precej terminoloških enot, ki vsebujejo ime. Islovar denimo vsebuje izraze *Bayerjevo drevo, datoteka Shape, Arnes, Bluetooth* itd., zato imen tudi iz seznama izluščenih kandidatov ne moremo neselektivno izločiti.

S primerjavo lematiziranih oblik izluščenih enot in lematiziranega geslovnika Islovarja ugotovimo, da je 1.154 slovarskih gesel tudi med izluščenimi kandidati. Ker sistem lušči le 2-4-besedne enote, to pomeni, da smo od skupno 2.898 2-4-besednih Islovarskih terminov iz korpusa izluščili okroglih 40 %. Če k temu dodamo še enobesedni inventar Islovarja, je od 1.744 enobesednih izrazov v Islovarju 1.489 prisotnih tudi v iKorpusu.

Sistem za luščenje je v svojih izhodiščnih nastavitvah usmerjen k čim večjemu priklicu (tj. čim manj spregledanih terminov) ob razmeroma nizki natančnosti. Ker so izluščeni terminološki kandidati urejeni po terminološki uteži, sestavljeni iz ključnosti posameznih besed in pogostosti cele enote, nas je zanimalo, kako se priklic spreminja z zviševanjem mejne terminološke uteži, ali z drugimi besedami, ali so "pravilne" terminološke enote iz Islovarja zgoščene v prvem delu seznama izluščenih enot. Prvi stolpec Tabela 6 navaja število izluščenih enot, na katerih smo računali priklic.

Št. enot	Priklic
5.000	0,09
10.000	0,12
50.000	0,20
100.000	0,26
150.000	0,30
200.000	0,32
250.000	0,35
309.418	0,40

Tabela 6: Priklic v odvisnosti od števila enot

Najopaznejši skok v priklicu se zgodi med 10.000 in 50.000 enotami, vseeno pa zgornji podatki kažejo, da se terminološke enote iz Islovarja pojavljajo tudi v spodnjem delu seznama izluščenih kandidatov. Iz tega sklepamo, da pogostost ni posebno dober kriterij terminološkosti, vsaj v našem primeru primerjanja izluščenih enot s slovarskimi gesli.

5. Razprava

Za evalvacijo sistemov luščenja terminologije ni enotne metodologije, nasprotno, različni poskusi poenotenja celo kažejo, da različnih sistemov praktično ni mogoče ocenjevati z istimi merili (Vivaldi in Rodriguez 2007). Največkrat se za ocenjevanje uspešnosti določenega sistema, tako kot za druge jezikovne tehnologije, uporabljata natančnost in priklic, pri čemer natančnost oceni

strokovnjak z ročnim pregledom izluščenih kandidatov, priklic pa je bistveno težje meriti, saj nihče vnaprej ne ve, koliko terminov v resnici vsebuje določeni korpus. Poleg tega je pojmovanje terminološkosti na eni strani izrazito subjektivno, saj se strokovnjak odloča drugače kot terminolog, na drugi pa odvisno od ciljne aplikacije, saj sistemi za iskanje podatkov izhajajo iz drugačne definicije termina kot terminografsko usmerjeni sistemi. Prav gotovo se načela, ki vodijo terminografa pri odločanju o terminološkosti določene besedne zveze in sistem za luščenje pri uvrščanju besednega niza v seznam kandidatov, razlikujejo že v temeljnih predpostavkah, prav zato so samodejno izluščeni spiski le redko neposredno uporabni za vključitev v aplikacijo.

Evalvacija izluščenih terminov s pomočjo referenčnega seznama izrazov področja, denimo geslovnika terminološkega slovarja, se uporablja redkeje; El Hadi in soavtorji (2006) pri enem takih eksperimentov poročajo o natančnosti okrog 4 %.

40-odstotno pokrivanje med geslovnikom Islovarja in samodejno izluščenimi termini iz iKorpusa lahko sproža različna vprašanja, med drugim tudi o terminološkosti preostalih 60 % izrazov v Islovarju, ki jih stroka očitno ne uporablja prav pogosto. Ker pa Islovarja - tudi po mnenju njegovih urednikov - kljub obsežnosti ne moremo jemati za zaključeno in popolno zbirko računalniško-informatičnih izrazov, je opisani eksperiment le korak v smeri korpusno podprte terminografije na tem področju. Ročna evalvacija izluščenih enot tako ostaja na seznamu načrtov za prihodnost.

6. Zaključek

V prispevku smo opisali nov jezikovni vir za področje računalništva in informatike iKorpus, ki je dostopen za spletno iskanje in bo v pomoč vsem, ki se na tak ali drugačen način ukvarjajo z računalniškim izrazjem. Čeprav korpus zdaj poleg konferenčnih zbornikov zajema tudi precej člankov iz različnih strokovnih revij, zanj še vedno ne moremo trditi, da je reprezentativen. Za celovitejšo predstavitev področja računalništva mu manjkajo še druge besedilne vrste, denimo srednješolski in univerzitetni učbeniki, diplomske, magistrske in doktorske naloge, priročniška besedila itd.

Eksperiment luščenja izrazja je pokazal, da iKorpus vsebuje večino besedišča Islovarja, da pa ga je s sedanjimi metodami mogoče samodejno izluščiti le določen segment. Ker pa Islovar še ni zaključena slovarska zbirka - kar glede na naravo področja najbrž tudi nikdar ne bo - pa lahko izluščene spiske uporabljamo kot vodilo pri vključevanju novih enot v slovar.

Literatura

Dias, G., Guilloré, S., Bassano, J.C., Lopes, J.G.P. (2000). Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association? In Proceedings of 6ème Conférence sur la Recherche d'Informations Assistée par Ordinateur (RIAO 2000). Paris, France, April 12-14. pp. 1-20.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics,*19:61-74.

El Hadi, W.M., Timimi, I., Dabbadie, M., Choukri, K., Hamon, O., Chiao, Y. (2006) Terminological resources acquisition tools: Towards a user-oriented evaluation model. Proceedings of the LREC2006 Fifth International Conference on Language Resources and Evaluation. 945-948.

Erjavec, T., Krek, S. (2008) The JOS morphosyntactically tagged corpus of Slovene. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'08. ELDA, Paris.

Erjavec, T., Vintar, Š. (2004) Korpus kot podpora slovarju informacijskega izrazja slovenskega jezika. Uporabna informatika. (Ljubljana), 12/2, 97-106.

Frantzi, K.T. in Ananiadou, S. (1999). The C-Value/NCValue domain independent method for multi-word term extraction. *Journal of Natural Language Processing,* 6(3): 145-179.

Gorjanc, V. in Logar, N. (2007). Od splošnih do specializiranih korpusov – načela gradnje glede na njihov namen. Orel, I: (ur.) Razvoj slovenskega strokovnega jezika. Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, 637-650.

Heid, U. (1999) Extracting Terminologically Relevant Collocations from German Technical Texts. V: Sandrini, Peter, ur.: Terminology and Knowledge Engineering (TKE99), Innsbruck. Dunaj: TermNet, 241-255.

Maynard, D., Ananiadou, S. (2001) Term Extraction using a Similarity-based Approach. V: Recent Advances in Computational Terminology, John Benjamins, 2001.

Puc, K. in Erjavec, T. (2006) Uporaba korpusa pri urejanju spletnega terminološkega slovarja. Zbornik mednarodne konference Language Technologies / Jezikovne tehnologije, Ljubljana: Institut Jožef Stefan, 156-161.

Turk, T. in Puc, K. (2007) Islovar kot model spletnega terminološkega slovarja. V: Orel, I. (ur.) Razvoj slovenskega strokovnega jezika. Ljubljana: Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik, 651-663.

Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. Memura 2004 - Methodologies and Evaluation of Multiword Units in Real-World Applications (LREC 2004): 54-57.

Vivaldi, J. in Rodriguez, H. (2007) Evaluation of terms and term extraction systems. Terminology 13:2 (2007), 225-248.

AvID: Audio–Video Emotional Database

Rok Gajšek[†], Anja Podlesek[*], Luka Komidar[*], Gregor Sočan[*], Boštjan Bajec[*], Vitomir Štruc[†], Valentin Bucik[*], France Mihelič[†]

[*]Psychological Methodology
Faculty of Arts
University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
{anja.podlesek, luka.komidar, gregor.socan, bostjan.bajec, valentin.bucik}@ff.uni-lj.si

[†]Laboratory for Artificial Perception, Systems and Cybernetics
Faculty of Electrical Engineering
University of Ljubljana
Tržaška 25, SI-1000 Ljubljana
{rok.gajsek, vitomir.struc, france.mihelic}@fe.uni-lj.si

Abstract

Initial attempts in design, recording strategies and collection of a multi-modal emotional speech database are presented. Our goal is to obtain a database to enable both the experiments in speaker identification/verification and detection of the emotional state of persons involved in communication. Especially we pay attention to gather data involving spontaneous emotions and therefore obtain more realistic training and testing conditions for experiments. Spontaneous emotions were induced with specially planned scenarios including playing computer games and adaptive intelligence tests. At the moment multi-modal speech from speakers was recorded and basic evaluations on data were processed.

1. Introduction

This study paper describes initial attempts to collect a multimodal emotional speech database as a part of our research under the ongoing interdisciplinary project *"AvID: Audiovisual speaker identification and emotion detection for secure communications"*. The goal of the project is to use speech and image technologies in video telecommunication systems for identification/verification and detection of the emotional state of persons involved in communication. Such a system should provide additional information about the identity and psychophysical condition displayed on the communication devices enabling more secure and credible exchange of information. Project partners are from the Faculty of Electrical Engineering – Department for Automatics from the University of Ljubljana and Jožef Stefan Institute – Department for Intelligent Systems from Ljubljana, the Faculty of Arts – Department for Psychology from the University of Ljubljana and the industrial R&D company Alpineon D.O.O. from Ljubljana.

Unfortunately most of the available speech databases with emotional speech were obtained by recording of acted different type of emotions usually from professional actors (LDC, 1999; LDC, 2002; Hozjan et al., 2001; Battocchi and Pianesi, 2004; Burkhardt et al., 2005; Martin et al., 2006a)[1] and therefore do not represent very adequately data for training and testing procedures for speaker psychophysical condition detection in real environment. To distinguish between normal and non-normal condition, and to compare speaker verification performances we also need quite a lot of speech material with normal speech, that is also usually not the case for available databases. Although we plan to perform our experiments on as much available data as possible, we also intend to obtain some data providing more realistic conditions for our task. Therefore we are planing to collect reasonable amount of audio and video recordings of spontaneous speech in normal (relaxed) and non-normal psychophysical conditions (the conditions of excitement and arousal with different valence, both positive and negative) from a representative group of speakers. Initial strategies to obtain desired speech corpora and recording setup along with the statistics of already recorded data are described in the following sections.

[1]There are of course some exceptions, as, for example, the German SmartKom database (Turk, 2001) which was recorded using a Wizard of Oz technique and tried to evoke different emotions in the participants.

2. Recording strategies

In the beginning, each participant was told that the main purpose of the experiments was to examine whether different measures of his/her state could be used in an adaptive test of intelligence. The biometric measures to be indicative of his/her psychophysical state at different moments were: psychophysiological response (the electrodermal and electrocardiographic response), verbal response, and facial expression. After a written consent to participate in the study was obtained from each individual, sensors were placed on the index and the middle finger on the left hand and the audio and visual recording started. The participant was instructed to speak loudly enough and not to move. With the right hand he/she had to hold a computer mouse in order to prevent the hand from excessive movement.

Subject	Sex	Age	Voice type	Health	Smoking	Overall mood	Dialect	Speech peculiarities
01	M	20	baritone	normal	NO	slightly tense	Central	
02	F	37	mezzo-sop.	cold	casual	relaxed	Eastern	
03	F	19	mezzo-sop.	normal	NO	relaxed	Littoral	
04	F	25	mezzo-sop.	normal	NO	relaxed	Eastern	
05	F	21	mezzo-sop.	normal	NO	relaxed	Central	
06	F	26	mezzo-sop.	normal	YES	NA	Central	
07	M	21	bass	normal	NO	relaxed	Central	rash speech
08	F	19	soprano	normal	stopped	distracted	Eastern	
09	F	20	mezzo-sop.	normal	NO	distracted	Littoral	
10	M	28	tenor	normal	NO	relaxed	Central	
11	F	26	mezzo-sop.	normal	NO	relaxed	Eastern	
12	F	20	mezzo-sop.	normal	NO	relaxed	Lower Car.	
13	F	27	mezzo-sop.	normal	NO	relaxed	Eastern	
14	F	20	mezzo-sop.	normal	YES	relaxed	Eastern	
15	F	27	soprano	normal	NO	relaxed	Central	

Table 1: Basic participants' data relevant for recorded speech analysis.

To obtain recordings of speech in both neutral and changed psychophysiological state of the participant, we designed an experiment composed of four parts. In Part I, after the participant introduced himself/herself with a few words (stated the name, the place of living, age, and main occupations), photographs with neutral content were presented on the screen. The participant was instructed to describe each photograph in detail, as if he/she were describing what he/she sees to a blind person. In this part we supposedly measured his/her verbal fluency. When he/she finished with the descriptions, he/she instructed the experimenter to continue with the presentation of the next photograph.

Before the start of Part II, the participant was told that we will be assessing the efficiency of his/her verbal instructions given to a teammate in order to achieve a common and specific goal. We explained to the participant that the team, which will involve himself/herself and the experimenter, will play a computer game (Tetris) and that he/she will observe the progression of the game on the computer monitor and will be giving verbal instructions, whereas the experimenter will not be able to observe the game and will carry out his/her orders by pressing the appropriate buttons on the keyboard. If the participant had no prior experience with the game, we explained the rules of Tetris and let him/her play for a few minutes. The participant who observed the ongoing game on the screen had to lead the experimenter through the game by uttering the following four commands: Left ('Levo' in Slovene), Right ('Desno'), Around ('Okrog'), and Down ('Dol'). The goal of the team was to achieve the highest score possible. Passive commands (e.g. 'Around' instead of 'Turn it around') were chosen in order to be suitable for use also in Part III of the experiment. At the end of the game the participant had to tell the experimenter what score they had achieved, what happened during the game, and why the game ended.

The ongoing game was recorded by CamStudio screen capture program. In Part III the recorded movie of the game was played and the participant had to describe what was happening on the screen by using the same four commands as in Part II. At the end the same description of the events on the screen had to be given as at the end of Part II. The aim of Part II was to obtain positive arousal (joy, satisfaction) as well as negative arousal (frustration, anger), whereas Part III was carried out to obtain exactly the same utterances in a relaxed, non-aroused state, because in Part III the participant was just a passive observer.

At the beginning of Part IV, the participant was told that he/she will be given an adaptive intelligence test where the difficulty of the task will be chosen by the computer according to (i) the correctness of the answer in the previous task, (ii) the mental strategy used for solving the task, and (iii) the biometric measures (EDR and heart rate). We explained to her that several values will be presented on the left part of the screen: the momentary IQ value, the arrows pointing upwards when the IQ estimate was increasing and downwards when it was decreasing, the momentary values of EDR and EKG measures, and the time that remained for solving the current task. On the right part of the screen, matrices with different figures or symbols were presented with one element absent. The participant had to reason aloud about the principles of the arrangement of matrix elements in rows and columns and find the proper solution among five to six possible answers. The participants believed they had to reason aloud so that the experimenter will be able to assess their mental strategy used for solving the task. After the experimenter showed two examples of matrices and explained how the reasoning should be verbalized, 20 matrices had to be solved, some of which were very difficult or did not have a known solution. If the participant ceased to speak aloud, the experimenter encouraged her to verbalize her thoughts. After the solution was found, the experimenter clicked some buttons to input the chosen solution and the presumed category of mental strategy. He could choose among six options with which he controlled the changes in the unfounded IQ estimate. He raised the IQ value only when the correctness of the solution was obvious, the reasoning was straightforward and solution was

derived quickly. In other cases the IQ value was decreased. The main purpose of decreasing the IQ value was to increase the participant's subjective stress level. Besides decreasing the temporary IQ value, the experimenter could also manipulate participants' stress level by increasing the EDR and heart rate values. In order to attract attention to the EDR and heart rate indicators during the test, the values changed its colour from black to red when a certain value was exceeded.

After Part IV was over we debriefed the participant. The experimenter explained that the temporary and final IQ scores were not valid estimates of her intelligence and that the real aim of the study was to obtain the recordings of speech in the normal, relaxed state and in the aroused, stress-induced emotional state. The participants were then asked to describe (freely) their feelings, thoughts, and involvement in each part of the experiment. In the end, some general data on participants and their speech characteristics were gathered (see Table 1).

3. Recording conditions and inventory

Recordings were done in a closed room using a digital video camera and three microphones (Mihelič et al., 2003). We used several microphones as in our previous GOPOLIS database (Mihelič et al., 2003) to enable some environment and channel normalization tests. For the reference also some physiological sensors were used to detect changes in the heart beet rate and skin conductivity during the tests on few test objects.

Each participant was asked to position himself/herself in front of the computer monitor shown in the upper right corner of Fig. 1. Behind the monitor a digital camera mounted on a tripod was placed to capture the video recording (i.e., video as well as one channel of audio data). To ensure an appropriate quality of the captured video data the participant was seated in front of a relatively homogeneous, white background and a light source was directed towards the participant's face. This setup resulted in the recorded video sequences showing a fairly "clean", i.e., without to much shadows, frontal view of the participant's face - see Fig. 2 where the recording setup is presented.

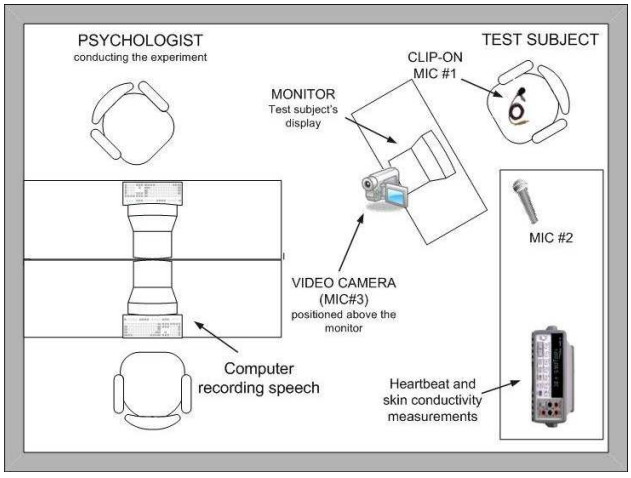


Figure 1: An outline of the recording setup.

Note that the quality of the recorded video data could have been further improved by using additional light sources or diffused light, however, as our goal was to collect a realistic database the employed setup fulfilled our requirements.

For capturing the audio signal two microphones were used in addition to the one integrated in the digital camera. The first, denoted as MIC #1 in Fig. 1, was attached on the participant's clothing near the chest, while the second, denoted as MIC #2 in Fig. 1, was positioned on the nearby desk. Both microphones were hooked to a computer which was used for recording and storing of the audio data.

Two people were supervising the acquisition of the database: (i) a psychologist who was in charge of the recording session and tried to induce a "non-normal" psychophysical condition in the participant by applying the strategies presented in Section 2. and (ii) a technician who overlooked the technical aspects of the acquisition process.

Figure 2: The recording setup.

3.1. Video data collection

The video part of the AvID emotion database was acquired using a high-definition Sony HDR-SR11E digital handycam which captures video at a resolution of 1920×1080 pixels and a bit-rate of 16Mb/s. The video data was recorded at a frame aspect ratio of $16 : 9$ and later on archived in the AVCHD (Advanced Video Codec High Definition) format[2]. Specifications of the employed format for the video data can be found in Table 2.

Video signal	$1080/50i$
Pixels	1920×1080
Aspect ratio	$16 : 9$
Compression	MPEG4 AVC/H.264
Luminance sampling frequency	74.25 MHz
Chroma sampling format	4:2:0
Quantization	8-bit

Table 2: Specifications of the employed AVCHD format.

A high-definition camera was chosen for capturing the video data of the database for several reasons: (i) different

[2]A high-definition format jointly established by Panasonic and the Sony Corporation.

kinds of experiments (e.g., biometric verification or identification, emotional state recognition, lip reading, etc.) can be performed on high quality video, (ii) with simple image- and video-processing techniques the quality of the video data can easily be degraded and research can be conducted on lower-quality video, and (iii) as high-definition technology is spreading with an increasing speed, it will soon find its way into peoples daily lives; with its widespread deployment the technology will also become easily affordable and, therefore, suitable for employment in low- (or medium-) cost recognition systems. A sample frame captured with Sony's HD camera in the AVCHD format is shown in Fig. 3.

Figure 3: A sample frame from the AvID audio–video emotional database.

3.2. Audio data collection

As mentioned above the audio signal was captured using three different microphones. Channel number one was recorded using a Sennheiser ew122-p G2 system with a clip-on microphone which transmitted the signal to the recording computer via radio waves. The microphone was pinned to the speakers chest roughly $10-20$cm away from the speaker's mouth.

The second channel was captured with a Shure PG81 microphone which was positioned approximately $30-40$ cm away from the speaker. Both microphones specify a frequency range of $40-18000$ kHz. Channels were recorded at a sampling rate of 16 kHz and 16-bit linear encoding.

The third channel was acquired from the employed video camera's built in microphones that record in Dolby Digital 5.1 and use AC-3 compression for audio storage.

4. Database description

A total of 15 native Slovenian speakers (12 female and 3 male) were recorded with one session lasting approximately an hour. After extracting only the speaker's speech from the session we got roughly a half an hour of usable audio per speaker. For the video part of the database one continuous recording was captured for each of the participants resulting in over 15 hours of high-definition video. As already mentioned in the previous section the participants were recorded in front of a white background and with frontal illumination. The average inter-ocular distance, which is the traditional measure of the size of the face in an image or video frame, is more than 150 pixels. Similar databases (uni- or multi-modal) used either for assessing biometric identification/verification or emotion recognition algortihms, such as the XM2VTS (Messer et al., 1999), the Cohn-Kanade (Kanade et al., 2000) and the eNTERFACE'05 (Martin et al., 2006b) databases, typically feature face images (or video sequences) with a distance of $40-60$ pixels between the left and the right eye. The AvID database is therefore suitable as the foundation for the development of recognition algorithms that make use of high resolution information.

The audio recordings were later split to shorter utterances - approximately one utterance per sentence. Transcriptions and labels describing emotional state of the speaker were made using Transcriber tool (Barras et al., 2001) and followed the LDC broadcast speech transcription conventions[3].

5. Evaluation results

Subjective reports were analyzed - descriptions of the states at the beginning of the experiment and within each part of the experiment were classified into five categories (see Figure 1) where possible. Where the category is composed of two arousal levels (e.g., moderately and highly aroused), the first one reflects the prevalent state and the second reflects temporary peaks of slightly elevated stress level. Mostly the participants reported of the relaxed or slightly tense state prior to the experiment (when the sensors were attached and the procedure was explained to them). In Part I, when describing photographs, most of them reported no tension, and some reported a slight tension that later vanished. Their arousal was increased slightly while playing Tetris. Some participants reported negative emotions (e.g., irritation) in Part III due to the inability to take over the control. The majority reported that Part IV was difficult and stressful because they had troubles with verbalizing their reasoning and were worried and puzzled about the calculated IQ value. This was reflected in a notable decrease of speech loudness. The change towards higher arousal during the experiment is indicated with the prevalence of darker pattern in Figure 4, whereas the transparent patterns represent a more relaxed state. It may be concluded that the experimental situations elicited the presumed levels of arousal: neutral emotional state with describing photographs and events, and arousal in playing an exciting computer game and in the situation where an individual wants to perform well under social pressure.

5.1. Future prospects

In future studies, a triangulation of different methods will be used to assess the arousal and the emotional state of the participants more systematically and objectively. To obtain additional indicators of the participants' emotional state we will use: (i) standardized instruments of subjective emotional experience, (ii) the psychophysiological measures, such as EDR and EKG, and (iii) behavioural expressions.

[3]LDC broadcast speech transcription conventions: http://projects.ldc.upenn.edu/Corpus_Cookbook/transcription/index.html

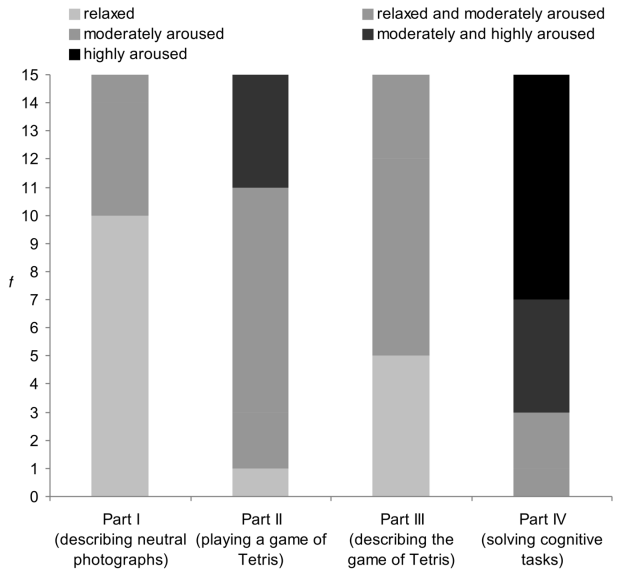


Figure 4: Levels of arousal during different parts of the experiment based on the participants' subjective report.

6. Conclusion

The goal of recording a multi-modal speech database containing different spontaneous emotions was achieved. Due to well selected experiments different levels of arousal were induced and measured by different biometric parameters: facial expression - video, verbal response - audio and psychophysical response - electrodermal and electrocardiographic response. Video and audio comprise the database, where psychophysical measures are only used to provide an objective information about the level of arousal.

Enough data was collected (which is especially important for speech research) to form a bases for future studies on speaker identification/verification, emotion recognition and spontaneous speech analysis research.

7. Acknowledgement

This work was supported by the Slovenian Research Agency (ARRS), development project M2-0210 (C) entitled "AvID: Audiovisual speaker identification and emotion detection for secure communications."

8. References

Claude Barras, Edouard Geoffrois, Zhibiao Wu, and Mark Liberman. 2001. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2):5–22.

A. Battocchi and F. Pianesi. 2004. Dafex: Un database di espressioni facciali dinamiche. *Proceedings of the SLI-GSCP Workshop "Comunicazione Parlata e Manifestazione delle Emozioni"*, pages 1–11.

F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. 2005. A database of German emotional speech. *Proceedings Interspeech 2005*, pages 1–4.

V. Hozjan, Z. Kačič, and B. Horvat. 2001. Prosody feature analysis for emotion modeling. *Electrotechnical Review*, 68:213–218.

T. Kanade, J.F. Cohn, and Y. Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings of the 4th AFGR'00*, pages 46–53, Grenoble, France.

LDC, 1999. *SUSAS (Speech Under Simulated and Actual Stress)*. Language Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S78.

LDC, 2002. *Emotional Prosody Speech and Transcripts*. Language Data Consortium, University of Pennsylvania. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006a. The enterface05 audio-visual emotion database. *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 1–8.

O. Martin, I. Kotsia, B. Macq, and I. Pitas. 2006b. The enterface'05 audio-visual emotion database. In *Proceedings of the 22nd International Conference on Data Engenireeng Workshops*, Washington D.C., USA. IEEE Computer Society.

K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. 1999. Xm2vtsdb: the extended m2vts database. In *Proceedings of AVBPA'99*, pages 72–77, Washington D.C., USA, March.

F. Mihelič, J. Žganec Gros, S. Dobrišek, J. Žibert, and N. Pavešić. 2003. Spoken language resources at LUKS of the University of Ljubljana. *Int. J. Speech Technology*, 6:221–232.

U. Turk. 2001. The technical processing in smartkom data collection: a case study. In *Proceedings of Eurospeech*, pages 1541–1544, Sweden.

Using the Web as a Corpus for Extracting Abbreviations in the Serbian Language

Vesna Satev [1], Nicolas Nikolov [2]

[1] Department for Informatics and Computer Science
Faculty of Mathematics, Belgrade University
Studentski trg 18, 11000 Beograd, Republic of Serbia
vesnasatev@ptt.rs

[2] Umbria Inc.
1655 Walnut St, Suite 300
Boulder, CO 80302, U.S.A.
nicolas@umbrialistens.com

Abstract

In this paper we discuss the results of extracting abbreviations in the Serbian language by using the web as a corpus. The results are compared to those retrieved by using the standard corpus of contemporary Serbian language. Using the web as a corpus is a very recent trend. It is a valuable source of data for research in computational linguistics and information extraction. Still, there are no adequate tools for searching the web, which are geared to linguistic needs. We chose crawling as a process for collecting data from the web, in order to extract abbreviations in the Serbian language. We show that, by using the web as a corpus, a higher number of abbreviations can be found and they are more recent.

Introduction

In the last decade, methods that are being used in computational linguistics research have drastically changed. Fifteen years ago, most research was focused on selected examples of sentences. Today, it is common to use huge corpora of text for research purposes.

One of the greatest advantages of corpus-based research is that corpora provide a way to discover facts about the language which are not observable or quantifiable by manual means. However, corpus-based research has some disadvantages:

- the design and creation of text corpora can be expensive;
- corpora are fixed at a point in time;
- corpora do not provide access to up-to-date information on language use or the changes which are occurring.

In an attempt to overcome these disadvantages, a lot of language scientists and technologists are increasingly turning to the web as a source of language data.

The web contains a lot of text resources, more than any known corpus. It is the only available source for some languages. It is free and instantly available.

In this paper, we will discuss these advantages, as well as when is the use of the web as a corpus a better alternative. We will show on the sample task of extracting abbreviations in the Serbian language that using the web as a data resource leads to better results than using a relatively large corpus.

The World Wide Web

The World Wide Web is a marvelous place, with a vast range of languages, content domains and media formats. The Web is constantly changing and growing, and even the best estimates can only approximate its extent and composition. The most reliable estimates suggest that the number of publicly indexable web pages in mid-2005 is in the range of 10 to 20 billion.

This is only the visible part of web. Far larger is the "invisible" Web, the information that is on the web, but cannot be accessed through search engines – these are documents or data that reside in databases and can be explored only by entering relevant queries in a form. There are a number of researchers who are trying to improve search engines and to pull out such data (a "deep web" concept), but this concept is still at the initial phase of development.

In addition to the web being a huge collection of text data, linguists have one more benefit from web which lies in its multilingualism. Data from Global Reach (2004) show the following percentages of users for the top ten languages: English 35.2%, Chinese 13.7%, Japanese 8.4%, Spanish 9.0%, German 6.9%, French 4.2%, Korean 3.9%, Italian 3.8%, Portuguese 3.1%, Dutch 1.7%, other 10.1%.

Another advantage of the web as a corpus is that, for some languages, the web is the only source of data.

Why Use the Web as a Corpus?

Different kinds of data are needed for different linguistic purposes. For many research questions, data from a standard corpus like the British National Corpus or the Corpus of Contemporary Serbian Language are sufficient. But there are cases in which the data needed to answer or explore a question cannot be found in a standard corpus. It could be because the phenomenon under consideration is rare, belongs to a genre or register not represented in the corpus, or stems from a time that the corpus data do not cover (for example, it is too new). In these cases the Web is a good and convenient source of data, because of some of its properties:

- Freshness and spontaneity: the content of compiled corpora ages quickly, while texts on contemporary is-

sues and authentic examples of current, non-standard, or emerging language usage resides on the web;

- Completeness and scope: existing corpora may lack a text genre or content domain of interest, or else may not provide sufficient examples of an expression or construction easily located online;
- Linguistic diversity: languages and language varieties for which no corpora have been compiled are found online;
- Cost and convenience: the web is virtually free, and desktop computers to retrieve and process web pages are available to researchers and students alike.

Computational Linguistics Approach

Using the web as a corpus is a very recent trend. The number of approaches that are relevant to Computational Linguistics is still rather small. But already the web has been tried for tasks on various linguistic levels: lexicography, syntax, semantics and translation.

Lexicography: Lexicographers find WWW as an interesting and very rich source for discovering and classifying new lexical material from the wealth of texts in the web. This includes finding and classifying new words or expressions and gathering additional information such as typical collocations, subcategorization requirements, or definitions. Jacquemin and Bush (2000) describe an approach of learning and classifying proper names from the web. This is a worthwhile undertaking since proper names are an open word class with new names being continually invented. Later in this paper we will describe an approach for finding new abbreviations in the Serbian language and their definitions from the web.

Syntax: Martin Volk (2002) discussed how the vast textual resources of the web can be exploited to improve parsing. He used frequencies obtained from search engines to resolve PP (prepositional phrase) attachment ambiguities (see also Volk 2000 and Volk 2001). An English sentence consisting of the sequence verb + NP + PP is a priori ambiguous. The PP in example sentence 1 is a noun attribute and needs to be attached to the noun, but the PP in 2 is an adverbial and thus part of the verb phrase:

(1) *Peter reads a book about computers.*
(2) *Peter reads a book in the subway.*

If the subcategorization requirements of the verb or of the competing noun are known, the ambiguity can sometimes be resolved. But often there are no clear requirements. Therefore, there has been a growing interest in using statistical methods that reflect attachment tendencies. For example for sentence 1 the triple frequencies for (*read*, *about*, *computer*) and (*book*, *about*, *computer*) are needed as well as the unigram frequencies for *read* and *book*. Obviously it is very difficult to obtain reliable frequency counts for such triples. Therefore the largest corpus available, the WWW, is used. With the help of a WWW search engine, frequency values ('number of pages found') are obtained and used to compute co-occurrence values, based on the following formula (X can be either the verb V or the head noun N1):

cooc(X,P,N2) = freq(X,P,N2) / freq(X)

For example, if some noun N1 occurs 100 times in a corpus and this noun co-occurs with the PP (defined by P and N2) 20 times, then the co-occurrence value cooc(N1,P,N2) will be 20 / 100 = 0.2. The value cooc(V,P,N2) is computed in the same way. The PP attachment decision is then based on the higher co-occurrence value.

Semantics: Retrieving semantic information from the web is a very difficult task, but also the most significant one with regard to practical applications. Agirre et al. (2000) describe an approach to enriching the WordNet ontology using the WWW. The authors show that it is possible to automatically create lists of words that are topically related to a WordNet concept. If a word has multiple senses in WordNet, it will be accompanied by synonyms and other related words for every sense. Agirre et al. query the web for documents exemplifying every sense by using these co-words. The query is composed by using the disjunction of the word in question and its co-words and by the exclusion of all co-words of the competing senses (via the NOT operator). The documents thus retrieved are locally processed and searched for terms that appear more frequently than expected using the X2 function. The resulting topic signatures (lists of related words) are evaluated by successfully employing them in word sense disambiguation. Moreover, the authors use the topic signatures to cluster word senses. For example, they were able to determine that some WordNet senses of boy are closer to each other than others. Their method could be used to reduce WordNet senses to a desired grain size.

Translation: The web is a very useful tool for looking up how a certain word or phrase is used. Queries to standard search engines allow for restrictions to a particular language or to a particular country. Therefore, it has become easy to obtain usage information. Grefenstette (1999) has shown that WWW frequencies can be used to find the correct translation of German compounds if the possible translations of their parts are known.

Also, there are a lot of translations that exist and are published in the web. The task is to find these text pairs, judge their translation quality, download and align them, and store them into a translation memory. Furthermore, parallel texts can be used for statistical machine translation. Resnik (1999) developed a method to automatically find parallel texts in the web.

Diachronic change: The web may also be used to observe language change over time and to provide access to today's language. Volk (2002) showed on example of two recent words in Swiss German, *Natel* and *Handy*, which are competing for prominence in Switzerland, how the web can be used to determine the language changes. He proved his hypothesis that *Handy* has become more frequently used, by checking the frequency of occurrence on the web before and after January 1st, 2000. He used the Hotbot search engine since it allows this kind of time span search.

Later in this paper, we will show how the web can be used to find out new abbreviations in the Serbian language, and potentially their definitions. This is something that can not be done by exploring a standard corpus, because of the nature of that class of words. For example, for a long time in every newspapers article in Serbia it

was common to use the word Kosovo for the Serbian province called Kosovo i Metohija (Kosovo and Metohija). But very recently, this trend has changed and almost every journalist started to use abbreviation KiM instead. This word can not be found in a standard corpus of the Serbian language, and therefore the web as a corpus is the only solution.

Information extraction approach

Information Extraction (IE) is the next step up from search engines in fulfilling information processing needs. It is a technology that is futuristic from the user's point of view in the current information-driven world. Rather than indicating which documents need to be read by a user, IE systems analyze unrestricted text in order to extract information about pre-specified types of events, entities or relationships. These facts are then usually entered automatically into a database, which may then be used to analyze the data for trends, to give a natural language summary, or simply to serve for on-line access. Links between the extracted information and the original documents are maintained to allow the user to reference context.

IE is closely related to Computational Linguistics, because it uses a lot of CL and NLP techniques and methods. Question Answering, perhaps better than any other area, represents this tight connection between these areas.

Depending on what data needs to be extracted, researchers could choose a standard corpus or the web as a source of data. For many purposes, a standard corpus can be a good place to search for facts. But also, there are a lot of facts that could not be found in corpora, maybe because they are too new or because a standard corpus does not cover the topic. For example, information about upcoming conferences is something that could be found only on WWW. If a user queries a standard search engine with "*conference CL 2006*", s/he will get a list of sites that contain those words, and than will have to download these pages, read them one by one to find out what is the date or the place where the conference will be held. But, he doesn't need all the other information he will get by the way. IE systems provide a way to get just the information user needs (the name, the place and the date of the conference) and a possibility to maybe sort that results by date, something that standard search engines can't do.

One good example of an IE system is Froogle (http://froogle.google.com). It is a search engine which, as a result, has a list of products available online, with pictures, prices, descriptions and links to particular product.

Accessing and Collecting Data from the Web

Currently, data on the web can be accessed only through search engines. The problem is that search engines are not tuned to the needs of linguists. For example, it is not possible to query for documents that contain the English word *back* as a noun. Thus, linguists have to find out other techniques for their research based on web data.

In principle, there are several options for using data from the web.

Searching the whole Web through a commercial engine directly

Researcher can use a commercial engine, for example Google or AltaVista, directly. Although those search engines are not adapted to the needs of linguists, some research can be done just by using them. There are many examples in the literature where researchers used frequencies returned by search engine to determine some facts about language (see Volk 2002).

Using frequency data from a search engine ("Google frequencies") can be useful, but is also problematic. For one thing, all search engines perform some sort of normalization: searches are usually insensitive to capitalization ("*jasna*" and "*Jasna*" return the same number of matches; the first is an adjective in the Serbian language, the second is a proper name), automatically recognize variants ("*white-space*" finds *white space*, *white-space* and *whitespace*) and implement stemming for certain languages (as in *lawyer fees* vs. *laywer's fees* vs. *lawyers' fees*). While such features can be helpful when searching information on the web, they may also distort the frequency counts. It is possible to deactivate some, but not all of these normalizations. However, this requires a detailed knowledge of the query syntax, which may change whenever Google decides to update its software. Another serious problem is duplicate texts found by search engines; the same text can be present in several different web pages (citation of someone's speech, for example). Such duplication, which is much more common on the Web than in a carefully compiled corpus, may inflate frequency counts drastically. Manual checking could in principle be used to correct the frequency counts, both for normalization and for duplication, but it is prohibitively time-consuming (since the original Web pages have to be downloaded).

Adding pre- and/or post-processing to the search engine, to refine query results

Using the features of a search engine is a good idea, but not enough for the majority of linguistic researches. There are several examples of systems that pre-process queries before they are sent to search engines and post-process the results to make them more linguist-friendly. Probably the most famous pre-/post-processing systems are WebCorp (Kehoe & Renouf, 2002) and KWicFinder (Fletcher 2001).

WebCorp is a Web-based interface to search engines such as Google and AltaVista, where the user can specify a query using a syntax that is more powerful and linguistically oriented than the one of the search engines. For example, it is possible to use wildcards such as * meaning "any substring" (as in: "*ing"). Moreover, WebCorp organizes the results returned by the search engine in a clean "keyword in context" format, similar to that of standard concordance programs.

However, ultimately these tools are interfaces to Google and other search engines, and as such they are subject to all the query limitations that the engines impose, they cannot provide information that is not present in the data returned by the engines, and they are subject to constant brittleness, as the nature of the services provided by the engines may change at any time.

Collecting pages from the Web (randomly or controlled) and searching them locally

Except for very small corpora, the process of downloading web pages to build the corpus (and any post-processing that is applied) must be automated. In principle, this is the optimal approach to using the Web as a corpus, given that it provides full control over the data. However, crawling, post-processing, annotating and indexing a sizeable portion of the Web is by no means a trivial task. Even though the idea of building a linguist search engine has been around for at least 4 years, to this date the only projects that have produced concrete results involved (relatively) small-scale crawling. For example, Ghani et al. (2001) described how to build corpora of minority languages by sending queries to the search engine using words "typical" of specific languages. Baroni and Bernardini (2004) used a similar approach to create specialized language corpora for terminographical work. Baroni and Kilgarriff (2006) have developed very large corpora from the Web for German and Italian languages by crawling the web and post processing retrieved pages (*DeWac* and *ItWac*). For generating seed URL's, they used results retrieved by Google search engine which was queried with randomly chose pairs of words. The crawling process and post processing of downloaded pages will be described later in this paper.

Exploring Abbreviations in the Serbian Language

Using the web as a corpus turned out to be the best alternative for analyzing named entities. Proper names, for example, are an open word class with new names being continually invented. Standard corpora can not serve as a data source, since they don't contain newly added names. A similar class of words is abbreviations.

In our research we try to collect acronyms, as one type of abbreviations in the Serbian language, together with their possible definitions. The task was to find newly acronyms, Available corpus of contemporary Serbian language (SrpKor), developed at Faculty of Mathematics, University of Belgrade, is made of different resources, such as books and news articles and it consists of 23,532,367 tokens. Newspaper texts date from 1993, textbooks and monographs date from 1980, while literature part dates from 1920 and it consists of both original and translated works. The most recent text in SrpKor dates from year 2000. Therefore, in order to find acronyms that are recent, we have to use some other resource. The web was the only adequate, because of its freshness and availability. Nevertheless, we did our research on the SrpKor as well, in order to compare the results.

Abbreviations in the Serbian language are often written with all letters in upper case. For example, very common abbreviations in Serbian language are:

SCG – Srbija i Crna Gora (Serbia and Montenegro)

SFRJ – Socijalisticka federativna republika Jugoslavija (Socialistic federal republic of Yugoslavia)

BIA – Bezbednosno informativna agencija (Agency for state security and information)

SANU – Srpska akademija nauka i umetnosti (Serbian academy of science and art)

NBS – Narodna banka Srbije (National bank of Serbia)

But there are also cases like *BiH* (Bosnia and Hertzegovina) or *KiM* (Kosovo and Metohia). So, in order to find abbreviations, we searched for words not longer then *n* letters (5 in our case), where more than half of the letters are upper case. We also excluded from the results words that satisfy the above condition and whose context is all written in upper case. Considering there is no search engine or tool which supports this kind of queries, we automatically collected web pages by crawling the web in order to get data for our research.

Crawling the web

A crawler is a program that traverses the web by following hyperlinks from one page to another. It usually starts from a predefined list of seed URLs. The set of URLs used to initialize the crawl and various parameter settings of the crawler (e.g., the number of pages to be downloaded from each domain) has a strong impact on the nature of the corpus being built.

Broad crawl of the Web requires considerable memory and disk storage resources. So crawling the whole or large part of web for linguistic purposes maybe is not such a good idea. Instead, we carefully choose seed URLs for the crawling process, having in mind our task to pull out recent acronyms. For that purpose, we put as seeds the URLs of daily newspapers from Serbia.

Post processing

After the crawling process had finished and pages were downloaded, the next step was to eliminate HTML content – tags, scripts and other "non Serbian" text. Web corpus, attained in the described way, contained about 500,000 words. The remaining text was then tokenized. Words, together with their context and source, were put into the database.

Exploring abbreviations

Once we had words stored into the database, the search could start. We extracted words not longer than five letters, with more than half letters in upper case, where the other words in the context are not written in upper case.
We also searched the standard corpus of contemporary Serbian language with the same query and joined results. When this process was finished, we had around 600 different acronyms. The results are shown in Table 1.

Percentage of abbreviations that appeared in both corpora	9%
Percentage of abbreviations that appeared only on web	33%
Percentage of abbreviations that appeared only in standard corpus	57%

Table 1. Percentage of found abbreviations

Among the results, there were strings that do not represent abbreviations, like Roman numbers (*XIV*). Also, because of the five letters limit, some abbreviations are not recog-

nized (*UNESCO*). But we assumed that percentage of wrong words that are recognized as abbreviations or that are not recognized at all is the same in both standard corpus and web, so we compared the results.

Of the union of the abbreviations found through both sources (web and the corpus of contemporary Serbian language) 9% represent abbreviations was present in both sources. Among the acronyms found on the web 33% were not found in the SrpKor, although the size of text collected from web was 50 times smaller then SrpKor. Those were the acronyms we were looking for, acronyms that are new and recent, and therefore they can not be found in SrpKor. On the other hand, among all acronyms found in the corpus of contemporary Serbian language 57% were not found on the web. The reason for the later is that many of the abbreviations in the standard corpus represent organizations, political parties or countries that no longer exist (SSSR, for example). This is not surprising, having in mind that we used URLs of daily newspapers in Serbia as a seed for crawling. The articles that can be found there are mostly about recent events and they don't contain historical texts. By extending the seed URLs for crawling, the percentage of abbreviations that appear only on web, and not in standard corpus, can only grow. Therefore, this is a kind of research that can not be done without using the web as corpus.

Conclusion

In this paper we have discussed at length the advantages and disadvantages of using the web as a data resource for research in computational linguistics. We considered the task of identifying recent acronyms in the Serbian language using the corpus of contemporary Serbian language and the web as data sources. We were able to find more abbreviations and more recent ones on the web. We argue that further linguistic research would benefit greatly from automated tools that can leverage the wealth of information found on the web to perform similar tasks that we programmed in our system.

References

Agirre, E., Olatz, A., Hovy, E. and Martinez, D. 2000. *Enriching very large ontologies using the WWW. ECAI 2000*, Workshop on Ontology Learning. Berlin.

Baroni, Marco and Silvia Bernardini 2004. *BootCaT*: *Bootstrapping corpora and terms from the Web*. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004), Lisbon

Baroni, Marco and Kilgarriff, Adam 2006. *Large linguistically-processed Web corpora for multiple languages.* In: Proceedings of EACL 2006, Trento, Italy

Fletcher, William H. 2001. *Concordancing the Web with KWiCFinder*. In: Proceedings of the 3rd North American Symposium on Corpus Linguistics and Language Teaching, Boston.

Ghani, Rayid, Rosie Jones and Dunja Mladenic. 2001. *Mining the web to create minority language corpora.* In: Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM), 2001.

Grefenstette, Gregory 1999. *The World Wide Web as a resource for example-based machine translation tasks.* Proc. of Aslib Conference on Translating and the Computer 21. London.

Jacquemin, C. and Bush C. 2000. *Combining lexical and formatting clues for named entity acquisition from the web.* Proc. of Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora. Hongkong. 181-189.

Kehoe, Andrew and Antoinette Renouf 2002. *WebCorp: Applying the Web to linguistics and linguistics to the Web*. In: Proceedings of the WWW 2002 Conference. Honolulu

Resnik, Philip 1999. *Mining the web for bilingual text*. Proc. of 37th Meeting of the ACL.Maryland. 527-534.

Volk, Martin 2000. *Scaling up. Using the WWW to resolve PP attachment ambiguities*. Proc. of Konvens-2000. Sprachkommunikation. Ilmenau.

Volk, Martin 2001. *Exploiting the WWW as a corpus to resolve PP attachment ambiguities*. Proc. of Corpus Linguistics 2001. Lancaster.

Volk, Martin. 2002. *Using the web as corpus for linguistic research.* In: Renate Pajusalu and Tiit Hennoste (eds.) Tähendusepüüdja. Hatcher of the Meaning. A Festschrift For Professor Haldur Õim. Tartu: University of Tartu.

Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata

Damir Ćavar,[1] **Ivo-Pavao Jazbec,**[2] **Siniša Runjaić**[2]

[1]University of Zadar
Linguistics Department
Obala kralja Petra Krešimira IV. 2, 23000 Zadar, Croatia
dcavar@unizd.hr

[2]Institute of Croatian Language and Linguistics
Ul. Republike Austrije 16, 10000 Zagreb, Croatia
{ipjazbec,srunjaic}@ihjj.hr

Abstract

We discuss the design and development of a finite state transducer for morphological segmentation, annotation, and lemmatization that allows for merging of three major functionalities into one high-performance monolithic automaton. It is designed to be flexible, extensible, and applicable to any language that allows for purely morphotactic modeling on the lexical level of morphological structure. The annotation schema used in an initial Croatian language model is a direct mapping from the GOLD ontology of linguistic concepts and features, which increases the potential for interoperability, but also opens up advanced possibilities for a DL-based post-processing.

1. Introduction

Quantitative and qualitative information about morphological properties of languages is hard to come by. For many languages information as for example contained in CELEX (Burnage, 1990) is not available. For many research questions, most of the available information about distributional properties of morphemes and their feature makeup is not sufficient.

Corpus annotations tend to be lexeme and word-form oriented, providing part-of-speech (PoS) tags for tokens in the corpus, rather than segmentation of word-forms into morphemes and allomorphs with their particular feature annotation. The notion of *morphological information* is used inconsistently in the literature, e.g. associated with lexeme and PoS information only. The documented Croatian morphological lexicon (Oliver and Tadić, 2004) for example does not provide information about the morphological structure and specific feature annotations of single morphemes, but rather word-forms and lexemes with PoS-annotation.

On the other hand, specific research questions, require detailed morphological analyses of lexical tokens in a corpus. In our particular case, the Croatian Language Corpus (Brozović-Rončević and Ćavar, 2008), as one of our major data sources needs to be annotated for subsequent analysis.

To be more precise, our understanding of a morphological lexicon and morphological corpus annotation, i.e. our desired type of information, consists of parses of lexemes on the morphological level, and explicit feature bundles associated with each single morpheme or allomorph, as shown in table 1 for the word *popijemo* (Croatian, "to drink (up)").

token	*popijemo*		
parse	*po*	*pije*	*mo*
	stem		suffix
	prefix	root	inflectional
	aspect	verb	1^{st}
	perfective	transitive	plural

Table 1: Morphological parse example

We refrain from providing a hierarchical tree structure for morpheme relations, although it might be useful to reveal scope ambiguities of semantic properties.[1] Parses with just a linear segmentation, including a quasi-hierarchical dependency with for example the prefix and root being contained in the stem, as shown in table 1, provide enough valuable information for some advanced research questions.

For specific research purposes, in particular questions about morphological ambiguity load, the desired morphological analysis of a corpus should provide all possible parses and feature bundles for each morpheme, be it a morphological root or affix (potentially a null-affix), as partially shown in table 2.

Looking at textual data from Croatian synchronic and diachronic dialects and variants, as well as the standard language, we are facing various problems, as for example:

– Different orthography standards have been, and are still used.

– The lexical environment is not static, with lexical

We are grateful to Thomas Hanneforth, Adrian Thurston, and Darko Veberič for their comments and response related to the code and technical realization, and to our colleagues at the IHJJ for lexical material and advice. Furthermore, we thank several anonymous reviewers for helpful hints and comments.

[1]An elegant description of the ambiguity of the word *unlockable* for example could be based on a hierarchical representation as [*un* [*lock able*]] or [[*un lock*] *able*]. A flat and linear segmentation representation is not as intuitive.

token	*kocka*	
parse 1	*kocka*	Ø
	root	suffix
	noun	singular
	feminine	nominative
parse 2	*kocka*	Ø
	root	suffix
	verb	3^{rd}
	intransitive	singular
	denominal	
...		

Table 2: Morphologically ambiguous parse example

items emerging and disappearing, their semantic properties changing etc.

– Lexical changes occurred, some might have affected the morphological makeup of individual word-forms (including changes in paradigms), some might be related to different feature bundles associated with them.

Since various domains of lexical and morphological properties and features in our particular case are still subject to ongoing research, the set of features is necessarily open and unspecified from the outset. We expect in particular semantic properties, new feature types that result from linguistic conceptual necessities, or marking of linguistic origin and cultural background to emerge during future studies. Also, identifiers for named entities appear to be natural extensions of the feature set. The annotations should be extensible with respect to these properties. Any extension of the morphological feature set should directly be integrated into the annotations of an entire corpus.

Once the morphological segmentation is available, further annotations and analyses can be incorporated in a trivial way. For example, the generation of lemmata for segmented word-forms can be achieved by appending the canonical inflectional suffix to the identified base, and potentially applying the necessary allomorphic change to the root. Furthermore, for establishing associations of word-forms to semantic fields, i.e. identifying the semantic root of a complex word-form, the lemma of the root provides a useful additional annotation information. For most Slavic and Germanic languages the rightmost root in a word-form is the semantic head of a complex morpheme. Thus, a general strategy for the identification of the root-lemma is to pick the rightmost root morpheme and append to it the canonical inflectional suffix.

This strategy is on the one hand attractive, because by identification of the lexical root one can relate complex word-forms to their underlying semantic concept. On the other hand, many word-forms are not strictly compositional to allow this. Consider the word-form *neprijatelj* ("enemy"):

– Surface form: *neprijatelj*

– Segmentation: *ne* <prefix> – *prijatelj* <root>

– Root lemma: *prijatelj*

– Base lemma: *neprijatelj*

The root-lemma in this case is not a direct derivation or composition of the negation operator and the lexical root. The meaning of ¬`prijatelj` is not `neprijatelj`, although ¬`prijatelj` might be one of many conceptually true properties of `neprijatelj`. Nevertheless, providing the root-lemma even in this case allows for associations of lexical items with fundamental concepts and word classes, in particular, if derivational morphology is involved in the word-formation process of a surface form. Our goal is thus to annotate the corpus for both lemma types, i.e. the root- and the base-lemma. The latter is achieved by inclusion of all prefixes in the lemma formation rule that are part of the morphological base.

To achieve this, a software solution is necessary. Manual annotation of large corpora is not feasible, it would lack consistency on the large scale, i.e. it is error prone. A software solution allows for systematic and consistent annotation of large corpora. Errors in the annotation should be systematically corrected in the grammar and formalism of some algorithmic annotation solution, rather than corrected manually in a corpus.

To sum up, a software solution should have the following properties:

a. It should provide parses of word-forms into morphemes.

b. It should provide annotations (feature bundles) for each single morpheme.

c. It should be extensible, wrt. the feature-bundles associated to morphemes, as well as to the list of morphemes as such.

A software solution, however, to our knowledge, does not exist for the languages we are interested in.

The specification might look like an all-in-one device for every purpose. However, its development appears to be feasible, with a very simple, and nevertheless efficient technical solution, i.e. finite state transducers (Berstel, 1979; Berstel and Reutenauer, 1988). In the following we describe the algorithmic specification of CroMo, the morphological parser, annotator and lemmatizer, developed for the Croatian standard, and synchronic and diachronic variants.

1.1. Previous approaches

Finite state methods for computational modeling of natural language morphology are wide-spread and well understood. Various commercial and open-source FSA-based development environments, libraries and tools exist for modeling of natural language morphology. A detailed discussion of their properties and application for various languages would be beyond the scope of this article. Some overview can be found in recent literature, e.g. (Sproat, 2000; Beesley and Karttunen, 2003; Roark and Sproat, 2007), further links to literature and implementations can be found in the context of the OpenFst library (Allauzen et al., 2007).

For Croatian there are various descriptions of the formalization and computational modeling of morphology in terms of finite state methods (Tadić, 1994; Lopina, 1999). However, an implemented testable application is not available.

Some solutions that have been implemented for example for German come close to the system requirements specified above. The SMOR (Schmid et al., 2004) and Mor-

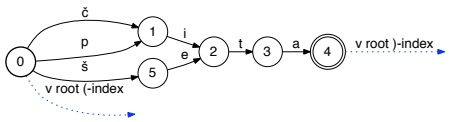
phisto (Zielinski and Simon, 2008) systems partially represent such a type of computational morphology application. An almost complete overlap of features and properties can be found in the implementation of the German morphology as described in the TAGH (Geyken and Hanneforth, 2005) system.

2. FST for morphological segmentation

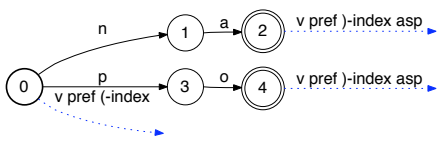
For various reasons, we decide to stick to the approach and implementation strategy of TAGH, while we apply our own experimental libraries and development environment.[2] Following the TAGH-approach (Geyken and Hanneforth, 2005), we model Croatian morphology by referring exclusively to morphotactic regularities, using morpheme and allomorph sets and regular morphological rules, such that a deterministic finite state transducer (FST) can be generated.

The initial modeling step consists of grouping of morphemes. While each application might involve different considerations about how morphemes have to be grouped, in general if should be based on criteria like (a.) having the same feature specification, and (b.) being subject to the same morphological rules.

In CroMo each morpheme group represents one deterministic and acyclic finite state transducer (DFST). The design is similar to the Mealy (Mealy, 1955) or Moore machine (Black, 2004). Every morpheme DFST emits on entry a tuple of the byte-offset in the input string, and the feature bundle that is associated with the DFSA path. In every final state the DFST emits the same tuple. This way morphemes are marked with a start and end index, as well as the corresponding feature bundle, representing the desired annotation. The following graph shows a simplified example of an acyclic DFST for verbal roots:



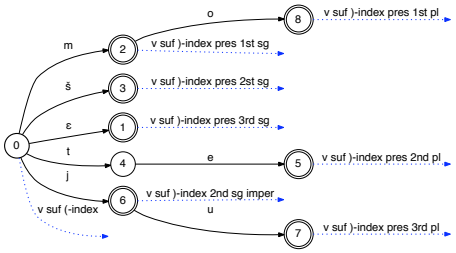
In the same way verbal (e.g. aspectual) prefixes are organized into acyclic DFSTs, as shown in the following graph:



The verbal inflectional paradigm is organized in the same way. Since the model is based on purely morphotactic distributional regularities, potential phonological phenomena are expressed using exclusively allomorphic variations.

[2]The automata and grammar definitions we use are compatible with several existing systems and libraries.

The following graph shows a simplified network for verbal suffixes:

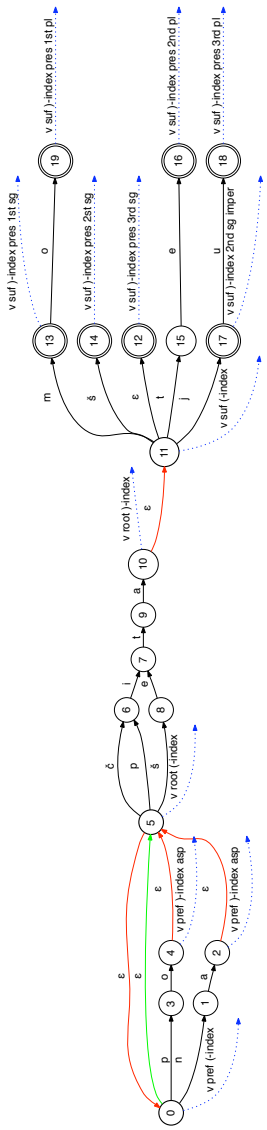


Once all morphemes are grouped into DFSTs, and the appropriate emission symbols (the annotations) are assigned to each entry and final state of the DFST, each morpheme group is assigned an arbitrary variable name, which is used in the definition of rules. A rule that makes use of the automata above could be defined as follows:

```
vAspectPref* . vAtiRoots . vInflSuf
```

This rule describes the concatenation of the verbal root DFST with the DFST for the verbal inflectional paradigm, using common regular expression notation. In this case we use the regular expression syntax as defined for the Ragel (Thurston, 2006) state machine compiler. Additionally, the prefixes are defined as optional and potentially recursive prefixes concatenated with the verbal root DFST. This definition generates a cyclic[3] deterministic transducer with every final state of beginning and intermediate sub-DFSTs into a non-final state, linked to the initial state of the concatenated DFSTs via an ϵ-transmission, as shown in the following graph:

[3]Cyclicality in this particular case leads to more compact automata. In principle, the depth of recursion of such prefixes could be limited (empirically and formally), and formalized using the appropriate regular expression syntax.



Such a DFST emits a tuple containing the byte-offset and the corresponding annotation symbols at the initial state, and at each morpheme boundary (former initial and final states of the sub-DFSTs).

Using this approach, all lexical classes are defined as complex (potentially cyclic) DFSTs, and combined, together with the closed class items, as one monolithic DFST.

The advantage of such a representation is not only that the resulting morphological representation is compressed, but also that it is processed in linear time, with the identification of morpheme boundaries and corresponding feature bundles being restricted by contextual rules.

In order to cope with morphological ambiguity, this approach is extended. In principle there are two major approaches to deal with ambiguity, either one has to allow for non-deterministic automata (two different transitions with the same input emit a different output tuple), or ambiguity is mapped on the emission of multiple annotation tuples. In the case of CroMo, the latter option is used in the modeling. Every emission is a tuple of length 0 to n, such that e.g. orthographically ambiguous nominal suffixes like *a* (genitive singular or plural) are modeled as a single transition in a DFST with the final state emitting two annotation tuples that contain the specific case and number features.

2.1. Interoperability and annotation standard

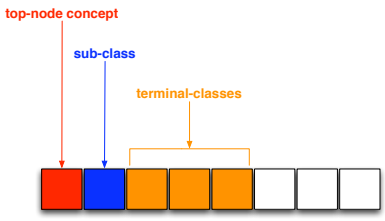
Current language resources face a serious problem, related to issues of interoperability and annotation compatibility. Various different tag-sets are used for particular languages, and some of those tend not to be straight-forward compatible. In the same way, linguistic annotation tools do not necessarily make use of some standardized tag-set, and such a tag-set actually does not even exist.

For our purposes here we decided to offer maximal interoperability in the resulting corpus annotation, as well as in the annotation tool as such, one that is maximally compatible with existing tag-sets, as language specific as necessary, and at the same time maximally extensible. The General Ontology for Linguistic Description (GOLD) (Farrar and Langendoen, 2003; Farrar et al., 2002; Farrar, 2003) *was originally envisioned as a solution to the problem of resolving disparate markup schemes for linguistic data*. GOLD specifies basic linguistic concepts and their interrelations, and can be used, to a certain extent, as a description logic for linguistic annotation.

We make use of three core concept classes in GOLD, and the necessary sub-concepts, i.e. MorphoSemanticProperty, MorphosyntacticProperty, and LinguisticExpression. The concepts defined therein relate to the notions that are expected to be emitted by CroMo, i.e. morphological properties of morphemes (e.g. prefix, suffix, root), morphosyntactic properties (e.g. case, number), and morphosemantic properties (e.g. aspect, mood, tense).

By using the labels for concepts as defined in GOLD, we should be able to maintain maximal compatibility with other existing tag-sets. While the logic of GOLD would burden a morphological parsing algorithm, the reference to the concepts doesn't seem problematic. Representing the concepts as pure emission strings associated to the emission states, as discussed above, might decrease memory and performance benefits of a DFST-based analyzer. To maximize the performance, the GOLD-concepts and relations are mapped on a bit-vector. Encoding of the relevant concepts can be achieved with bit-vectors of less than 64 bit.

The mapping defines constants that correspond to bit-masks that are pre-compiled into the DFST. The bit-mask for example for `Genitive` might be defined as one that corresponds to set first and second bits of the terminal-class bit-field, plus the corresponding bits that indicate that the sub-class `CaseProperty` is set, as well as the bit for the corresponding top-node class `MorphosyntacticProperty`, as shown in the following graphic:



In a limited way, via definitions of constants and mapping of linguistic annotation in the morpheme dictionaries, one can maintain implicatures and inheritance relations, as defined in the ontology, via bit-vector representations and appropriate bit-masks.

For the morphological analyzer this does not imply any additional processing load, i.e. the emission tuples consist of bit-vectors in form of 4-byte numerical integer values. Converting the emission tuples (i.e. individual bit-vectors) into literal string representations can be achieved efficiently, once an input string is analyzed completely.

2.2. Implementation

CroMo consists of two sets of code-bases. The first component converts a lexical base into a formal automaton definition. The second compiles together with the automaton definitions into a binary application.

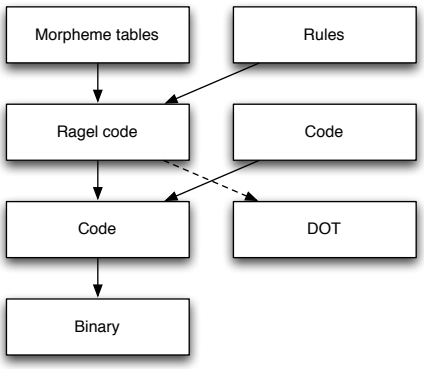
The lexical base is kept either in database tables, spreadsheets, or textual form. The different formats allow us to maintain a minimally invasive lexical coding approach. Linguists or lexicologists are not required to learn a formal language for DFST definitions. Furthermore, they are free to use their individual way of annotation, being guided by GOLD concepts, but free to define their own, should these not be part of GOLD. CroMo provides guidelines for the data-format, but also the possibility to use individual scripts for data conversion and annotation mappings.

The individual morpheme lists, annotations and rule definitions are compiled into Ragel (Thurston, 2006) automata definitions, as described above. Besides rules that are related to concrete morpheme lists and the corresponding DFSTs, there are also guessing rules that define general properties of nouns, verbs and adjectives. The features that are used, be they specified in GOLD or not, are mapped on bit-vectors, and C-header files with the constant literal and bit-vector mask definitions are generated.

Ragel generates a monolithic DFST as C-code, using highly efficient C-jump code (`goto`-statements), as well as a DOT-file for visualization of the resulting automaton (using e.g. Graphviz[4]). The generated code is wrapped in a C++ class that handles input and output, and controls the program logic.

In the current version the generation of the root- and the base-lemma is encoded in the emission bit-vector. One byte is reserved to mark the reverse offset for string concatenation, while two bytes are reserved to point to an element in a string array with the corresponding string that needs to be appended. The form *čitamo* would be associated with an offset of -2 and a corresponding suffix *ti*. This solution doesn't match the general paradigm, and is just temporary. In the next release the output characters of the corresponding lemma will be integrated in the emission of the transducer, associated with each single transition. Thus every emission will be a tuple that contains tuples of output characters and optional annotation bit-vectors.

The following graphic shows the general automaton compilation workflow, implemented in a single shell-command:



CroMo expects a token list as input. Tokens are processed sequentially. For each token, all emitted tuples are collected in a stack. Only matching start- and end-tuples are returned, if there are compatible sub-morpheme analyses that span over the complete input token length.

The significant implementation features that differentiate CroMo from other solutions, are:

1. The code-base is platform independent and free open-source, using only free and open tools like GCC and Ragel for compilation.

2. CroMo doesn't depend on any particular encoding standard. The lexical base can be encoded in any common encoding, since the automaton processing is purely binary (i.e. byte-based). The character encoding of the morpheme definitions and the input tokens should match. In principle it is possible to use multiple different encodings.

The extension of the morphological base is kept trivial, along the lines of the requirements specified above, i.e. the necessity to be able to add newly identified morphemes or paradigms from diachronic and synchronic variants.

3. Evaluation

The evaluation version of CroMo contains approx. 120,000 morphemes in its morpheme-base, using UTF-8 character encoding. The number of strings it can recognize is infinite, due to cyclic sub-automata. Unknown word-forms can be analyzed due to incorporated guessing rules.

For the following evaluation results we used a 2.4 GHz 64-bit Dual-Core CPU. In the evaluation version only a single core is used during runtime of the FST, while both CPU cores are used during compilation.

Compilation of the morphology requires min. 4 GB of RAM using GCC 4.2. This is expected due to the monolithic architecture, and since the Ragel-generated C-code of the transducer gets very large. The compilation process takes less than 5 minutes, using both CPU cores. The resulting binary footprint is less than 5 MB of size.

The final automaton consists of approx. 150,000 transitions and 25,000 states.

[4] See `http://www.graphviz.org/` for details.

We selected randomly 10,000 tokens with an average morpheme length of 2.5 morphemes. CroMo processes in average approx. 50,000 tokens per second (real 10,000 tokens per 150 millisec.), including runtime instantiation in memory, mapping of the analysis bit-vectors to the corresponding string representations, generation of lemmata, and output redirection to a log-file. An extension of the morpheme base has no significant impact on memory instantiation time, neither on the runtime behavior. The memory instantiation can be marginalized for a large processing sample.

CroMo doesn't implement transitional or emission-probabilities, due to missing quantitative information from training data. Once an annotated corpus is available, these weights can trivially be implemented as additional weights in the emission tuple.

A relevant evaluation result is the coefficient of the ratio between all and relevant emissions, i.e. the percentage of relevant (possible) morpheme analyses and all generated ones. Due to certain limitations, we did not perform such an evaluation, neither a recall evaluation on a predefined evaluation corpus. The results of these evaluations, together with the source code, will be made available on CroMo's web site `http://personal.unizd.hr/~dcavar/CroMo/`.

4. Comments

CroMo manifests a highly efficient morphological segmentation and annotation algorithm, with little margin for efficiency improvement in the code base.

The use of GOLD as an annotation standard has been shown to be feasible, on the implementation level. However, GOLD is still under development. Many necessary concepts have not yet been implemented. A benefit of using a DL for annotation has yet to be shown. Once necessary syntactic concepts are specified in GOLD, syntax-based disambiguation would be feasible. GOLD mappings to different tag-sets have still to be established, to fulfill the promise of interoperability and annotation compatibility.

5. References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. OpenFst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*, pages 11–23. Springer-Verlag.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Publications, Stanford, April.

Jean Berstel and Christophe Reutenauer. 1988. *Rational Series and Their Languages*. EaTCS Monographs on Theoretical Computer Science. Springer-Verlag, Berlin, December.

Jean Berstel. 1979. *Transductions and Context-Free Languages*. Teubner Studienbücher, Stuttgart.

Paul E. Black. 2004. Dictionary of algorithms and data structures. Online publication: U.S. National Institute of Standards and Technology, Available from http://www.nist.gov/dads/HTML/mooreMachine.html, December.

Dunja Brozović-Rončević and Damir Ćavar. 2008. Hrvatska jezična riznica kao podloga jezičnim i jezičnopovijesnim istraživanjima hrvatskoga jezika. In *Vidjeti Ohrid*, Hrvatska sveučilišna naklada, pages 173–186, Zagreb. Hrvatsko filološko društvo.

Gavin Burnage. 1990. CELEX - A guide for users. Technical report, Centre for Lexical Information, University of Nijmegen, Nijmegen.

Scott O. Farrar and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *Glot International*, 7(3):1–4, March.

Scott O. Farrar, William D. Lewis, and D. Terence Langendoen. 2002. A common ontology for linguistic concepts. In N. Ide and C. Welty, editors, *Semantic Web Meets Language Resources: Papers from the AAAI Workshop*, pages 11–16. AAAI Press, Menlo Park, CA.

Scott O. Farrar. 2003. *An Ontology for Linguistics on the Semantic Web*. Ph.D. thesis, The University of Arizona, Tucson, Arizona.

Alexander Geyken and Thomas Hanneforth. 2005. TAGH: A complete morphology for german based on weighted finite state automata. In Anssi Yli-Jyrä, Lauri Karttunen, and Juhani Karhumäki, editors, *FSMNLP*, volume 4002 of *Lecture Notes in Computer Science*, pages 55–66. Springer, September.

Vjera Lopina. 1999. Strojna obrada imenične morfologije u pisanome hrvatskom jeziku. Ma thesis, Centar za postdiplomske studije Dubrovnik, Dubrovnik, October.

George H. Mealy. 1955. A method for synthesizing sequential circuits. *Bell System Technical Journal*, 34(5):1045—1079, September.

Antoi Oliver and Marko Tadić. 2004. Enlarging the croatian morphological lexicon by automatic lexical acquisition from raw corpora. In *Proceedings of LREC 2004*, volume IV, pages 1259–1262, Lisbon, May. ELRA.

Brian Roark and Richard Sproat. 2007. *Computational Approaches to Syntax and Morphology*. Oxford University Press, Oxford.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A german computational morphology covering derivation, composition, and inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266, Lisbon, Portugal.

Richard Sproat. 2000. *A Computational Theory of Writing Systems*. AT&T Bell Laboratories, New Jersey, July.

Marko Tadić. 1994. *Računalna obradba morfologije hrvatskoga književnog jezika*. doctoral dissertation, Filozofski fakultet Sveučilišta u Zagrebu, Zagreb, Croatia.

Adrian D. Thurston. 2006. Parsing computer languages with an automaton compiled from a single regular expression. In *11th International Conference on Implementation and Application of Automata (CIAA 2006)*, volume 4094 of *Lecture Notes in Computer Science*, pages 285–286, Taipei, Taiwan, August.

Andrea Zielinski and Christian Simon. 2008. Morphisto – an open-source morphological analyzer for german. In *Proceedings of FSMNLP 2008*, Ispra, Italy, September.

Productivity of concepts in Serbian Wordnet

Jelena Tomašević, Gordana Pavlović-Lažetić

Faculty of Mathematics, University of Belgrade
Studentski trg 17, 11 000 Belgrade, Serbia
{jtomasevic, gordana}@matf.bg.ac.yu

Abstract

Wordnet is an online lexical database designed for use under program control. It is based on word meaning, rather than word forms. All of the words that can express a given sense are grouped together in a synonym set (synset) representing a concept. All concepts are linked with semantic relations forming semantic network. The network is basically a forest consisting of concept hierarchies rooted in top ontology concepts. In this paper we describe several measures for determining productivity of some concept in order to find those concepts that most effectively represent hierarchy they belong to. They are different from top ontology concepts (which are too general), and could be considered as ontological concepts associated with classes characterized by hierarchies rooted in them. Determining most productive concepts may be applied to text classification in different ways. Information retrieval and information extraction could be made more efficient if they are based on such kind of classification.

1. Introduction

The main goal of the Princeton Wordnet (PWN) (Miller, 1995), or simply Wordnet, developed by George Miller and his associates was to serve as a mental lexicon in the scope of psycholinguistic research projects. In a traditional dictionary, lexical items are listed alphabetically, giving definitions for each sense. Wordnet, in contrast, is based on word meaning; all of the words that can express a given sense are grouped together in a synonym set, or synset representing a concept. It is a set of approximately 100.000 concepts interconnected by semantic relations to form a semantic network. A new dimension to the Wordnet project was added by EuroWordnet (EWN) (Vossen 1998), a multilingual database with wordnets for Dutch, Italian, Spanish, German, French, Czech and Estonian. The aim of BalkaNet, the Balkan wordnet project (BWN) (Stamou et al., 2002), was to develop aligned semantic networks for several Balkan languages, namely Bulgarian, Greek, Romanian, Serbian and Turkish, as well as to extend the existing network for Czech, initially developed within the EWN project. The BWN and EWN wordnets are all linked to an Inter-Lingual-Index (ILI), which makes it possible to relate similar concepts in different languages, a feature that can be used, among others, for cross-language information retrieval. The BWN, as a multilingual database, in its initial phase followed the basic pattern set by EWN. The formation of databases started from a translation of a common set of concepts named “basic concepts” (see paragraph below) in EWN. Further development of the BWN databases was aimed at maintaining a common set of concepts to be shared among languages, while at the same time, giving liberty to each Wordnet to introduce specific concepts for its own language on an as needed basis. In the absence of both an explanatory dictionary for Serbian and an English/Serbian dictionary in electronic form, translation of English synsets from PWN into Serbian was done manually. It brought up the question of validation of Serbian synsets on corpora. The use of monolingual and multilingual corpora for synset validation led to introduction of new literals or removal of existing ones from a synset (Krstev et al., 2003; Obradović et al., 2003).

Wordnet is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. When psychologists think about the organization of lexical memory it is nearly always the organization of nouns that they have in mind. Nominal concepts are organized hierarchically into levels, from specific to generic. The top-most, or most generic level of the hierarchy is almost vacuous semantically. If these hierarchies were inheritance systems, they seldom go more than ten levels deep, and those cases usually contain technical levels that are not part of the everyday vocabulary. These hierarchies of nominal concepts have been said to have a level, somewhere in the middle, where most of the distinguishing features are attached. It is referred to as the base level of the noun lexicon, and concepts at this level are “basic concepts”. These concepts are not too specific and not too general. For lexical concepts at the base level, people can list many distinguishing features. Above the base level, descriptions are brief and general. Below the base level, little is added to features that distinguish “basic concepts”.

In order to formalize the notion of "basic concept", we introduce the notion of "productivity" of concept that represents how much that concept effectively represents hierarchy which belongs to. In this manner, “basic concepts” could be defined as all the concepts that have value of productivity in some range near the largest value of the defined measure of “productivity”. These “basic concepts” could be considered like the most representative concepts of the hierarchy which belong to.

Measures of productivity could be defined in many different ways. For example, for some selected concept, it could be defined as a product of the number of all descendants (or just all leaves) of that concept and the distance of that concept from the root of the hierarchy (number of concepts that are in its path to the root). Measure of productivity of a concept could also be defined as a ratio of the number of its in the tree rooted in that concept, to the number of descendants in trees rooted in his siblings (concepts at the same level as the considered concept). The choice between those measures depends on the purpose of productivity measurement.

Wordnet has been used in numerous natural language processing tasks. It is widely used in text classification which is an important application of machine learning.

Most productive concepts and hierarchies rooted in them may support a new approach to text classification.

The paper proceeds as follows. In section 2 we present the structure of Serbian wordnet. Section 3 presents different measures of concept productivity. In section 4 we present comparison of Serbian and English Wordnet. In section 5 one example of most productive concepts is presented for hierarchy rooted in "entity". Finally, section 6 presents some examples of applying most productive concepts to natural language processing and section 7 presents the conclusion and future work.

2. Serbian wordnet - SWN

Serbian wordnet (SWN) is a lexical-semantic network of Serbian language (Pavlović-Lažetić, 2006). The structure of SWN is basically the same as the structure of American wordnet for English, PWN. It is based on concepts called synsets. Only words of the same Part of Speech (PoS) **(e.g. nouns, verbs)** can belong to the same synset. A synset ID is assigned to every word. Words in the same synset have the same synset ID. Since one word can have several meanings, it can belong to more than one synset. Each literal string is accompanied by a unique sense number denoting **its meaning within a particular synset**.

SWN contains only nouns, verbs, adjectives and adverbs. It presently contains approximately 14000 sets of synonyms (Krstev et al, 2008). Except for the synonymy relation, defining the concept of a synset, the most important relations between concepts is the hypernym / hyponym (is-a) relation. Because there is usually only one hypernym, this semantic relation organizes the meanings of nouns into a hierarchical structure. The synsets for nouns refering to actions are much more structured than synsets for verbs (more exactly, there is a long and unstructured list of top-level synsets for verb meanings). So, we took into consideration just noun concepts. There are 9 top-level synsets in Serbian Wordnet for noun meanings that correspond to relatively distinct semantic fields: "entity" ("entitet" in Serbian), "abstraction" ("apstrakcija"), "group" ("grupa"), "act" ("akt"), "psychological feature" ("psihičko svojstvo"), "state" ("stanje"), "event" ("događaj"), "phenomenon" ("fenomen") and "possession" ("svojina").

All national wordnets share the same data structure in XML. An example of a synset from the SWN and PWN, corresponding to the concept "deity" ("božanstvo") is represented in Figure 1\.

The meaning of (some of) the elements in this figure is the following: ID is the synset identifier, unique across different languages, POS is the part of speech, ILR is the ID of a concept (synset) related to this one by a relation of the type TYPE, etc.

Table 1 represents the number of synsets in SWN per specific PoS, PoS related distribution of synsets in SWN, number of synsets and PoS related distribution of synsets in PWN, and last column in the table represents ratio between number of synsets in SWN and PWN per specific PoS.

```
<SYNSET>
 <ID> ENG20-08904620-n </ID>
 <SYNONYM>
   <LITERAL>bozxanstvo<SENSE>1</SENSE></LITERAL>
 </SYNONYM>
 <DEF>
 Natprirodno bicxe koje se obozxava zbog verovanxa da upravlxa
 nekim delom sveta ili nekim aspektima zxivota ili zato sxto
 personifikuje silu.
 </DEF>
 <POS>n</POS>
 <ILR>ENG20-08903509-n<TYPE>hypernym</TYPE></ILR>
 <ILR>ENG20-07660421-n<TYPE>holo_member</TYPE></ILR>
</SYNSET>

<SYNSET>
 <ID> ENG20-08904620-n </ID>
 <SYNONYM>
   <LITERAL>deity <SENSE>1</SENSE></LITERAL>
   <LITERAL>divinity <SENSE>1</SENSE></LITERAL>
   <LITERAL>god <SENSE>2</SENSE></LITERAL>
   <LITERAL>immortal <SENSE>2</SENSE></LITERAL>
 </SYNONYM>
 <DEF>
 any supernatural being worshipped as controlling some part of the
 world or some aspect of life or who is the personification of a force.
 </DEF>
 <POS>n</POS>
 <ILR>ENG20-08903509-n<TYPE>hypernym</TYPE></ILR>
 <ILR>ENG20-07660421-n<TYPE>holo_member</TYPE></ILR>
 <ILR>ENG20-00670299-v <TYPE>eng_derivative</TYPE></ILR>
</SYNSET>
```

Figure 1. A synset of the Serbian and its corresponding synset of the English Wordnet

PoS	SWN	percents (SWN)	PWN	percents (PWN)	PWN/SWN
nouns	11 155	80.1%	79 689	69%	7.1
verbs	1 945	14%	13 508	12%	6.9
adjectives	793	5.7%	18 563	16%	23.4
adverbs	27	0.2%	3 664	3%	135.7
total	13 920	100%	115 424	100%	8.3

Table 1: PoS related distribution of synsets in SWN and PWN

3. Measures of concept productivity

As we mentioned earlier, we took into consideration just noun concepts and hypernym / hyponym relations between them.

In order to formalize the notion of "basic concept", we have introduced the notion of "productivity" of concepts. In this manner, the "basic concepts" could be defined as all concepts that have value of productivity in some range close to the largest value of defined measure of productivity. These "basic concepts" could be considered

as the most representative concepts of the hierarchies which belong to.

Measures of concept productivity could be defined in different ways. For example, for a selected concept, it could be defined as a product of the number of all descendants (or just leaves) of that concept and distance of that concept from the root of the corresponding hierarchy (number of concepts that are in its path to the root). Formally, for a selected concept *c* which belongs to a hierarchy rooted in the concept *root*, it could be represented as follows:

Measure 1:

$$ProductivityOfConcept(c) = ND(c)^{alpha} * d(c)^{beta}$$

Measure 2:

$$ProductivityOfConcept(c) = NL(c)^{alpha} * d(c)^{beta}$$

where *ND(c)* is the number of descendants in a tree rooted in the selected concept *c*, *NL(c)* is the number of leaves in the tree rooted in *c,* and *d* is a distance of *c* from the root of the hierarchy. *alpha* and *beta* are parameters which determine weights (or significance) that the number of descendants (or leaves), and the distance from the root, respectively, have in the overall measure. In our experiments we have used *Measure1* with *alpha*=1 and *beta*=2. If *alpha* has bigger and *beta* has lower value, then concept with the largest value of this measure will go upper to the hierarchy which belongs to.

At the other hand, measure of productivity of a concept could be defined as a ratio of the number of descendants in a tree rooted in the selected concept, to the number of descendants in trees rooted in his siblings. Formally, it could be represented as follows:

Measure 3:

$$ProductivityOfConcept(c)=ND(c)/summ(ND(siblings))$$

where *ND(c)* is the number of descendants in the tree rooted in the concept *c*.

It is possible to define measure of productivity at many other deferent ways. The choice between these measures depends on the purpose of productivity measurement.

"possession" "svojina"	most productive concept (SWN)	percents, level (SWN)	most productive concept (PWN)	percents, level (PWN)
Measure 1	"cost" "potrošnja"	10%, 6	"cost" "potrošnja"	30%, 6
Measure 2	"cost" "potrošnja"	10%, 6	"cost" "potrošnja"	30%, 6
Measure 3	"estate" "imanje"	6.25%, 4	"financial loss" "finansijski gubitak"	33%, 4

Table 2: Most productive concepts for the hierarchy rooted at "possession" in SWN and PWN using deferent measures

For example, the most productive concept for hierarchy rooted at "entity" ("entitet") (which has the largest number of concepts), if we use *Measure 1* with *alpha*=1 and *beta*=2, is the concept "organism" or "being". Tree rooted in that concept contains 44% of all concepts in that hierarchy in English wordnet (PWN) and 55% in SWN and it is at the third level of hierarchy (see section 5). For the hierarchy rooted at "possession" ("svojina") (which has the least number of concepts), if we use the same measure, most productive concept is "cost" ("potrošnja"). Tree rooted in that concept contains 10% of all concepts in that hierarchy in SWN and 30% in PWN and it is at the sixth level of hierarchy. Table 2 represents the most productive concepts for the concept "possession" ("svojina") using all three previously defined measures.

4. Comparison of Serbian and Englsih wordnets

Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. The database is restricted to the relations suggested by the psycholinguistic data: synonymy, antonymy, hyponymy, hypernymy, and three types of meronymy and holonymy. These and other similar relations serve to organize the mental lexicon. They can all be thought of, or represented by, pointers or labeled arcs from one synset to another.

The relations between synsets in SWN and PWN are summarized in Table 3. These relations in SWN have automatically been inherited from the PWN, but then they were all manually checked.

relation	PoS/PoS (N=Nouns, V=Verbs, A=Adjectives)	SWN	PWN	PWN /SWN
hypernym	N/N V/V	13034	94844	7.28
near_antonym	N/N A/A V/V	680	7642	11.2
holo_part	N/N	701	8636	12.3
verb_group	V/V	163	1748	10.72
holo_member	N/N	3309	12205	3.69
be_in_state	A/N	268	1296	4.84
subevent	V/V	76	409	5.38
causes	V/V	58	439	7.57
derived	A/N	326	6809	20.89
particle	A/V	10	401	40.1

Table 3: Distribution of relations between synsets in SWN and PWN

Wordnet includes the following semantic relations:

- *Synonymy* is Wordnet's basic relation, because Wordnet uses sets of synonyms (*synsets*) to represent word senses. It can be used to map words with similar meanings together e.g. "happy" and "glad".
- *Hyponymy* (subordination) and its inverse, h*ypernymy* (superordination) can be used to generalize noun and verb meanings to a higher level of abstraction[1]. The hyponym inherits all the features of the more generic concept and adds at least one feature that distinguishes it from its

[1] The hypernymy relation is defined only for nouns and verbs.

superordinate and from other hyponyms of that superordinate. The Figure 2 shows an idealized model of a part of the Wordnet semantic hierarchy and Figure 3 shows it as a part of an XML document.

- *Antonymy* (opposing-name) is especially important in organizing the meanings of adjectives. The antonym of a word x is sometimes not-x, but not always. For example, “rich” and “poor” are antonyms, but to say that someone is not rich does not mean that he must be poor.
- *Meronymy* (part-name) and its inverse, *holonymy* (whole-name), are complex semantic relations. Wordnet distinguishes *component* parts, *substantive* parts, and *member* parts. For example, “paper” is a meronym of “book”, since paper is a part of a book.
- *Troponymy* (manner-name) is for verbs what hyponymy is for nouns. The "kind-of" relation among nouns corresponds to a "manner-of" relation among the verbs. For example, “walk” is a troponym of “move” and “limp” is a troponym of “walk”.
- *Entailment* refers to a relationship between two verbs. A verb x entails y if the truth of y follows logically from the truth of x. The relation of entailment is unilateral, i.e., it is one way relation. For example, concept "kill" entails concept "die".

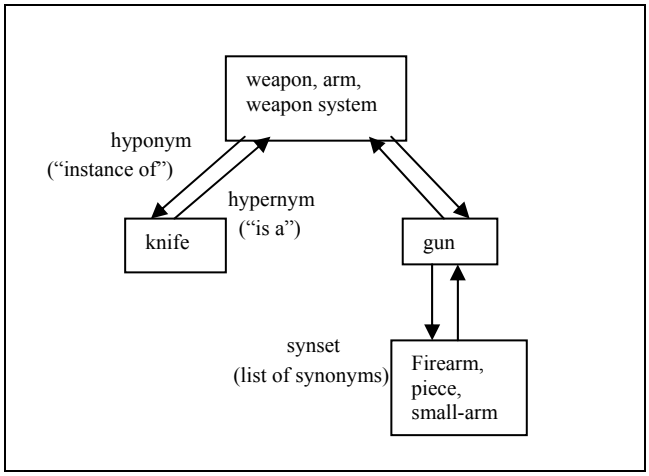


Figure 2. A piece of the Wordnet semantic hierarchy (idealized model)

In the previous section we have defined several measures for a concept that may determine its “productivity” or how much that concept effectively represents hierarchy which belongs to. “Basic concepts” could be considered as the most productive concepts for the corresponding hierarchy.

```
<SYNSET>
 <ID> ENG20-04387884-n </ID>
 <SYNONYM>
   <LITERAL>weapon<SENSE>1</SENSE></LITERAL>
   <LITERAL>arm<SENSE>2</SENSE></LITERAL>
   <LITERAL>weapon system<SENSE>1</SENSE></LITERAL>
 </SYNONYM>
 <DEF>
 any instrument or instrumentality used in fighting or hunting
 </DEF>
 <USAGE>he was licensed to carry a weapon</USAGE>
 <POS>n</POS>
 <ILR>ENG20-03443087-n<TYPE>hypernym</TYPE></ILR>
</SYNSET>

<SYNSET>
 <ID> ENG20-03489316-n </ID>
 <SYNONYM>
   <LITERAL>knife <SENSE>2</SENSE></LITERAL>
 </SYNONYM>
 <DEF>a weapon with a handle and blade with a sharp point</DEF>
 <DOMAIN>military</DOMAIN>
 <POS>n</POS>
 <ILR>ENG20-04387884-n <TYPE>hypernym</TYPE></ILR>
</SYNSET>

<SYNSET>
 <ID> ENG20-03340812-n </ID>
 <SYNONYM>
   <LITERAL>gun<SENSE>1</SENSE></LITERAL>
 </SYNONYM>
 <DEF> a weapon that discharges a missile at high velocity (especially
       from a metal tube or barrel)</DEF>
 <DOMAIN>military</DOMAIN>
 <POS>n</POS>
 <ILR>ENG20-04387884-n <TYPE>hypernym</TYPE></ILR>
</SYNSET>

<SYNSET>
 <ID> ENG20-03222124-n </ID>
 <SYNONYM>
   <LITERAL> firearm <SENSE>1</SENSE></LITERAL>
   <LITERAL> piece <SENSE>7</SENSE></LITERAL>
   <LITERAL> small-arm <SENSE>1</SENSE></LITERAL>
 </SYNONYM>
 <DEF> a portable gun </DEF>
 <USAGE> he wore his firearm in a shoulder holster </USAGE>
 <DOMAIN>military</DOMAIN>
 <POS>n</POS>
 <ILR> ENG20-03340812-n <TYPE>hypernym</TYPE></ILR>
</SYNSET>
```

Figure 3. A piece of the Wordnet semantic hierarchy (XML representation)

In Table 4 results are given for most productive concepts (MPC) in SWN and PWN for all hierarchies rooted in 9 top-level synsets for noun meanings. Also, number of concepts in the tree rooted in the most productive concept is given in brackets, along with the percentage of those concepts with respect to all the

concepts in the corresponding hierarchy. Results in this table represent most productive concepts using *Measure 1* defined in the previous section, with parameters *alpha*=1 and *beta*=2.

roots	MPC(PWN)	MPC(SWN)
"entity" ("entitet")	"organism" ("organizam") (18997 – 44%)	"organism" ("organizam") (2884 – 55%)
"abstraction" ("apstrakcija")	"communication" ("komunikacija") (4638 – 42%)	"natural language" ("prirodni jezik") (610 – 34%)
"group" ("grupa")	"genus" ("rod") (3652 – 44%)	"taxon" ("takson") (1826 – 81%)
"human action" ("ljudska aktivnost")	"activity" ("aktivnost") (3164 – 47%)	"activity" ("aktivnost") (405 – 55%)
"psychological feature" ("psihičko svojstvo")	"content" ("saznajni sadržaj") (2214 – 46%)	"biology" ("biologija") (31 – 8%)
"state" ("stanje")	"disease" ("oboljenje") (514 – 16%)	"condition" ("postojanje") (99 – 40%)
"event" ("događaj")	"happening" ("dešavanje") (941 – 44%)	"group acton" ("grupna akcija") (115 – 50%)
"phenomenon" ("fenomen")	"physical phenomenon" ("fizička pojava") (513 – 32%)	"light" ("svetlost") (5 – 5%)
"possession" ("svojina")	"cost" ("potrošnja") (233 – 31%)	"cost" ("potrošnja") (8 – 10%)

Table 4: Most productive concepts (MPC) in noun hierarchies, using *Measure 1* of productivity with *alpha*=1 and *beta*=2 in SWN and PWN

From this table some differences in most productive concepts in SWN and PWN could be seen. For example, most productive concept for hierarchy rooted in "psychological feature" ("psihičko svojstvo") in PWN is "content" ("saznajni sadržaj") and in SWN is "biology" ("biologija"). Reason for this is that more concepts were added in SWN that are ancestors of "biology" than of some other concepts, in comparison with PWN. These differences indicate deficiency of SWN and they are good pointers to direction in which SWN should be recharged.

5. One example of the most productive concepts

In this section we present an example of determining the most productive concept for hierarchy rooted in the concept "entity" ("entitet"). In this example we have used *Measure 1* with parameters *alpha*=1 and *beta*=2.

As we could see in Figure 4, the most productive concept for the "entity" ("entitet") hierarchy is "organism" ("organizam"). It contains 44% of all concepts in the corresponding hierarchy in PWN and 55% in SWN and it could be considered as the most productive concept for that hierarchy. This concept belongs to the third level in the whole hierarchy.

entity
(pwn 43255 – 100%)(swn 5278 – 100%)
=> object, physical object
(pwn 31306 – 72%) (swn 4210 – 80%)
=> whole, whole thing, unit
(pwn 10346 – 24%)(swn 1085 – 21%)
=> artifact, artefact
(pwn 10342 – 24%)(swn 1083 – 21%)
=> instrumentality, instrumentation
(pwn 5291 – 12%)(swn 464 – 9%)
=> device
(pwn 2617 – 6%)(swn 206 – 4%)
=> living thing, animate thing
(pwn 19129 – 44%) (swn 2957 – 56%)
=> organism, being
(pwn 18997 – 44%) (swn 2884 – 55%)
=> person, individual, someone, somebody, mortal, human, soul
(pwn 9894 – 23%) (swn 1074 – 20%)
=> plant, flora, plant life
(pwn 4781 – 11%) (swn 952 – 18%)
=> animal, animate being, beast, creature, fauna
(pwn 3989 – 9%) (swn 669 – 13%)

Figure 4. Most productive concept for hierarchy rooted in the concept entity ("entitet")

6. Applications of the most productive concepts

Because meaningful sentences are composed of meaningful words, any system that tends to process natural languages as people do must have information about words and their meanings. This information is traditionally provided through dictionaries, and machine-readable dictionaries are now widely available. But dictionary entries evolved for the convenience of human readers, not for machines. Wordnet provides a more effective combination of traditional lexicographic information and modern computing. It is designed for use under program control.

Wordnet has been used in numerous natural language processing tasks, such as part of speech tagging, word sense disambiguation, text classification, information extraction, information retrieval and so on, with considerable success.

Text classification is an important application of machine learning that can be used in a variety of contexts, including e-mail and news filtering, personal information agents and assistants, information retrieval, and automatic indexing. Text classification is the task of assigning a text document to one or more categories, based on its contents. Usually, text documents were represented as the set of words that occurred in the document, an approach that is often referred to as the *bag of words* model. A number of authors have experimented with automatic text classification systems.

Text classification could be done in different ways by applying measures of productivity. For example, some

well chosen most productive concepts could be associated with some predefined classes in the process of text classification. Text document then could be classified by considering frequency of literals (or some specific literals, e.g., named entities) in hierarchies of those most productive concepts. Information Retrieval (IR) and Information Extraction (IE) could be made more efficient if they are based on that kind of classification.

Information retrieval is concerned with locating documents relevant to user's information needs from a collection of documents. The user describes his/her information needs with a query which consists of a number of words. The information retrieval system compares the query with documents in the collection and returns the documents that are likely to satisfy the user's information requirements. It could be very useful to extend those queries by adding all literals from hierarchy of some well chosen productive concepts.

7. Conclusions and future work

In this paper we described several measures for determining productivity of a Wordnet concept in order to find those concepts that most effectively represent hierarchies which they belong to. For example, experimental results for concept "entity" show that concept "organism" could be considered as the most representative concept for the whole hierarchy rooted in "entity". That is the concept that has maximum value of defined measure of productivity. If we took into consideration all concepts that have value of productivity near maximum value, we could get more concepts that represent this hierarchy, and we could consider them as "basic concepts".

Abundant future work stems from this preliminary study. The first task is to complete this analyzing process. Instead of one most productive concept, we could take into consideration all concepts that belong to some given range of productivity. Recent experiments have suggested that text classification process in that case could be made more efficient. We plan to compare this classification scheme with existing classifying systems such as EAGLES (Sinclair 1996), to elaborate criteria for classifying documents based on most productive concepts and to experiment with this kind of text classification on a real corpus. We also plan to define adequate database structure. Another area for consideration is the IR and IE. Our experience indicates that in some cases, better results can be obtained if IR and IE are based on this kind of classification.

References

G. Pavlović-Lažetić, Electronic Resources of Serbian: Serbian WORDNET, 36th International Slavic Conference, MSC, Belgrade, Serbia, september 2006.

Stamou, S., Oflazer, K., Pala, K., Christodoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigori-Adou, M., BALKANET: A Multilingual Semantic Network for Balkan Languages, in 1st International Wordnet Conference, Mysore, India, January 2002, http://www.ceid.upatras.gr/Balkanet/files/balkanet-elsnet-ko-accept.pdf

Mariano Sigman and Guillermo A. Cecchi, Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences, 2002.

G.A. Miller, "Wordnet: A Lexical Database," Comm. ACM, vol. 38: no. 11, pp. 39-41, Nov. 1995.

Fellbaum, C. (ed.), Wordnet: An Electronic Lexical Database, The MIT Press, 1998.

Vossen, P. (ed.), EuroWordnet: A Multilingual Database with Lexical Semantic Networks, Dordrecht, Kluwer Academic Publishers, 1998.

Obradović, I., et al.: Application of Intex in Refinement and Validation of Serbian Wordnet. *6th Intex Workshop*, 28.30th May, Sofia (2003).

Krstev, C., Pavlović-Lažetić, G., Obradović, I., Vitas, D., Corpora Issues in Validation of Serbian Wordnet, in Matoušek, V., Mautner, P. (eds.), Text, Speech and Dialogue, LNAI 2807, Springer, 132-137, 2003.

Krstev , C. et al. Cooperative work in further development of Serbian Wordnet, Infotheca, ISSN 1450-9687, No 1-2, Vol IX, May 2008, pp 59a-78a

Sinclair J, Ball J 1996 EAGLES Preliminary Recommendations onText Typology. (http://www.ilc.pi.cnr.it/EAGLES96/texttyp/texttyp.html)

A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators

Tim vor der Brück, Sven Hartrumpf, Hermann Helbig

Intelligent Information and Communication Systems (IICS)
FernUniversität in Hagen
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

Abstract

Checking for readability or simplicity of texts is important for many institutional and individual users. Formulas for approximately measuring text readability have a long tradition. Usually, they exploit surface-oriented indicators like sentence length, word length, word frequency, etc. However, in many cases, this information is not adequate to realistically approximate the cognitive difficulties a person can have to understand a text. Therefore we use deep syntactic and semantic indicators in addition. The syntactic information is represented by a dependency tree, the semantic information by a semantic network. Both representations are automatically generated by a deep syntactico-semantic analysis. A global readability score is determined by applying a nearest neighbor algorithm on 3,000 ratings of 300 test persons. The evaluation showed, that the deep syntactic and semantic indicators lead to quite comparable results to most surface-based indicators. Finally, a graphical user interface has been developed which highlights difficult-to-read text passages, depending on the individual indicator values, and displays a global readability score.

1. Introduction

Readability checkers are used to highlight text passages that are difficult to read. They can help authors to write texts in an easy-to-read style. Furthermore they often display a global readability score which is derived by a readability formula. Such a formula describes the readability of a text numerically. There exists a large amount of readability formulas (Klare, 1963). Most of them use only surface-oriented indicators like word frequency, word length, or sentence length. Such indicators have only indirect and limited access to judging real understandability. Therefore, we use deep syntactic and semantic indicators[1] in addition to surface-oriented indicators. The semantic indicators operate mostly on a semantic network (SN); in contrast, the syntactic indicators mainly work on a dependency tree containing linguistic categories and surface text parts. The SNs and the dependency trees are derived by a deep syntactico-semantic analysis based on word-class functions.

Furthermore, we collected a whole range of readability criteria from almost all linguistic levels: morphology, lexicon, syntax, semantics, and discourse[2] (Hartrumpf et al., 2006). To make these criteria operable, each criterion is underpinned by one or more readability indicators that have been investigated in the (psycho-)linguistic literature and can be automatically determined by NLP tools (see (Jenge et al., 2005) for details). Two typical readability indicators for the syntactic readability criterion of *syntactic ambiguity* are the *center embedding depth of subclauses* and the *number of argument ambiguities* (concerning their syntactic role[3]).

[1] In this paper, an indicator is called *deep* if it depends on a deep syntactico-semantic analysis.

[2] In this paper, discourse criteria are subsumed under the heading semantic because they form only a small group and rely directly on semantic information.

[3] Such ambiguities can occur in German because of its relatively free constituent order.

2. Related Work

There are various methods to derive a numerical representation of text readability. One of the most popular readability formulas was created in 1948: the so-called Flesch Reading Ease (Flesch, 1948). The formula employs the average sentence length and the average number of syllables for judging readability. The sentence length is intended to roughly approximate sentence complexity, while the number of syllables approximates word frequency since usually long words are less used. Later on, this formula was adjusted to German (Amstad, 1978). Despite of its age, the Flesch formula is still widely used.

Also, the revised Dale-Chall readability index (Chall and Dale, 1995) mainly depends on surface-type indicators. Actually, it is based on sentence length and the occurrences of words in a given list of words which are assumed to be difficult to read.

Recently, several more sophisticated approaches which use advanced NLP technology were developed. They determine for instance the embedding depth of clauses, the usage of active/passive voice or text cohesion (McCarthy et al., 2006; Heilman et al., 2007; Segler, 2007). The method of (Chandrasekar and Srinivas, 1996) goes a step beyond pure analysis and also creates suggestions for possible improvements.

Usually, those approaches are based on surface or syntactic structures but not on a truly semantic representation which represents the cognitive difficulties for text understanding more adequately. Moreover, readability checkers normally focus on English texts which means that grammatical phenomena typical for German like separable prefixes are not handled (see Sect. 5.2.). Moreover, only few of those approaches (e.g., (Rascu, 2006)) integrate their readability checkers into a graphical user interface, which is vital for practical usage.

Readability formulas usually combine several so-called readability indicators like sentence or word length by a pa-

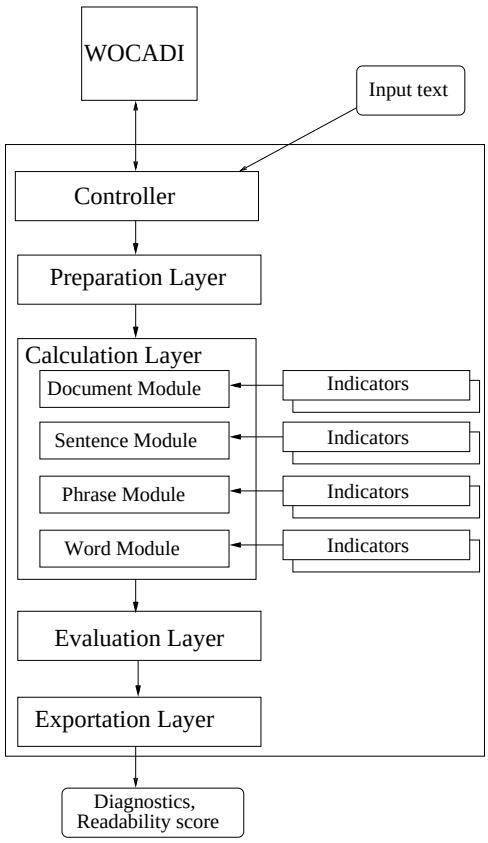


Figure 1: System architecture of the readability checker DeLite.

rameterized sum. Non-linear readability formulas are currently quite rare. Examples of the latter type are the nearest neighbor approach of (Heilman et al., 2007) and the employment of support vector machines by (Larsson, 2006). Larsson used them to separate the vectors of indicator values for given texts into the three different readability classes *easy*, *medium*, and *difficult*. A drawback of this method is that the classification into only three levels is rather rough.

3. System Architecture

A text is processed in several steps (see Figure 1) by our readability checker DeLite (an association of *Lite* as in light/easy reading and *De* as in Deutsch/German; there is also a prototype EnLite for English). First, the Controller passes the text to a deep syntactico-semantic analysis (WOCADI[4] parser, (Hartrumpf, 2003)), which is based on a word-class functional analysis and is supported by a large semantically oriented lexicon (Hartrumpf et al., 2003). The parser output for each sentence is a morpho-lexical analysis, one or more (in case of ambiguities) syntactic dependency trees, one or more SNs, and intrasentential and intersentential coreferences determined by a hybrid rule-statistical coreference resolution module. An example of the resulting SNs, which follow the MultiNet formalism (multilayered extended semantic network, (Helbig, 2006)), is shown in Figure 2. On the basis of this analysis, the text is divided into sentences, phrases, and words in the Preparation Layer.

The individual indicator values are determined by the Calculation Layer. DeLite currently uses 48 morphological, lexical, syntactic, and semantic indicators; in the following sections, we concentrate on some deep syntactic and semantic ones. Each indicator is attached to a certain processing module depending on the type of required information: words, phrases, sentences, or the entire document. Each module iterates over all objects of its associated type that exist in the text and triggers the calculation of the associated indicators. Examples for indicators operating on the word level are the indicators *number of characters* or *number of word readings*. Semantic and syntactic indicators usually operate on the sentence level. As the result of this calculation step an association from text segments to indicator values is established.

In the Evaluation Layer, the values of each indicator are averaged to the so-called *aggregated* indicator value. Note that there exists for each indicator only one aggregated indicator value per text. The readability score is then calculated (see Sect. 4.) by the k-nearest neighbor algorithm of the machine learning toolkit RapidMiner (Mierswa et al., 2006) . In spite of surface-type indicators a deep indicator can usually only be determined for a given sentence (usually, deep indicators operate on sentences) if certain prerequisites are met (e.g., full or chunk parse available). If this is not the case, the associated sentence is omitted for determining the aggregated indicator value. If an indicator could not be calculated for any sentence of the text at all, its value is set to some fixed constant.

Finally, all this information is marked up in XML and in a user-friendly HTML format and is returned to the calling process by the Exportation Layer.

4. Deriving a Readability Score Using the k-Nearest Neighbor Algorithm

A nearest neighbor algorithm is a supervised learning method. Thus, before this method can be applied to new data, a training phase is required. In this phase, a vector of aggregated indicator values is determined by RapidMiner (see previous section) for each text of our readability study. The vector components are normalized and multiplied by weights representing the importance of the individual indicators where the weights are automatically learned by an evolutionary algorithm. All vectors are stored together with the average user ratings for the associated texts.

To derive a readability score for a previously unseen text, the vector of weighted and normalized aggregated indicator values is determined for this text first (see above). Afterwards, the k vectors of the training data with the lowest distance to the former vector are extracted. The readability score is then given as a weighted sum of the user ratings associated with those k vectors (the k nearest neighbors).

5. Syntactic Indicators

5.1. Clause Center Embedding Depth

A sentence is difficult to read if the syntactic structure is very complex (Groeben, 1982). One reason for a high com-

[4]WOCADI is the abbreviation of **Wo**rd-**C**l**a**ss based **Di**sambiguating.

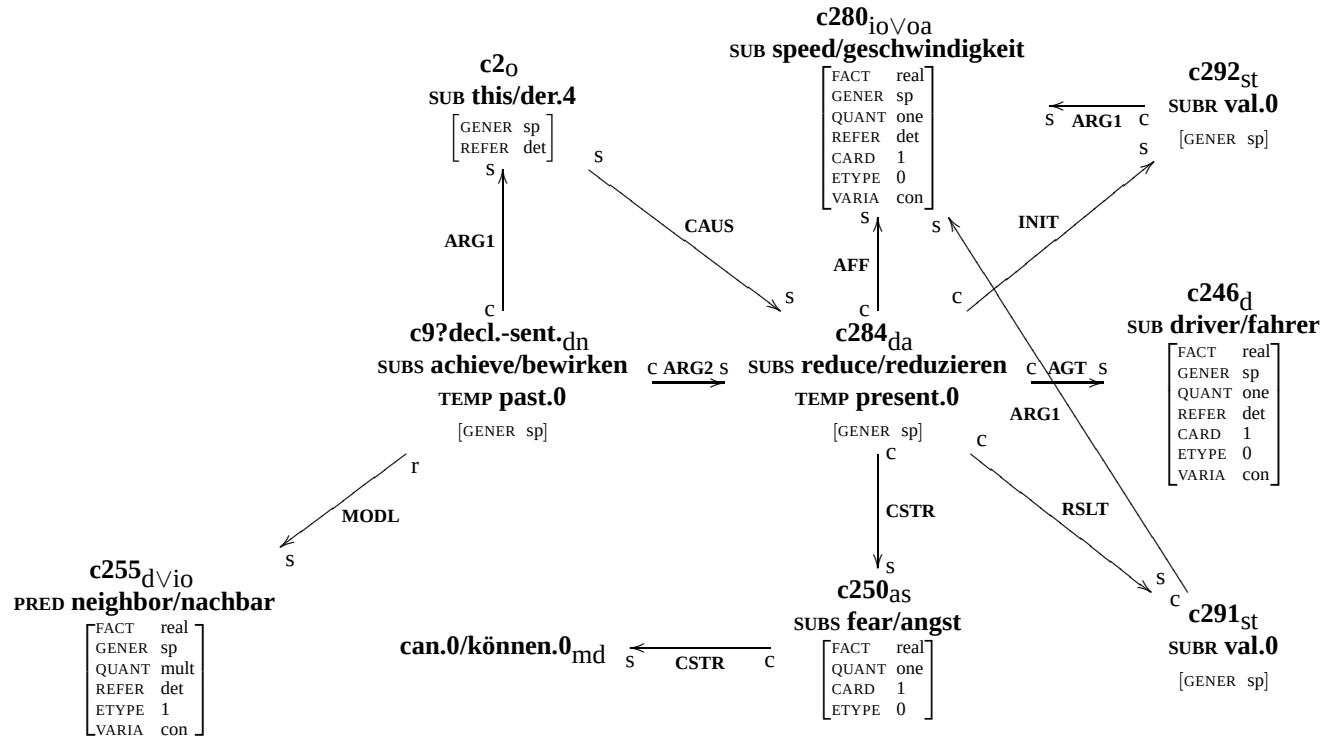


Figure 2: SN for the corpus sentence *Das könnte bewirken, dass der Fahrer aus Angst vor den Nachbarn die Geschwindigkeit reduziert.* (*This could achieve that the driver reduces the speed for fear of the neighbors.*)

plexity can be that the sentence contains deeply embedded subordinate clauses. The difficulty can be increased if the subordinate clause is embedded into the middle of a sentence since the reader has to memorize the superior clause until its continuation after the termination of the subordinate clause, for example: *Er verließ das Haus, in dem die Frau, die er liebte, wohnte, sofort.* (literally: *He left the house where the woman he loved lived immediately.*) Thus, we employ the center embedding depth of a main verb as a readability indicator and calculate its value in the following way. First, we determine the path from the root of the dependency tree to each main verb. Then, we count the occurrences of the dependency relations for relative or other subordinated clauses on this path. However, we only take them into account if the embedded clause is not located on the border of the superior clause which we can verify by comparing the start/end character indices of both clauses.

5.2. Distance between Verb and Separable Prefix

In German, so-called separable prefix verbs are split into two words in clauses with main clause word order. Example: *einladen* (*invite*) ⇒ *Er lädt . . . ein.* (*He invites . . . .*). If the verb is far away from the verb prefix, it can be difficult to associate both parts.

5.3. Number of Words per Nominal Phrase

According to Miller (Miller, 1962), long NPs degrade readability. Hence, a part of the information given in the long NP should better be placed in a subordinate clause or a new sentence. Therefore we count the average number of words contained in an NP where a larger number results in a worse readability score. Note that we only consider maximal NPs (i.e., NPs not contained in a larger NP). Otherwise a large indicator value for the long NP could be compensated by small indicator values for the contained NPs which should be avoided.

6. Semantic Indicators

6.1. SN Quality

The fact that a sentence could not be completely parsed is caused mainly by syntactic or semantic defects since the parser builds the syntactic structure as a dependency tree and the semantic representation as an SN in parallel. Therefore, the indicator *SN quality* is a mixed one: semantic and syntactic. Consider for instance the two sentences *Das Werk kam vor allem bei jungen Theatergängern an. Schulbusse reisten an, um es sich anzusehen.*[5] (*The work was very well accepted by young visitors of the theater. School buses arrived to watch it.*) The second sentence, which is syntactically correct, is semantically incorrect and therefore difficult to read. The semantic lexicon, which is employed by the parser, requires that the first argument (which plays the semantic role of the agent) *ansehen.1.1*[6] (*to watch*) is of type *human*. Thus, this sentence is rejected by the parser as incorrect. In other cases the sentence might be accepted but considered as semantically improbable. This information, which is provided by the parser, is used by the readability checker DeLite and turned out to be very valuable for estimating text readability.

Three parse result types are differentiated: complete parse (around 60% of the sentences; note that this means

[5] from the newspaper *Schleswig-Holstein am Sonntag, 2007*

[6] Note that the readings of a lexeme are distinguished by numerical suffixes.

complete syntactic structure *and* semantic representation at the same time), chunk parse (25%), failure (15%).[7] Those three cases are mapped to different numerical values of the indicator *SN quality*. Additionally, if a full parse or a chunk parse is available, the parser provides a numerical value specifying the likelihood that the sentence is semantically correct which is determined by several heuristics. This information is incorporated into the quality score of this indicator too. Naturally, this indicator depends strongly on the applied parser. A different parser might lead to quite different results.

6.2. Number of Propositions per Sentence

DeLite also looks at the number of propositions in a sentence. More specifically, all SN nodes are counted which have the ontological sort *si*(tuation) (Helbig, 2006, p. 412) or one of its subsorts. In a lot of cases, readability can be judged more accurately by the number of propositions than by sentence length or similar surface-oriented indicators. Consider for instance a sentence containing a long list of NPs: *Mr. Miller, Dr. Peters, Mr. Schmitt, Prof. Kurt, ... were present.* Although this sentence is quite long it is not difficult to understand (Langer et al., 1981). In contrast, short sentences can be dense and contain many propositions, e.g., concisely expressed by adjective or participle clauses.

6.3. Number of Connections between SN Nodes/Discourse Entities

The average number of nodes which are connected to an SN node is determined. A large number of such nodes often indicates a lot of semantic dependencies. For this indicator, the arcs leading to and leaving from an SN node are counted. Note that the evaluation showed that better results (stronger correlation and higher weight) have been achieved if only SN nodes are regarded which are assigned the ontological sort *object* (Helbig, 2006, p. 409–411). In this case, these SN nodes roughly represent the discourse entities of a sentence.

6.4. Length of Causal and Concessive Chains

Argumentation is needed to make many texts readable. But if an author puts too many ideas in too few words, the passage becomes hard to read. For example, the following sentence from a newspaper corpus has been automatically identified as pathologic because it contains three causal relations (CAUS and CSTR in Figure 2) chained together: *Das könnte bewirken, dass der Fahrer aus Angst vor den Nachbarn die Geschwindigkeit reduziert.* (*This could achieve that the driver reduces the speed for fear of the neighbors.*). Again, length measurements on the surface will not help to detect the readability problem, which exists for at least some user groups. Splitting such a sentence into several ones is a way out of too dense argumentation.

[7] Note that the absence of a complete parse is problematic only for a part of the indicators, mainly deep syntactic and semantic ones. And even for some of these indicators, one can define fallback strategies to approximate indicator values by using partial results (chunks).

Indicator	Weight	Type
Number of words per sentence	0.679	Sur
Passive without semantic agent	0.601	Syn
Number of readings	0.520	Sem
Distance between verb and complement	0.518	Syn
SN quality	0.470	Syn/Sem
Number of connections between discourse entities	0.467	Sem
Inverse concept frequency	0.453	Sem
Clause center embedding depth	0.422	Syn
Number of sentence constituents	0.406	Syn
Maximum path length in the SN	0.395	Sem
Number of causal relations in a chain	0.390	Sem
Number of compound simplicia	0.378	Sur
...	...	...
Word form frequency	0.363	Sur
...	...	...
Number of connections between SN nodes	0.326	Sem

Table 1: Indicators with largest weights in our readability function (Syn=syntactic, Sem=semantic, and Sur=surface indicator type).

7. Evaluation

To judge the viability of our approach, we conducted an online readability study with 500 texts, more than 300 participants, and around 3,000 human ratings for individual texts where the participants rated the text readability on a 7 point Likert scale (Likert, 1932).

Almost 70 % of the participants were between 20 and 40 years old; the number of participants over 60 was very small (3 %). The participants were mainly well-educated. 58 % of them owned a university or college degree. There is none who had no school graduation at all.

Our text corpus originated from the municipal domain and differs significantly from newspaper corpora, which are widely used in computational linguistics. So the text corpus we used contains a lot of ordinances with legal terms and abbreviations, e.g., § *65 Abs. 1 Satz 1 Nr. 2 i.V.m.* § *64 Abs. 1 Satz 2 LWG NRW* (*section 65.1.1 (2) in connection with section 64.1.2 LWG NRW*). This corpus has been chosen because local administrations in Germany have committed themselves to make their web sites accessible; one central aspect of accessibility is simple language.

Figure 4 shows the mean average error (MAE) and the root mean square error (RMSE) of DeLite's global readability score in contrast to the average user rating determined by a 10 fold cross-validation over all 500 test documents. The ordinate contains MAE and RMSE, the abscissa, on a logarithmic scale, the number of neighbors used. The lowest errors (MAE: 0.126, RMSE: 0.153) were obtained when using the 40 nearest neighbors. The nearest neighbor algorithm determined the weights of each indicator using an evolutionary algorithm. The resulting indicator weights are given in Table 1.

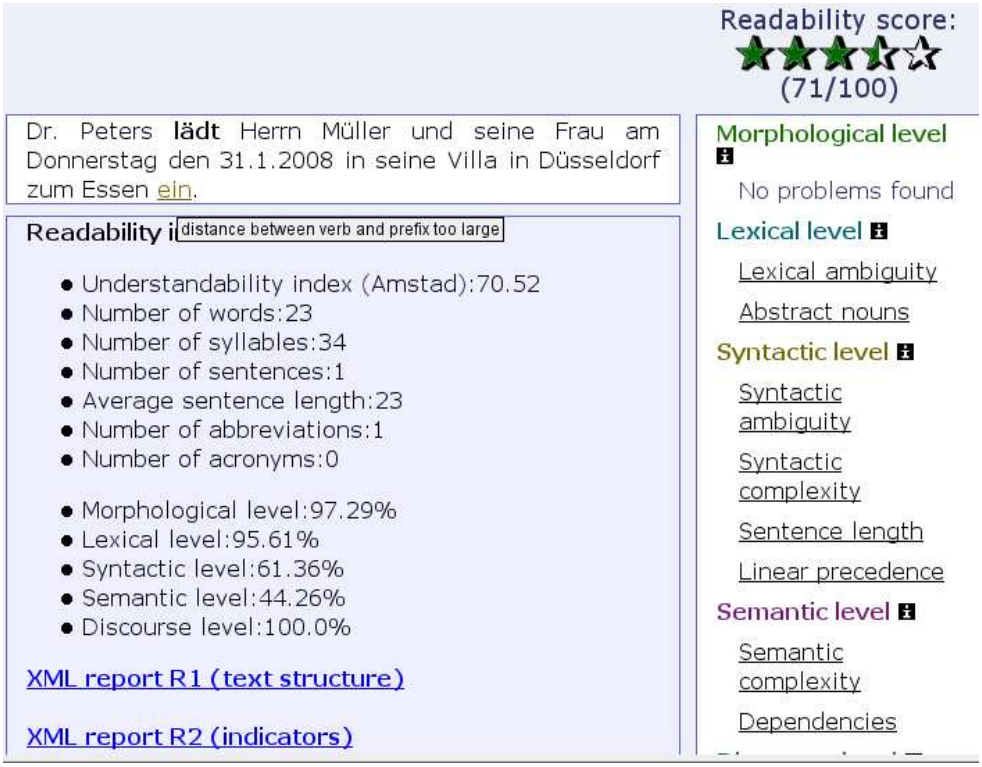


Figure 3: DeLite screenshot showing a sentence which contains a large distance between verb (*lädt*) and separable verb prefix (*ein*). English translation for the example sentence: *Dr. Peters invites Mr. Müller and his wife for dinner on Thursday, Jan. 31, 2006 to his villa in Düsseldorf.*

Indicator	Correlation	Type
Number of words per sentence	0.430	Sur
SN quality	0.399	Syn/Sem
Inverse concept frequency	0.330	Sem
Word form frequency	0.262	Sur
Number of reference candidates for a pronoun	0.209	Sem
Number of propositions per sentence	0.180	Sem
Clause center embedding depth	0.157	Syn
Passive without semantic agent	0.155	Syn
Number of SN nodes	0.148	Sem
Pronoun without antecedent	0.140	Sem
Number of causal relations in a chain	0.139	Sem
Distance between pronoun and antecedent	0.138	Sem
Maximum path length in the SN	0.132	Sem
Number of connections between discourse entities	0.132	Sem

Table 2: Indicators most strongly correlated with user ratings (Syn=syntactic, Sem=semantic, and Sur=surface indicator type).

The correlations of the indicators in comparison with the user ratings are displayed in Table 2. Correlation and weights of deep syntactic and semantic indicators turned out to be quite comparable to surface-type indicators.

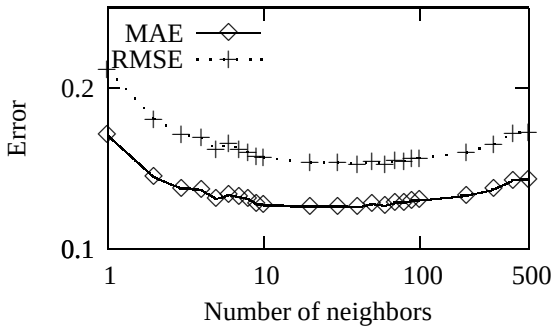


Figure 4: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between the DeLite readability score and the average user ratings of a text depending on the number of neighbors.

Finally as a baseline, DeLite was compared to the readability index resulting from employing the nearest neighbor approach only on the indicators of the Flesch readability index, i.e. average sentence length and number of syllables per word. The correlation of DeLite with the user ratings is 0.501 which clearly outperforms the Flesch indicators (0.432).

8. User Interface

Besides a low-level server interface, DeLite provides a graphical user interface for comfortable usage. In Fig-

ure 3, a screenshot of this interface is shown.[8] The types of readability problems found in the text are displayed on the right side. If the user clicks on such a type, the associated difficult-to-read text segments are highlighted. Additional support for the user is provided if he/she wants to have more information about the readability problem. Moving the mouse pointer over the highlighted text segment, a fly-over help text with a more detailed description is displayed. Moreover, if the user clicks on the highlighted segment, additional text segments are marked in bold face. These additional segments are needed to fully describe and explain specific readability problems.

The example in Figure 3 shows the readability analysis of a verb which is too far away from its separable prefix (see Sect. 5.2.). The prefix *ein-* is highlighted as problematic and additionally the main verb *lädt* is marked in bold face for better understanding.

9. Conclusion

An overview of some typical examples of deep syntactic and semantic readability indicators has been given. In our evaluation, it turned out that these indicators have comparable weights and correlations to most surface-type indicators in accurately judging readability.

In the future, the parser employed in DeLite will be continually improved. Currently, DeLite is only an authoring tool; we will investigate the addition of the ability to reformulate a sentence to be better to understand. Finally, the automatic distinction between real ambiguities that exist for humans and spurious ambiguities that exist only for machines (e.g., NLP methods like PP attachment and interpretation) must be sharpened.

Deep syntactic and semantic indicators turned out to be quite valuable for assessing readability and are expected to be a vital part of future readability checkers.

Acknowledgments

We wish to thank our colleagues Christian Eichhorn, Ingo Glöckner, Johannes Leveling, and Rainer Osswald for their support for this work. The research reported here was in part funded by the EU project Benchmarking Tools and Methods for the Web (BenToWeb, FP6-004275).

10. References

T. Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich, Zurich, Switzerland.

J. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Brookline, Massachusetts.

R. Chandrasekar and B. Srinivas. 1996. Automatic induction of rules for text simplification. Technical Report IRCS Report 96-30, University of Pennsylvania, Philadelphia, Pennsylvania.

R. Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

N. Groeben. 1982. *Leserpsychologie: Textverständnis – Textverständlichkeit*. Aschendorff, Münster, Germany.

S. Hartrumpf, H. Helbig, and R. Osswald. 2003. The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

S. Hartrumpf, H. Helbig, J. Leveling, and R. Osswald. 2006. An architecture for controlling simple language in web pages. *eMinds: International Journal on Human-Computer Interaction*, 1(2):93–112.

S. Hartrumpf. 2003. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany.

M. J. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Human Language Technology Conference*, Rochester, New York.

H. Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin.

C. Jenge, S. Hartrumpf, H. Helbig, G. Nordbrock, and H. Gappa. 2005. Description of syntactic-semantic phenomena which can be automatically controlled by NLP techniques if set as criteria by certain guidelines. EU-Deliverable 6.1, FernUniversität in Hagen.

G. Klare. 1963. *The Measurement of Readability*. Iowa State University Press, Ames, Iowa.

I. Langer, F. Schulz von Thun, and R. Tausch. 1981. *Sich verständlich ausdrücken*. Reinhardt, München, Germany.

P. Larsson. 2006. Classification into readability levels. Master's thesis, Department of Linguistics and Philology, University Uppsala, Uppsala, Sweden.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

P. McCarthy, E. Lightman, D. Dufty, and D. McNamara. 2006. Using Coh-Metrix to assess distributions of cohesion and difficulty: An investigation of the structure of high-school textbooks. In *Proc. of the Annual Meeting of the Cognitive Science Society*, Vancouver, Canada.

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. 2006. Yale: Rapid prototyping for complex data mining tasks. In *Proc. of KDD*, Philadelphia, Pennsylvania.

G. Miller. 1962. Some psychological studies of grammar. *American Psychologist*, 17:748–762.

E. Rascu. 2006. A controlled language approach to text optimization in technical documentation. In *Proc. of KONVENS 2006*, pages 107–114, Konstanz, Germany.

T. M. Segler. 2007. *Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German*. Ph.D. thesis, School of Informatics, University of Edinburgh.

[8] Note that the classification of indicators is slightly different in the screenshot than in this paper. This is caused by the fact that we want to evaluate surface-oriented indicators in comparison to linguistically informed indicators.

Rapid development of data for shallow transfer RBMT translation systems for highly inflective languages

Jernej Vičič

University of Primorska
Glagoljaška 8, SI-6000 Koper
jernej.vicic@upr.si.si

Abstract

The article describes a new way of constructing rule-based machine translation systems (RBMT), in particular shallow-transfer RBMT suited for related languages. The article describes methods that automate parts of the construction process. The methods were evaluated on a case study: the construction of a fully functional machine translation system of closely related language pair Slovenian - Serbian. The Slovenian language and The Serbian language belong to the group of southern Slavic languages that were spoken mostly in the former Yugoslavia. The economies of the nations where these languages are spoken are closely connected and younger generations, the post-Yugoslavia breakage generations, have difficulties in mutual communication, so there is a big interest in construction of such translation system. The system is based on Apertium (Oller and Forcada, 2006), an open-source shallow-transfer RBMT toolkit. Thorough evaluation of the translation system is presented and conclusions present the strong and the weak points of this approach and explore the grounds for further work.

1. Introduction

Slovenian language and Serbian language belong to the group of southern Slavic languages that are spoken mostly on the territory of former Yugoslavia. Slovenian language is mostly spoken in Slovenia, Serbian language is mostly spoken in Serbia. The languages share common roots and even more importantly they share common recent historical environment, these languages were spoken in the same country, even taught in schools as languages of the surroundings. Economies of both countries are closely connected. Younger generations, the post-Yugoslavia breakage generations, have difficulties in mutual communication, so there is a big interest in the construction of an automatic machine translation system for this language pair. Both languages are highly inflective and morphologically and derivationally rich languages and differ greatly from mostly used languages in electronic materials like English, Arabic, Chinese, Spanish and French. This means that most of the data and translation methods must be at least revisited or even worse rewritten. This language pair is closely related lexicographically and syntactically which simplifies most of the translation system production steps. All methods and materials discussed in this paper were tested on a fully functional machine translation system based on Apertium (Oller and Forcada, 2006; Corbi-Bellot et al., 2005), an open-source shallow-transfer RBMT toolkit. Apertium is an open-source machine translation platform, initially aimed at related-language pairs but recently expanded to deal with more divergent language pairs (such as English-Catalan). The platform provides a language-independent machine translation engine, tools to manage the linguistic data necessary to build a machine translation system for a given language pair and linguistic data for a growing number of language pairs. All these properties make Apertium a perfect choice in a cost effective machine translation system development.

The construction of a machine translation system for a new language pair falls roughly into two categories:

- A long and not particularly interesting job of manual dictionary and rule construction in case of classic Rule-Based Machine Translation (RBMT) (Hutchins, 2005) system construction approach, including all similar approaches.

- Automatic machine translation system construction in case of corpus-based machine construction systems such as Statistical Machine Translation (SMT) (Brown et al., 1993; Och and Ney, 2003) or Example-Based Machine Translation (EBMT) (Nagao, 1984) and (Hutchins, 2005). Several other examples of corpus-based machine translation systems are available.

The SMT seems like a perfect choice as some of the best performing machine translation systems are based on the SMT technologies (NIST, 2006), but it has a few drawbacks that cannot be ignored; the SMT systems, to be efficient, require huge amount of parallel text (Och, 2006) that is available only for a few of the widely used languages like English, Spanish, French, Arabic, etc. The morphologically rich and highly inflective languages like the pair presented in this paper (Slovenian and Serbian language) present an even bigger problem as shown on table 1 3. in section 3.1.. The rest of the article is organized as follows: state of the art is presented in section 2., follows a presentation of the used methods in section 3.. The evaluation methodology with results is presented in section 4., the article concludes with the discussion.

2. State of the art

According to (NIST, 2006), today's best performing automatically constructed machine translation systems , we

will concentrate on SMT systems like (Google, 2008), require huge amounts of parallel data to learn from. Systems for language pairs with big parallel corpora yield good results, for some language pairs even best results overall. Such big corpora are not available for most of the languages. The smaller corpora, even linguistically annotated, are easier to be found, at least for most of the European languages, see (Dimitrova et al., 1998; Multext, 2007).

2.1. Available technologies and materials

A research of already available and accessible language processing tools and materials, mostly corpora, revealed that there is a reasonably big amount of work already done for the Slovenian language, less for the Serbian language. The tools for the Slovenian language are (reasonable or even good quality): a part of speech tagger (Erjavec, 2006; Brants, 2000), a lemmatizer (Erjavec, 2006; Erjavec, 2004), a stemmer (Popovic and Willett, 1992; Popovic and Willett, 2000). None of these tools exists for Serbian language. Both languages have solid monolingual reference corpora (going into hundreds of millions) and a small bilingual corpus (Dimitrova et al., 1998).
This research focuses mostly on the lexical level mainly for these reasons:

- The lexical level presents the starting ground for written text translation.
- The related languages, particularly the language pair we based our study upon, usually share the same sentence structure.
- Most of the translation takes place on the lexical level.
- Unlike some well-known languages, like English, southern Slavic languages express most of the meaning by inflecting words and less by word order.

Only the lexicographic modules were taken into consideration in this case study as the work on the project is still in progress. We concentrated the research on preceding modules, the lexicographic modules, as they present the basis for all translation stages. Still some basic structural transfer rules were constructed to greatly enhance translation performance at a small cost in expert hours.

3. Intention

Quite a few methods that automate some parts of the RBMT machine translation system construction have been presented and are even used as part of the construction toolkits. This article presents an attempt to automate all data creation processes of a shallow transfer machine translation system based on RBMT. The Apertium (Oller and Forcada, 2006) shallow transfer machine translation toolbox was used in our experiments although most of the methods could be applied to other systems. The data:

1. The monolingual source dictionary with morphological information for source language parsing.
2. the monolingual target dictionary with morphological information for target language generation.
3. The bilingual translation dictionary.
4. The shallow transfer rules.
5. The disambiguation data.

The monolingual dictionaries are used in shallow parsing of the source text and generation of the translation text in the target language. The bilingual dictionary is used for word-by-word translation, in our case the translation is based on lemmata. The shallow transfer rules are used to address local syntactical and morphological rules such as local word agreement and local word reordering. The morphological disambiguation of the source language morphological parsing phase was done using implicit disambiguation rules, in our case in form of the Hidden Markov Model (HMM) parameters (stochastic POS tagger), but other alternatives are possible such as methods described in (Homola and Kubon, 2008). Each item from the list was addressed by applying a known method or by introducing a new method. The methods are further presented in a separate subsection. A fully functional system was constructed using presented methods and overall performance of the whole system was evaluated.

3.1. Monolingual source and target dictionary creation

Let us look at an example from the English language; the transformation of the word walk into walked can be achieved by a morphological transformation rule (for past tense). A variation of the same rule would be used for the irregular word sleep, changing into slept. For languages that employ concatenative morphology [1] such as the majority of European languages, different forms of the same word are realized by changing the prefix and suffix of the word. Thus, slept can be derived from sleep by changing the suffix -ep to the suffix -pt. The same phenomenon, but to a much greater extent, occurs in highly inflectional languages, an example for Slovenian language is shown in table 1.

3.1.1. Paradigm creation

The words were grouped into paradigms in order to deal with multiple word-forms as Slovenian and Serbian language are both highly inflectional languages. Each paradigm is represented by:

- a typical lemma, the lemma the paradigm was constructed from
- a stem, the longest common prefix of all words in the lemma
- a set of all words split into stems and postfixes and Morpho-Syntactical Descriptors (MSDs) (Erjavec, 2004)

The annotated lexicons, lists of unique words with lemma descriptor and MSD, were extracted from corpus for both languages and paradigms were constructed using the algorithm in figure 1.

[1] words are composed of a number of morphemes concatenated together; the morphemes include the stem plus prefixes and suffixes

Table 1: All word forms for Slovenian lemma mesto (place/city)

word form	number	case
mest-o	Singular	nominative
mest-a	Singular	genitive
mest-u	Singular	dative
mest-o	Singular	accusative
mest-u	Singular	locative
mest-om	Singular	instrumental
mest-a	Plural	nominative
mest-∅	Plural	genitive
mest-om	Plural	dative
mest-a	Plural	accusative
mest-ih	Plural	locative
mest-i	Plural	instrumental
mest-i	Dual	nominative
mest-∅	Dual	genitive
mest-oma	Dual	dative
mest-i	Dual	accusative
mest-ih	Dual	locative
mest-oma	Dual	instrumental

```
//par - paradigms
for(i = 0; i < par.size; i++){
  for(j = i; j < par.size; j++){
    if(par[i].POS == par[j].POS){
      if(all entries agree){
        join(par[i], par[j])
      }
    }
  }
}
```

Figure 1: Paradigm construction algorithm

All word forms of a lemma present in the corpus are grouped into a class representing the lemma. A paradigm is constructed from each class; for each lemma. Two paradigms are joined together if lemmata of both paradigms, in the first step just two lemmata, later the number increases, have the same POS and if all entries agree: entries with same MSD have same postfix. Sets of entries of both paradigms are joined into a new set. The information about all lemmata that generated the paradigm is stored in a list enabling easy lookup. The monolingual source and target dictionaries were constructed using joined paradigms resulting in a roughly 20 times bigger lexicon than the starting.

3.2. Bilingual translation dictionary creation

The Number of word forms in a text is much bigger for highly inflective languages like the Slavic languages. Figure 4 shows the difference in number of word forms for the same corpus (Dimitrova et al., 1998) in four languages; three highly inflective Slavic languages: Slovenian, Serbian, Czech and English language as a reference.

The reduction of search space obviously increases the accuracy of the model (the word-by-word translation model). This result is not surprising, but a lot of information about the word form was lost in the process. Let us observe the phenomenon to a greater extent. The word alignment model as described in (Brown et al., 1993; Och and Ney, 2003) can be used as the basis for a new model that uses lemma+POS descriptions of the actual word forms used in the bilingual parallel corpus.

Table 2: Number of lemmata in corpus MULTEXT-EAST (Dimitrova et al., 1998)

language	number of words	lemmata
Slovenian	22134	6512
Serbian	21435	6832
Czech	23654	7263
English	11293	8182

Some simple definitions that will help the formulation of the equation 1
L - language, all words
E_L - lemmata of the language L
$E_{L(i)}$ - i^{th} lemma with all word forms

$$|L| = \sum_{i=0}^{|E_L|} E_{L(i)} \tag{1}$$

The search space is reduced from $|L|$ to $|E_L|$.
Observe the example:
Assuming that George Orwell's novel "1984", which comprises the multilingual sentence-aligned part of the (Dimitrova et al., 1998) corpus, is a good sample of a language, in our case the Slovenian language, we observe the values in figure 2 taken from table 2. The search space has been reduced from 22134 word forms to 6512 lemmata.

Original language $|L| = 22134$
Lematized language $|E_L| = 6512$

Figure 2: The reduction of the search space for the Slovenian language (small corpus MULTEXT-EAST (Dimitrova et al., 1998))

The bilingual parallel annotated corpus (Dimitrova et al., 1998) comprises original text with additional information in form of XML tags according to the TEI-P4 (Consortium, 2007) and the EAGLES (Leech and Wilson, 1996) guidelines. An example excerpt is shown on figure 3.

Each word is represented by the *lemma* (lemma of the word), *ana* (morphosyntactical description - MSD (Erjavec, 2004)) and the word form used in corpus. Only the lemma and the POS, first feature of MSD, of each word were extracted from the corpus for this task, leaving parallel sentences in lemmatised form with the POS tag. Figure 4 shows the prepared data.

An SMT word-to-word model (Brown et al., 1993; Och and Ney, 2003) was trained on the parallel, sentence aligned list extracted from the corpus, shown on figure 4. The lemmata alignment ensures much better alignment performance due to the search space reduction as described in equation 1 and in figure 2. The words from the monolingual dictionaries are aligned to the translations (bilingual lemmata pairs) through paradigms that retain the information about the included lemmata, see section 3.1.1..

```
<s id="Osl.2.3.5.11">
<w lemma="priti" ana="Vmps-dma">Prisla</w>
<w lemma="biti" ana="Vcip3d--n">sta</w>
<w lemma="do" ana="Spsg">do</w>
<w lemma="podrt" ana="Afpnsg">podrtega</w>
<w lemma="drevo" ana="Ncnsg">drevesa</w>
<c>,</c>
<w lemma="o" ana="Spsl">o</w>
<w lemma="kateri" ana="Pr-nsl----a">
katerem</w>
<w lemma="on" ana="Pp3msd--y-n">mu</w>
<w lemma="biti" ana="Vcip3s--n">je</w>
<w lemma="praviti" ana="Vmps-sfa">
pravila</w>
<c>.</c>
</s>
```

Figure 3: A sentence in the corpus

```
priti_V biti_V do_S podrt_A
drevo_N , o_S kateri_P on_P
biti_V praviti_V .
```

Figure 4: Prepared data: lemmata and POS of each word from the corpus

3.3. Transfer rules induction

This experiment focused on morphologically annotated data. The creation of shallow-transfer translation rules has been systematically avoided. A few test rules have been manually created to observe the translation quality performance boost. Shallow transfer translation rules will be automatically constructed using already available software (Sanchez-Martinez and Forcada, 2007) and automatically ordered according to (Vicic and Forcada, 2008).

3.4. Implicit disambiguation rules training

The POS tagger has been used to disambiguate source language parsing options. Two POS taggers were tested: the TnT (Brants, 2000) from TOTALE (Erjavec, 2006) toolkit and the Apertium POS tagger (Sanchez-Martinez et al., 2007). The first was already trained on the same corpus while the second was trained in an unsupervised method on an automatically harvested text from the internet. As expected, better results were achieved by (Brants, 2000) due to better training data. The experiment was not conducted thoroughly due to lack of time and due to satisfactory results achieved by the available tools.

4. Evaluation methodology and results

The evaluation of the translations was performed in four parts, each part is further described in a separate subsection in the continuation of this chapter:

1. The automatic objective evaluation using BLEU (Papineni et al., 2001) metric.

2. The automatic objective evaluation using METEOR (Banerjee and Lavie, 2005; Lavie and Agarwal, 2007) metric.

3. The non-automatic evaluation by counting the number of edits needed to produce a correct target sentence from automatically translated sentence.

4. Non-automatic subjective evaluation following (LDC, 2005) guidelines.

Subjective evaluation was performed after first poor BLEU results triggered some distrust. Many authors agree that BLEU metric systematically penalizes RBMT systems (Callison-Burch et al., 2006) and it is not suited for highly inflective languages. Authors of METEOR (Banerjee and Lavie, 2005), (Lavie and Agarwal, 2007) state that their system fixes most of the problems encountered using BLEU metric; they state that METEOR correlates highly with human judgement. Unfortunately METEOR did not support our language pair, additional software had to be written. The bilingual parallel corpus (Dimitrova et al., 1998) was used in automatic evaluation of translations. The K-fold cross-validation (Kohavi, 1995) was used as the method for estimating the generalization error as it is most suitable for small data sets. In our case five-fold cross validation was used instead of more frequently used ten-fold cross validation as construction of a fully functional system was not automated. The corpus was divided into five parts, each part consisting of roughly 1700 sentences. The evaluation consisted in selecting one part of the corpus as testing set and remaining four parts as training set. The translation system was constructed according to the methodology presented in 3. using the selected training set. The evaluated values in each fold and the average final values are presented.

4.1. Automatic objective evaluation using BLEU metric

The publicly available implementation of the BLEU metric (NIST, 2008) version v11b was used. Results are presented in table 3. These scores are relatively low, especially considering the relatedness of the language pair. Low values are partly to be attributed to high inflexibility of the language pair and partly to the fact that the BLEU metric penalizes RBMT systems (Callison-Burch et al., 2006).

Table 3: The BLEU metric scores, each fold is presented in a separate line, last two lines present average values with standard deviation

fold	BLEU value
1	0.1167
2	0.1211
3	0.1206
4	0.1198
5	0.1201
Average	0.1196
STDEV	0.0017

4.2. Automatic objective evaluation using METEOR metric

The publicly available implementation of the METEOR metric (Lavie and Agarwal, 2007) version v0.6 was used.

The METEOR uses stemming mechanism as one of the algorithms that enhance correlation between METEOR metric and human evaluation for highly inflectional languages. The stemming mechanism that is a side-product of the described translation system was used. Results are presented in table 4.

Table 4: The METEOR metric scores, each fold is presented in a separate line, last two lines present average values with standard deviation

fold	METEOR value
1	0.6344
2	0.6296
3	0.6316
4	0.6297
5	0.6352
Average	0.6321
STDEV	0.0026

4.3. Non-automatic evaluation using edit distance

The edit-distance (Levenshtein, 1965) was used to count the number of edits needed to produce a correct target sentence from automatically translated sentence. This procedure shows how much work has to be done to produce a good translation. The metric roughly reflects the complexity of post-editing task. The evaluation comprised of selecting 100 sentences from testing data, translating these sentences using the translation system and manually counting the number of words that had to be changed in order to obtain a perfect translation. By perfect translation we mean a translation that is syntactically correct and expresses the same meaning as the source sentence. 22% of all words had to be corrected in order to achieve the perfect translation. The results of this evaluation can be compared to results of the same metric used on a similar system; (Homola and Kubon, 2008). Language pair's properties and similarities of our system in comparison to (Homola and Kubon, 2008) make the comparison feasible. The (Homola and Kubon, 2008) system presents one of the best performing related language translation systems with only 3.55 % of errors and therefore presents a good reference point for our system's final goal. This evaluation was conducted as a test on a low number of test translations due to time limitations.

4.4. Non-automatic subjective evaluation following (LDC, 2005) guidelines

Subjective manual evaluation of translation quality was performed according to the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium guidelines. The most widely used methodology when manually evaluating MT is to assign values from two five-point scales representing fluency and adequacy. These scales were developed for the annual NIST Machine Translation Evaluation Workshop by the Linguistic Data Consortium (LDC, 2005).

The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a hypothesis translation:

- 5 = All
- 4 = Most
- 3 = Much
- 2 = Little
- 1 = None

The second five-point scale indicates how fluent the translation is. It expresses weather the translation is syntactically correct. When translating into Serbian the values correspond to:

- 5 = Flawless translation
- 4 = Good Serbian
- 3 = Non-native Serbian
- 2 = Disfluent Serbian
- 1 = Incomprehensible text

Separate scales for fluency and adequacy were developed under the assumption that a translation might be disfluent but contain all the information from the source. Four independent evaluators (two native speakers) evaluated sets of 100 sentences using this methodology. The results are presented in 5.

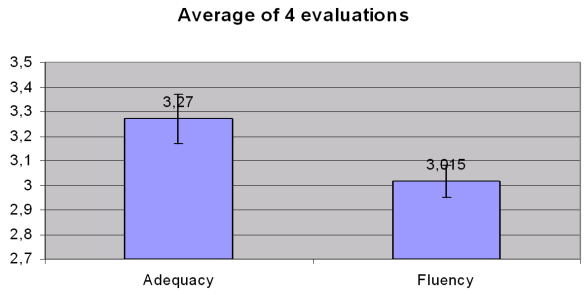


Figure 5: Evaluation results using (LDC, 2005) guidelines. Average values of four independent evaluations show high scores for adequacy and lower values for fluency.

5. discussion

The article presents an ongoing research of rapid construction of shallow-transfer machine translation systems for related languages. The evaluation shows promising results although there is still a lot of space for improvement. All described methods were tested on a fully-functional translation system, the latest version of the system is available online at the following address: http://jt.upr.si/guat/index.php.

The automatic construction of shallow-transfer translation rules has not been addressed in this research and will, in addition to automatic ordering of the rules, present the next step of the research.

6. References

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*. Ann Arbor, Michigan.

Thorsten Brants. 2000. Tnt – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference*. Seattle, WA.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational linguistics*, 19(2):163–311.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of EACL*.

TEI Consortium. 2007. Tei p5: Guidelines for electronic text encoding and interchange. Technical report, TEI consortium.

Antonio M. Corbi-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Prez-Ortiz, Gemma Ramirez-Snchez, Felipe Sanchez-Martnez, Inaki Alegria, Aingeru Mayor, and Kepa Sarasola. 2005. An open-source shallow-transfer machine translation engine for the romance languages of spain. In *Proceedings of the Tenth Conference of the European Association for Machine Translation*, pages 79–86, May.

Ludmila Dimitrova, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaalep, and Dan Tufis. 1998. Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *COLING-ACL*, pages 315–319.

Tomaz Erjavec. 2004. Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. In *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04*.

Tomaz Erjavec. 2006. Multilingual tokenisation, tagging, and lemmatisation with totale. In *Proceedings of the 9th INTEX/NOOJ Conference*.

Google. 2008. Google translator.

Petr Homola and Vladislav Kubon. 2008. A method of hybrid mt for related languages. In *Intelligent information systems XVI : proceedings of the International IIS '08 conference*, pages 269–278.

John Hutchins. 2005. Towards a definition of example-based machine translation. In *MT Summit X, Proceedings of Workshop on Example-Based Machine Translation*, pages 63–70.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 1137–1143.

A. Lavie and A. Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report.

GN Leech and A. Wilson. 1996. Eagles recommendations for the morphosyntactic annotation of corpora. Technical report, ILC-CNR, Pisa.

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

Multext. 2007. The multext project.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and Human Intelligence*.

NIST. 2006. Nist 2006 machine translation evaluation official results.

NIST. 2008. Evaluation software.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29:19–51.

Franz Josef Och. 2006. Challenges in machine translation. In *Proceedings of ISCSLP*.

Carme Armentano Oller and Mikel L. Forcada. 2006. Open-source machine translation between small languages: Catalan and aranese occitan. In *Strategies for developing machine translation for minority languages (5th SALTMIL workshop on Minority Languages)*, pages 51–54.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical report.

M. Popovic and P. Willett. 1992. The effectiveness of stemming for natural language access to slovene textual data. *Journal of the American Society for Information Science*, 43(5).

M. Popovic and P. Willett. 2000. Krnjenje kot osnova nekaterih nekonvencionalnih metod poizvedovanja. *Knjinica, Ljubljana*, 44.

Felipe Sanchez-Martinez and Mikel L. Forcada. 2007. Automatic induction of shallow-transfer rules for open-source machine translation. In Andy Way and Barbara Gawronska, editors, *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, volume 2007:1, pages 181–190. Skovde University Studies in Informatics, September.

Felipe Sanchez-Martinez, Carme Armentano-Oller, Juan Antonio Perez-Ortiz, and Mikel L. Forcada. 2007. Training part-of-speech taggers to build machine translation systems for less-resourced language pairs. In Vctor J. Daz Madrigal and Fernando Enrquez de Salamanca Ros, editors, *Procesamiento del Lenguaje Natural (XXIII Congreso de la Sociedad Espanola de Procesamiento del Lenguaje Natural)*, volume 39, pages 257–264, September.

Jernej Vicic and Mikel Forcada. 2008. Comparing greedy and optimal coverage strategies for shallow-transfer machine translation. In *Proceedings of Intelligent information systems*.

Part-of-Speech Tagging of Slovenian, 12 years after

Primož Jakopin*, Aleksandra Bizjak Končar*

* Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Novi trg 4, 1000 Ljubljana, Slovenia
primoz.jakopin@guest.arnes.si, aleksa@zrc-sazu.si

Abstract

The paper begins with a brief overview of the efforts and accomplishments in the field of part-of-speech tagging of Slovenian texts. Quite a few research institutions have participated, and the most prominent Slovenian language technology enterprise. An overview of the POS-tagged 1.3 mil. word text corpus at the Fran Ramovš Institute of the Slovenian Language ZRC SAZU follows. The tags of the machine-tagged texts have been verified by linguists and serve as a resources for the POS-tagging of the 240 mil. *Nova beseda* corpus.

1. Introduction

The term of the paper topic is known in a quite a few incarnations, here they are given by descending order of their quoted Google frequencies (June 18, 2008): *part-of-speech tagging* (207.000), *POS tagging* (107.000), *grammatical tagging* (6.770), *morphosyntactic tagging* (1.880) and *word-class syntactic tagging* (8). It is a procedure during which every word of text is assigned an unambiguous, one-and-only morphosyntactic tag out of a set of all possible tags, associated to a particular word - or wordform, to be more precise. An example is the popular sentence *Danes je lepo vreme* [*Today the weather is fine*], where the following tags could be linked to words: *Danes*{A} *je*{GPce, GOce, Gce, ZOcže2} *lepo*{A, Pse1, Pse4, Pže4, Pže6} *vreme*{Sse1, Sse4}. Here the tag values, proposed by the authors in 1996 (Jakopin and Bizjak, 1997, also see http://bos.zrc-sazu.si/cgi/pos_tagging.html) have been used, with the following interpretations: A = adverb, S = noun, G = main verb, GO = the verb to be, GP = auxiliary verb to be in present, ZO = personal pronoun, c = third person, e = singular, ž = female gender, s = neuter gender, 1 = nominative case, 2 = genitive case, 4 = accusative case, 6 = instrumental case. There are 1 x 5 x 5 x 2 = 50 different ways to associate a single tag to every word and the correct one is *Danes*{A} *je*{GPce} *lepo*{Pse1} *vreme*{Sse1}.

2. Flow of events

Part-of-speech tagging first emerged for the English language in the 1980s; the language is morphologically not so very rich which reflects in a very manageable tagset of around 50 tokens. Languages with higher degree of inflection, such as Slovenian, require much larger tagsets for proper morphological description.

2.1. Early developments

The Slovenian story started in 1996 (Jakopin and Bizjak, 1997) with the development of a stochactic tagger at the Fran Ramovš Institute of the Slovenian Language (ISJ) and a tagset of around 5.000 different tokens, and at the Dept. of Knowledge Technologies of the Jozef Stefan Institute in the course of the Multext-East project (Erjavec and Ide, 1998, Džeroski et al., 2000), which aimed at establishing language resource, based on the annotated novel 1984 (0.1 mil. words), by George Orwell, for the 7 languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, Slovenian, and English (around 2.000 different tags for Slovenian). Later on the third institution joined the effort, namely Centre of Language Technology, Faculty of Electrical Engineering and Computer Science at the University of Maribor (Verdonik et al., 2002).

The purpose was practical in all cases: to build a tool, useful for research and for monolingual dictionary development, to foster research of language technologies and to build a resource, useful in speech recognition of Slovenian. Levels of tagging accuracy rose from 80-85% to 90-95% in more specific environments. Resources and tools developed during these early POS-tagging efforts were applied in different fields such as enhancement of entropy calculation (Jakopin, 2002), genre classification (Bizjak, 2005), analysis of the 16th century grammar (Bizjak, 2008), streamlining the annotation of the large text corpus FIDA, prepared by a four partner industry-university consortium and in different speech technology projects (Verdonik and Rojc, 2004).

2.2. Nova beseda and FIDA

Two sizeable text corpora of Slovenian, each targeted at a different audience, emerged around 2000, the monitor corpus *Nova beseda* (Jakopin and Michelizza, 2007), currently comprising 242 mil. words, and the reference corpus *FIDA*, 100 mil. words till 2007 and 623 mil. words since 2007, as FIDAplus (Arhar et. al., 2007).

2.3. Recent state of affairs

Resources that could be devoted to POS-tagging were however limited and further research concentrated either in targeting specific language fields (Erjavec and Sárossy, 2006), techniques (Erjavec and Džeroski, 2004), tackling related research areas (Džeroski et al., 2006) and evaluating the current state of affairs (Lönneker, 2005). During the application of the Tree-Tagger, developed mainly for German and English and hence for much smaller tagsets, and applied to a 100 mil. word part of text corpus *Nova beseda*, accuracy estimate of 85% could be obtained (Lönneker, 2005, Lönneker and Jakopin, 2004).

During the time from 1996 to 2006 a reasonable collection of texts (1.3 mil. words), discussed in more detail in Chapter 3, has been machine-tagged and verified by linguists at the ISJ.

More recently the FIDAplus corpus has been tagged by Amebis, the leading Slovenian enterprise in the field of written language technology, using an own, rule-based tagger. The accuracy estimate of 85% has been given during the discussion after the presentation of a recent paper (Arhar and Holozan, 2008).

Whether 85% is a good figure or not depends on the intended use: for some applications it is more useful than for the other.

2008 however witnesses a revival in the field of Slovenian POS-tagging. A new project, JOS (*jezikoslovno označevanje slovenskega jezika* or *linguistic annotation of the Slovenian language*, Erjavec and Krek, 2008), though modest in scale, with a new tagset brought the two main tagsets (Multext-East and ISJ) much closer and a JOS-ISJ interface has been developed and tested during the first Slovenian POS-workshop (Jakopin, 2008).

1	** **		
2	**<p id="F0000015.13"/>**		
3	**<s id="F0000015.13.1"/>**		
4	**Postopanje**	postopanje	**Sosei**
5	**pred**	pred	**Do**
6	**afganistanskim**	afganistanski	**Ppnseo**
7	**veleposlaništvom**	veleposlaništvo	**Soseo**
8	**,**		
9	**širjenje**	širjenje	**!Sosei**
10	**govoric**	govorica	**Sozmr**
11	**o**	o	**Dm**
12	**tem**	ta	**!Zk-mem**
13	**,**		
14	**kdo**	kdo	**Zv-mei**
15	**bo**	biti	**Gp-pte-n**
16	**med**	med	**Do**
17	**prvimi**	prvi	**!Kbvmmo**
18	**letel**	leteti	**Ggnd-em**
19	**v**	v	**Dt**
20	**Afganistan**	Afganistan	**Slmetn**
21	**in**	in	**Vp**
22	**kakšne**	kakšen	**Zv-zmi**
23	**so**	biti	**Gp-stm-n**
24	**možnosti**	možnost	**Sozmi**
25	**,**		
26	**da**	da	**Vd**
27	**se**	se	**Zp------k**
28	**sploh**	sploh	**L**
29	**uvrstiš**	uvrstiti	**Ggdsde**
30	**na**	na	**Dt**
31	**seznam**	seznam	**Sometn**
32	**čakajočih**	čakajoč	**Ppnmmr**
33	**,**		
34	**so**	biti	**Gp-stm-n**
35	**postali**	postati	**Ggdd-mm**
36	**del**	del	**!Somei**
37	**novinarskega**	novinarski	**Ppnmer**
38	**vsakdana**	vsakdan	**Somer**
39	**.**		

Table 1: JOS-tagged sentence in vertical format

In Table 1 an example of a sentence is given, with token numbers, tokens, lemmas and tags. Approximate translation: *Hanging out in front of the Afghan embassy, spreading rumors about who will be among the first to fly to Afghanistan and what are the chances to join the waiting list have all become part of the journalist's everyday life*.

Table 2 shows the same sentence in horizontal format, as is used at the ISJ, and with tags converted to ISJ tagset.

```
* /<div id="F0000015"/
* /<POS tagged>
      <p><s>Postopanje pred afganistanskim
            Sse1        E6    Pse6
veleposlaništvom, širjenje govoric o
Sse6              Sse4      Sžp2     E5
tem   , kdo   bo    med prvimi letel v
ZKse5  ZVme1 GFPce E6  ŠVmp6  GLme  E4
Afganistan in  kakšne so   možnosti, da
IZme4       Vpr ZVžp1  GPcp Sžp1      Vpo
se  sploh uvrstiš na seznam čakajočih,
Gmp A     Gbe      E4 Sme4    PČžp2
so    postali del  novinarskega vsakdana.</s>
GPcp GLmp     Sme4 Pme2        Sme2
```

Table 2: Sentence from Table 1 in ISJ format

In Table 2 the same sentence, where tags have been converted to the ISJ tagset, is shown.

Some very welcome simplifications for the future tagging, such as in the adjective-verb relation, have been proposed (Krek, 2008).

3. ISJ POS-tagged corpus

In the decade from 1996 to 2006 a selection of texts from the *Nova beseda* text corpus has been tagged using the own stochastic tagger, supplemented by a database of 3 mil. wordforms and their tags, derived from 270.000 open-class lemmas. The lemmas were taken from the two monolingual dictionaries, The Dictionary of Standard Slovenian Language (SSKJ) and The Dictionary of Lesser Used Slovenian Words (BSJ), produced at the ISJ.

The tagged texts were manually checked by two linguists, the second author and by L. Uršič. The tagger has been integrated into a text editor and text with POS-tags was not shown in the more usual vertical format, shown in Table 1, but a line of text, followed by a line of tags, vertically aligned to words (Table 2). This screen arrangement made the task of manual checking considerably easier, as the text contingency, word context, was not lost to the viewer. In short, tagged text in such a layout was more human-readable.

Besides the mentioned database 3 other lists, extracted from the already tagged and checked texts are used in the disambiguation phase. The first is the list of wordforms with tags and frequencies.

```
tak • M,28; ZKme1,199; ZKme4,57; ZK,19
taka • ZKmd1,3; ZKsp1,9; ZKsp4,9; ZKže1,144
takale • ZKže1,5
take • ZKmp4,51; ZKže2,40; ZKžp1,46; ZKžp4,70
takega • ZKme2,40; ZKme4,18; ZKse2,110
takegale • ZKme4,1
takele • ZKmp4,2
takem • ZKme5,29; ZKse5,11
takemle • ZKse5,1
takemu • ZKme3,9; ZKse3,2
taki • ZKmp1,71; ZKžd1,1; ZKže3,4; ZKže5,25
```

Table 3: Part of the wordform-tag-frequency list

Here code M is used for interjection and ZK for demonstrative pronoun; other codes, used in tags are explained in the discussion of the Table 7.

The second list contains word n-grams (n=2-5) with corresponding tag n-grams and frequencies, as shown in Table 4.

modrim in • Pmp3 Vpr,2;Pse6 Vpr
modrim in bistrim • Pmp3 Vpr Pmp3,2
modrim in bistrim razodel • Pmp3 Vpr Pmp3 GLme,2
modrim in bistrim razodel pa • Pmp3 Vpr Pmp3 GLme Vpr,2
modrim in drugim • Pse6 Vpr ZDse6
modrim in drugim zelenim • Pse6 Vpr ZDse6 Pse6
modrim loncem • Pme6 Sme6
modrim loncem zajel • Pme6 Sme6 GLme
modrim loncem zajel vodo • Pme6 Sme6 GLme Sže4
modrim loncem zajel vodo in • Pme6 Sme6 GLme Sže4 Vpr
modrim lončkom • Pme6 Sme6
modrim lončkom nato • Pme6 Sme6 A
modrim lončkom nato šele • Pme6 Sme6 A Č

Table 4: Wordform-tag-frequency n-grams

Vpr Gmp GLme • 786
Vpr Gmp GLme A • 34
Vpr Gmp GLme A E3 • 3
Vpr Gmp GLme A E4 • 5
Vpr Gmp GLme A E5 • 1
Vpr Gmp GLme A GNE • 10
Vpr Gmp GLme A GNme1 • 1
Vpr Gmp GLme A Pme1 • 1
Vpr Gmp GLme A Vpo • 1
Vpr Gmp GLme A Vpr · 2
Vpr Gmp GLme A ZObe2 • 1
Vpr Gmp GLme A ZV • 1
Vpr Gmp GLme E ZV • 2
Vpr Gmp GLme E2 • 21

Table 5: Tag n-grams with frequencies

Tag codes for both tables are again explained on top of the next page.

The fact that open-class words in Slovenian occur in many inflected forms is the main reason for the ample size of the tagset. This is also the culprit for small frequencies and ample size of the tables. Especially Table 4 and Table 5 are not as dense as one would wish – a lot of entries are Hapax legomena.

An overview of the tagged text collection is given in Table 6.

Collected works of Ciril Kosmač	0.40 mil.
Tomo Križnar: O iskanju ljubezni	0.13 mil.
George Orwell: 1984	0.09 mil.
Plato: Republic	0.09 mil.
The Bible - New Testament	0.15 mil.
Gustave Flaubert: Bouvard and Pécuchet	0.09 mil.
DELO nespaper (1997)	0.05 mil.
Monitor computer magazine (2001)	0.07 mil.
DELO newspaper (2004)	0.07 mil.
Marko Uršič: Four Seasons	0.17 mil.
National Assembly session transcripts (2003)	0.05 mil.
Total	1.36 mil.

Table 6: ISJ POS-tagged corpus overview

It is mainly composed of fiction, there are four translations, a philosophical monograph (Four Seasons), a small amount of general newspaper language (DELO), computerese (Monitor) and formal speech (National Assembly session transcripts).

	Multext East	JOS	ISJ	freq.	Description
1.	Ccs	Vp	Vpr	61.639	coordinating conjunction
2.	Rgp	Rnn	A	50.520	adverb
3.	Vcip3s-n	Gp-ste-n	GPce	46.225	auxiliary verb to be in present, third person singular: *je*
4.	Vmps-smp	Ggdd-em	GLme	37.735	verb participle ending in -l, masculine, singular
5.	Spsl	Dm	E5	33.534	preposition requiring locative case
6.	Q	L	Č	33.351	particle
7.	Css	Vd	Vpo	33.211	subordinating conjunction
8.	Px------y	Zp------k	Gmp	25.651	separate verbial morpheme: *se*
9.	Voip3s-n	Ggnste	Gce	24.905	verb, present tense, third person singular
10.	Spsa	Dt	E4	23.065	preposition requiring accusative case
11.	Ncmsn	Somei	Sme1	19.357	noun, masculine gender, singular, nominative
12.	Spsi	Do	E6	17.662	preposition requiring instrumental case
13.	Ncfsa	Sozet	Sže4	16.244	noun, feminine gender, singular, accusative
14.	Ncfsn	Sozei	Sže1	16.065	noun, feminine gender, singular, nominative
15.	Npmsn	Slmei	Ime1	15.003	proper noun, masculine gender, singular, nominative
16.	Spsg	Dr	E2	14.577	preposition requiring genitive case
17.	Ncmsa	Somet	Sme4	13.843	noun, masculine gender, singular, accusative
18.	Vmps-pmp	Ggdd-mm	GLmp	12.737	verb participle ending in -l, masculine, plural
19.	Ncfsg	Sozer	Sže2	12.692	noun, feminine gender, singular, genitive
20.	Vmps-sfp	Ggdd-ez	GLže	11.951	verb participle ending in -l, feminine, singular

Table 7: Top 20 POS-tags with frequencies per million

In Table 7 the most frequent POS-tags form this corpus are given, in all three codings, the Multext-East, JOS and the coding developed at the ISJ. The Multext-East coding of POS-tags is English language based, i.e. V = verb, N = noun, P = pronoun, s = singular, p = plural, m = masculine, f =feminine. As the coding scheme had to fulfill all the grammatic criteria of the 7 languages involved, the codes are also relatively long – average length of the top 20 is 5.25. The ISJ coding has been developed with just Slovenian grammar in mind, G = glagol (verb), GP = pomožni glagol biti (verb to be, auxiliary), S = samostalnik (noun), P = pridevnik (adjective), A = prislov (adverb), Č = členek (particle), e = ednina (singular), p = množina (plural), m = moški spol (masculine), ž = ženski spol (feminine).

Resort to English, or better, foreign abbreviations was only used when it was the only possible option. An example is the abbreviation A for adverb, instead of Slovenian P (prislov). A was used because P found a better use for adjective (pridevnik). As no information, pertaining to other languages and grammars, hat to be taken into account, tag lengths also tend to be shorter. Average length of the top 20 ISJ tag codes is 3.10, more than two characters less than the value for the Multext-East coding. Shorter tags, especially if they tend to be self-explaining to the native linguist, are also easier to remember and the uneasy task of hand-checking less tedious. JOS tags are somewhere inbetween, with an average length of 4.50, better than 5.25. Many English abbreviations in coding have been substituted by Slovenian equivalents, such as S for noun (samostalnik), G for verb (glagol). Some fear to use Slovenian however remains, Č was not used for particle (členek) – second letter in the name, L was used instead, ž not for feminine gender (ženski spol), but z – ž without a hacek. Yet, all in all, a considerable step to make the conversion process easier has been made.

It is also worth noting that the top 20 tags add up to 52% of the total, and 2.180 tags actually appeared in the corpus (out of possible 5.000). Growth curve is shown in Figure 1. Top 2 POS-tags combined cover over 10%

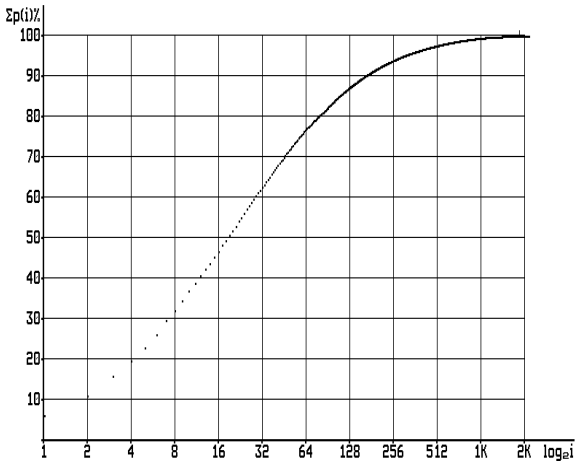


Figure 1: Growth curve for the 2.180 different tags with a total frequency of 1.36 million.

of the whole lot, top 8 combined about 33%, 19 over 50%, 100 about 80% and 200 close to 90% of the entire text. Over one half of the tags, 1.320, have a frequency 10 or above.

POS-tags	freq.	examples
IOme1	1843	*Bruno. Janez. Angel. Drejc! Stane!*
Č	674	*Da. Seveda. Vsekakor. Hvala lepa. Hvala.*
A	366	*Gotovo. Nujno. Dobro. Očitno. Jasno.*
M	334	*Aha! Amen. Halo! Hm? Mhm.*
GPae GLme	243	*sem rekel. sem nadaljeval, sem vprašal. sem rekel, sem nadaljeval.*
IOme1 IOme1	215	*Filotej (Bruno). Elpin (Janez). Matjaž Klančar Saša Vidmajer*
ZK GPce	196	*Tako je. Tako je, Tako je!*
Sme1	178	*Pozdrav. Oče! Stric! Mir! Molk.*
IOže1	174	*Marija. Tilčka! Ana! Božena! Justina!*
ZV	146	*Kako? Zakaj? Kam? Kdaj? Kje?*
Sže1	143	*Tišina. Mati! Neumnost! Mama! Pomlad!*
KURL	137	*www.microsoft.com www.ibm.com www.apple.com www.compaq.com*
GVbe	126	*Na! Pridi! Nehaj! Počakaj! Govori!*
IZže1 Š K Š	120	*Avstrija (100 ATS) 1294,8968 Francija (100 FRF) 2700,9905 Hrvaška (100 HRK) 2554,4020*
Pže1 Sže1	108	*Vremenska napoved. Gadja zalega! Cela mavrica! Kraška burja. Lepa smrt!*
GPce GLme IOme1	102	*je rekel Glavkon. je odvrnil Glavkon. je vprašal Winston. je odvrnil Adeimant.*
ČZ	97	*Ne. Ne! Ne, Ne?*
IOme1 GPce GLme	94	*Jezus je odgovoril: Peter je molčal. Jezus je rekel: Pécuchet je nadaljeval:*
KP KP	91	*C. R., G. R., M. B., D. G., D. V.*
GPce GLme	86	*je rekel. je odvrnil. je vprašal. je rekel, je odgovoril.*

Table 8: Top 20 sentence types with frequencies and example

In Table 8 top 20 sentence types are shown – the total number of sentences is 102.000. Very simple discourse-type sentences prevail. The majority of the entries in the table are minor clauses, in the role of calls, greetings and exclamations such as *Drejc!* (nickname), *Oče!* (*Father!*), *Počakaj* (*Wait!)* and sentences with elliptical clauses, that are grammatically incomplete.

POS-tags	freq.	*examples*
GPae GLme	243	*sem rekel. sem nadaljeval, sem vprašal. sem rekel, sem nadaljeval.*
ZK GPce	196	*Tako je. Tako je, Tako je!*
GVbe	126	*Na! Pridi! Nehaj! Počakaj! Govori!*
GPce GLme IOme1	102	*je rekel Glavkon. je odvrnil Glavkon. je vprašal Winston. je odvrnil Adeimant.*
IOme1 GPce GLme	94	*Jezus je odgovoril: Peter je molčal. Jezus je rekel: Pécuchet je nadaljeval:*
GPce GLme	86	*je rekel. je odvrnil. je vprašal. je rekel, je odgovoril.*
IOme1 Gmp Gce	81	*Bruno se nasmehne. Bruno se zasmeje. Bruno se muza. Bruno se smehlja.*
Gae	77	*Prosim. Razumem. Grem! Poslušam Mislim.*
IOme1 Gce	73	*Janez bere: Bruno vzdihne. Janez vstane. Najdù beži. Bruno okleva.*
GLme GPce	68	*Odgovoril je: Rekel je: Govoril je: Prisluhnil je. Dejal je:*
GVbp	68	*Izvolite. Pijte! Počakajte! Poslušajte! Berite!*
GLme ZOcmp3 GPce	54	*Rekel jim je: Odgovoril jim je: Dejal jim je: Dovolil jim je. Odvrnil jim je:*
GPce GLme Sme1	44	*je pribil stric. je prikimal stric. je vprašal Kovač. je vprašal oče.*
IOme1 ZOcme3 GPce GLme	40	*Jezus mu je rekel: Jezus mu je odgovoril: Jezus mu je odvrnil: Peter mu je rekel:*
GPce GLže Sže1	39	*je rekla teta. je vprašala mama. je vzkliknila teta. je poskočila teta. je puhnila teta.*
IOme1 ZOcmp3 GPce GLme	39	*Jezus jim je odgovoril: Jezus jim je rekel: Jezus jim je odvrnil: Pilat jim je rekel:*
Gce	38	*Bere. Gori! Bode! Pride! Reče.*
Sže1 GPce GLže	38	*Brzostrelka je zadrdrala. Množica je jeknila. Množica je molčala. Teta je umolknila.*
M Č Gmp GPce GLme Sme1	36	*Hm, kajpak, se je popraskal kmet. Hm, kajpak, se je oglasil kmet.*
ČZ Gae	33	*Ne vem. Ne razumem. Ne morem! Ne smem! Ne maram!*

Table 9: Top 20 sentences containing at least one verb with frequencies and examples

In Table 9 the most frequent sentences, which contain at least one major clause, are shown. It is easy to notice that most clauses are exclamative, imperative and declarative. Declarative clauses have a very simple choice of Subjects (1st person and 3rd person masculine). The Predicator most often expresses verbal processes, usually in past tense.

Conclusion

POS-tagging of a morphologically-rich language such as Slovenian brings to surface all the issues, coming from the fluidity and fuzziness of a scientific phenomenon called language, obviously not a very suitable arena for the application of "hard-core" techniques and algorithms so successful in areas such as physics, chemistry or engineering. Yet a combined application of knowledge, accumulated in the available POS corpus and the availability of new resources (Žele, 2008), could definitively bring a positive, acceptable outcome within reach.

4. References

Arhar, Š., Holozan, P. 2008. ASES - Lexical Database as the Language Resource for Slovene Language Technology Development . *Glasnik ZRS Koper*. 3: 30

Arhar, Š., Gorjanc, V., Krek, S. 2007. FidaPLUS corpus of Slovenian. The New Generation of the Slovenian Reference Corpus: Its Design and Tools. In *Proceedings of the Corpus Linguistics Conference CL 2007*. Birmingham 1/219.

Bizjak Končar, A. 2008. Dvojina v Dalmatinovem Pentatevhu. Symposium *Slovenski knjižni jezik v 16. stoletju*. Ljubljana: ZRC SAZU. 17-19 April.

Bizjak, A. 2005. *Pridiga kot žanr*. Ljubljana: Založba ZRC

Džeroski, S., Erjavec, T., Ledinek, N., Pajas, P., Žabokrtsky, Z., Žele, A. 2006. Towards a Slovene Dependency Treebank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, LREC'06. Paris: ELRA, 1388-1391.

Džeroski, S., Erjavec, T., Zavrel, J. 2000. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. *Second International Conference on Language Resources and Evaluation*, LREC'00, Paris: ELRA, 1099-1104.

Erjavec, T., Ide, N. 1998. The MULTEXT-East Corpus. *First International Conference on Language Resources and Evaluation*, LREC'98, Granada: ELRA, 971-974.

Erjavec, T., Sárossy, B. 2006. Morphosyntactic Tagging of Slovene Legal Language. *Informatica* 30:483-488.

Erjavec, T., Džeroski, S. 2004. Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words. *Applied Artificial Intelligence*, 18(1): 17-40.

Erjavec, T., Krek, S. 2008. The JOS Morphosyntactically Tagged Corpus of Slovene. *Sixth International Conference on Language Resources and Evaluation*, LREC'08, http://www.lrec-conf.org/proceedings/lrec2008/.

Jakopin, P., Bizjak, A. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična Revija* 45/3-4:513-532.

Jakopin, P. 2002. *Entropija v slovenskih leposlovnih besedilih*. Ljubljana: Založba ZRC

Jakopin, P., Michelizza M. 2007. Besedilni korpus Nova beseda. *Mostovi*, glasilo DZTPS, in printJakopin, P. 2008. Oblikoslovno označevanje besedil. *Workshop*. Ljubljana: ZRC SAZU, 5-6 May.

Krek, S. 2008. Kratek pregled jezikoslovnih zadreg pri oblikoslovnem označevanju slovenščine. *Posvet o sistemih oblikoskladenjskega označevanja za slovenščino*. UP Koper. April 4.

Lönneker, B. 2005. Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo. *Slavistična revija* 53/2:193-210.

Lönneker, B., Jakopin, P. 2004. Checking POSBeseda, a Part-of-Speech tagged Slovenian corpus. *Zbornik 7. mednarodne multikonference Informacijska Družba IS 2004*. vol. B:48-55.

Verdonik, D., Rojc, M., Kačič Z., Horvat, B. 2002. Zasnova in izgradnja oblikoslovnega in glasoslovnega slovarja za slovenski knjižnji jezik. *Zbornik 6. mednarodne multikonference Informacijska Družba IS 2002*. vol. B, 44-48.

Verdonik, D., Rojc, M. 2004. Jezikovni viri projekta LC-STAR. *Zbornik 7. mednarodne multikonference Informacijska Družba IS 2002*. vol. B, 43-47.

Žele, A. 2008. *Vezljivostni slovar*. Ljubljana: Založba ZRC.

Improving morphosyntactic tagging of Slovene by tagger combination

Jan Rupnik, Miha Grčar, Tomaž Erjavec

Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
jan.rupnik@ijs.si, miha.grcar@ijs.si, tomaz.erjavec@ijs.si

Abstract

Part-of-speech (PoS) or, better, morphosyntactic tagging is the process of assigning morphosyntactic categories to words in a text, an important pre-processing step for most human language technology applications. PoS-tagging of Slovene texts is a challenging task since the size of the tagset is over one thousand tags (as opposed to English, where the size is typically around sixty) and the state-of-the-art tagging accuracy is still below levels desired. The paper describes an experiment aimed at improving tagging accuracy for Slovene, by combining the outputs of two taggers – a proprietary rule-based tagger developed by the Amebis HLT company, and TnT, a tri-gram HMM tagger, trained on a hand-annotated corpus of Slovene. The two taggers have comparable accuracy, but there are many cases where, if the predictions of the two taggers differ, one of the two does assign the correct tag. We investigate training a classifier on top of the outputs of both taggers that predicts which of the two taggers is correct. We experiment with selecting different classification algorithms and constructing different feature sets for training and show that some cases yield a meta-tagger with a significant increase in accuracy compared to that of either tagger in isolation.

1. Introduction

Morphosyntactic tagging, also known as part-of-speech tagging or word-class syntactic tagging is a process in which each word appearing in a text is assigned an unambiguous tag, describing the morphosyntactic properties of the word token. Such tagging is the basic pre-processing step for a number of applications or more advanced analysis steps, such as syntactic parsing. Morphosyntactic tagging is, in general, composed of two parts: the program first assigns, on the basis of a morphological lexicon all the possible tags that a word form can be associated with (morphological look-up), and then chooses the most likely tag on the basis of the context in which the word form appears in the text (disambiguation). For words not appearing in the lexicon, various taggers can either ignore them or employ various heuristics to guess at their tag.

Unlike English, morphologically richer Slavic languages such as Czech (Hajič and Hladka, 1998) or Slovene typically distinguish more than a thousand morphosyntactic tags. In the multilingual MULTEXT-East specification (Erjavec, 2004) almost 2,000 tags (morphosyntactic descriptions, MSDs) are defined for Slovene. MSDs are represented as compact strings, with positionally coded attribute values, so they effectively serve as shorthand notations for feature-structures. For example, the MSD `Agufpa` expands to `Category = Adjective, Type = general, Degree = undefined, Gender = feminine, Number = plural, Case = accusative`.

Having such a large number of tags makes assigning the correct one to each word token a much more challenging task than it is e.g. for English. The problem for Slovene has been exacerbated by the lack of large and available validated tagged corpora, which could serve as training sets for statistical taggers.

Recently, new annotated language resources have become available for Slovene. FidaPLUS[1] (Arhar & Gorjanc, 2007) is a 600 million word monolingual reference corpus automatically annotated with MULTEXT-East MSDs by the Slovene HLT company Amebis[2]. But while FidaPLUS is freely available for research via a Web concordancer, it is not generally available as a dataset. In order to remedy the lack of publicly available annotated corpora for HLT research on Slovene, the JOS project (Erjavec and Krek, 2008) is making available two corpora under the Creative Commons license. Both contain texts sampled from FidaPLUS, with the smaller jos100k containing 100,000 words with fully validated JOS morphosyntactic annotations, and the larger, jos1M having 1 million words, and partially hand validated annotations – project resources preclude fully validating the latter.

Previous experiments (Erjavec et al., 2000) showed that from various publicly accessible taggers the best results were achieved by TnT (Brants, 2000). TnT is a Hidden Markov Model tri-gram tagger, which also implements an unknown-word guessing module. It is fast in training and tagging, and is able to accommodate the large tagset used by Slovene.

Having the validated jos100k at our disposal, we experimented with training TnT and seeing how its errors compare to the ones assigned by the Amebis tagger. It turned out that the two taggers are comparable in accuracy, but make different mistakes. This gave us a method of selecting the words that should be manually corrected in jos1M – only those tokens where the annotations between the taggers differ are selected for manual inspection. This approach concentrates on validating the words where state-of-the-art taggers are still able to make correct decisions, at the price of ignoring cases where both taggers predict the same but incorrect tag, i.e. the truly difficult cases.

Having several tags for each word also offers the possibility of combining their outputs in order to increase accuracy, say, over the whole FidaPLUS corpus. Experiments in combining PoS taggers have been attempted before, using various learning strategies, and for various languages, e.g. voting, stacking, etc. for Swedish (Sjöbergh, 2003) or multi-agent systems for Arabic (Othmane Zribi et al., 2006). An experiment, more similar

[1] http://www.fidaplus.net/

[2] http://www.amebis.si/

to ours, is reported in Spoustová et al. (2007) for Czech, also using a rich positional tagset, where several stochastic taggers are combined with a rule based one; the rule based tagger is used predominantly as a pre-disambiguation step, to filter out unacceptable tags from the ambiguity classes of the tokens.

This paper presents a similar experiment, which, however, uses only two independent taggers, so precluding combination methods such as voting or pipelining. But as in the Czech case, we also need to deal with a very large, although positionally encoded tagset.

The rest of this paper is structured as follows: Section 2 presents the dataset used in the experiments, Section 3 explains the methods used to combine the output of the taggers, Section 4 gives the experimental results performed with different methods and features, and Section 5 gives the conclusions and directions for further work.

2. The Dataset

The dataset used in the experiments is based on the jos100k corpus; the corpus contains samples from almost 250 texts from FidaPLUS, cca. 1,600 paragraphs or 6,000 sentences. The corpus has just over 100,000 word tokens, and, including punctuation, 120,000 tokens. jos100k contains only manually validated JOS MSDs, of which there are 1,064 different ones.

For the dataset we added MSDs assigned by Amebis and TnT. Two sentences from the dataset are given in Figure 1. Annotations marking texts and paragraphs have been discarded and end of sentence is marked by an empty line. Punctuation is tagged with itself.

Prišlo	Vmep-sn	Vmep-sn	Vmep-sn
je	Va-r3s-n	Va-r3s-n	Va-r3s-n
do	Sg	Sg	Sg
prerivanja	Ncnsg	Ncnsg	Ncnsg
in	Cc	Cc	Cc
umrla	Vmep-sf	Vmep-sf	Vmep-sf
je	Va-r3s-n	Va-r3s-n	Va-r3s-n
.	.	.	.
Tega	Pd-nsg	Pd-msa	Pd-msg
se	Px------c	Px------c	Px------c
sploh	Q	Q	Q
nisem	Va-r1s-y	Va-r1s-y	Va-r1s-y
zavedel	Vmep-sm	Vmep-sm	Vmep-sm
.	.	.	.

Figure 1. Example stretch of the corpus dataset ("*Prišlo je do prerivanja in umrla je. Tega se sploh nisem zavedel.*"). First column is the word-form, second the gold standard tag, third the one assigned by TnT, and the fourth by Amebis.

The source FidaPLUS corpus also contains, for each word token, all possible MSDs that could be assigned to it, i.e. its ambiguity class. Based on this information, we computed the average per-word MSD ambiguity which turns out to be 3.13 for the jos100k corpus. So, on the average, a tagger needs to choose the correct MSD tag between three possibilities. Note that disambiguation is only possible for known words.

2.1. Amebis MSDs

The Amebis MSDs were taken from the source FidaPLUS corpus; as mentioned, the Amebis tagger is largely a rule-based one, although with heuristics and quantitative biases. The tagger uses a large lexicon, leaving only 2% of the word tokens in jos100k unknown. Amebis doesn't tag these words, and they have all been given a distinguished PoS/MSD "unknown". Furthermore, FidaPLUS is annotated according to the MULTEXT-East specification, while the JOS corpus uses a modification, based on, but different from the MULTEXT-East/FidaPLUS one. Differences concern reordering of attribute positions, changes in allowed values, etc., as well as lexical assignment. For the most part an information-preserving conversion is possible, but for MSDs (attributes) of some lexical items only heuristics are possible. Taking into account that all Amebis "unknowns" are by definition wrong, as all words are manually annotated with specific MSDs, and that a certain number of errors is introduced by the tagset mapping, Amebis obtains 87.9% accuracy on all tokens in the dataset.

2.2. TnT MSDs

The TnT tagger was trained on the dataset itself, using 10-fold cross-tagging. The dataset was split into 10 parts, and 9 folds were used for training, and the remaining fold was tagged with the resulting model, and this process repeated for all 10 folds. As the lexical stock of jos100k is small, the tagging model used a backup lexicon which was extracted from the FidaPLUS corpus and its annotations. In other words, tri-gram statistics and lexicon containing uni-gram statistics of word-forms (their ambiguity classes) of frequent words were learned from jos100k, while less frequent words obtained their ambiguity classes from MSDs assigned by the Amebis tagger. Given such a tagging set-up, the obtained accuracy over the all dataset tokens for TnT is 88.7%, slightly better than Amebis; but TnT has the advantage of learning how to correctly tag at least some unknown words (such as those marked as "foreign", i.e. tokens in spans of non-Slovene text), as well as having less problems with tagset conversion. Nevertheless, on the dataset it performs better than Amebis, so the TnT accuracy can be taken to constitute the baseline for the experiment.

2.3. Error comparison

Table 1 compares the errors made by the taggers against the gold standard. The first line gives the complete size of the corpus in words. The second gives the number of correct MSD assignment to word tokens for TnT (86.6% per-word accuracy), and the third for Amebis (85.7%). The fourth line covers cases where both taggers predict the correct MSD, for 78% of the words.

Lines 5 and 6 cover cases where one tagger correctly predicts the tag, while the other makes a mistake. These two lines cover a significant portion (2/3) of all the errors, so if such mistakes can be eliminated by deciding which tagger made the correct choice, the gains in accuracy are considerable.

The last two lines indicate upper bounds on the gains achieved by concentrating on choosing the correct tag.

Line 7 gives cases where both taggers agree, but on an incorrect tag (3.2%), and line 8 the number of cases where both are wrong, but in different ways (2.4%); the upper bound on combination accuracy is thus 94.3%.

Let us look at two typical examples of cases 7 and 8. An example of both taggers being wrong, but agreeing on the assigned tag is exemplified in the fragment *"ni mogoče povedati" (it is not possible to tell)* where *"mogoče"* is correctly tagged as an adverb but both taggers assign it an adjectival tag. As an example of both taggers being wrong in different ways is the fragment *"ni priporočene/Adj zgornje/Adj mejne/Adj vrednosti/Adj" (there is no recommended upper boundary value)*. The correct tag for the noun is *Ncfsg*, i.e. feminine singular genitive, the genitive being determined by the (long distance) dependency on *"ni"*. The Amebis tagger correctly predicts this tag, while TnT makes a mistake, and assigns to the noun the plural accusative. As adjectives must agree with the noun in gender, number and case, the three adjectives preceding the noun must also be tagged as feminine singular genitive. Here both taggers are wrong: while TnT correctly posits the agreement between the noun and adjectives, all the adjective tags are wrong, due to the noun being incorrectly tagged. Amebis, on the other hand, does not pick up the agreement, and tags all three adjectives as masculine ones.

	Words	Gold	Amebis	TnT	Gloss
1	100,003	MSD1			Words in dataset
2	86,617	MSD1		MSD1	TnT tagger correct
3	85,719	MSD1	MSD1		Amebis tagger correct
4	78,011	MSD1	MSD1	MSD1	Both taggers correct
5	7,708	MSD1	MSD1	MSD2	Amebis correct, TnT error
6	8,606	MSD1	MSD2	MSD1	Amebis error, TnT correct
7	3,238	MSD1	MSD2	MSD2	Both wrong, and identical
8	2,440	MSD1	MSD2	MSD3	Both wrong, and different

Table 1: Comparison of tagging accuracy of Amebis and TnT over the 100k dataset.

3. Combining the taggers

As mentioned, our meta-tagger is built on top of two taggers, TnT and the Amebis rule-based tagger. The sole task of the meta-tagger is to decide which tag to consider correct. The meta-tagger is implemented as a classifier which, if the two underlying taggers disagree, classifies the case into one of the two classes indicating which of the two taggers is more likely to be correct. To train the classifier, we needed two things: a way to describe a case with a set of features, and a classification algorithm. The following section describes the feature construction process and the subsequent section the classification algorithms we tried out for this task.

3.1. Feature construction

To be able to train the classifier we needed to describe each case with a set of features. We decided to keep our meta-tagger relatively simple and to construct features solely out of tags predicted by the underlying taggers. Alternatively, we could compute content features as well (such as *n*-grams, prefixes, and suffixes) as it is the case with the SVM-based taggers such as SVMTool (Giménez & Márquez, 2004).

For training we used a sequence of 100,000 Slovene words (see Section 2 for more details), with each word assigned three tags: the correct tag (assigned manually), a tag assigned by TnT, and a tag assigned by the Amebis tagger. Each of these three tags can be decomposed into 15 attributes such as the part-of-speech category, type, gender, number, and so on. For a given tag, not all attribute values are set, therefore the data is sparse in this sense (e.g. the value of gender and number for prepositions is "undefined").

The attributes of the tags assigned by the two taggers (but not those of the manually assigned tags) were directly used as features for training. In addition, we constructed features that indicate whether the two taggers agree on a particular attribute value or not (the so called agreement features). The example was labeled according to the tagger which correctly tagged the word (the label was thus either TnT or Amebis). Note that we built a training feature vector only when the two taggers disagreed and one of them was correct (if none of the taggers was correct, we were unable to label the feature vector). The entire feature construction process is illustrated in Figure 2.

For the first set of experiments we used the tag attributes and agreement features of the current word to construct a feature vector (termed non-contextualized features in Figure 2). In the second set of experiments, on the other hand, we also added tag features (from both, TnT and Amebis) from the previous and the next word (termed contextualized features in Figure 2). Also important to mention is that we ran a set of experiments where we excluded punctuation from the text and a set of experiments where each different type of punctuation was treated as a separate part-of-speech category (e.g. POS_T=,) with all the other attributes set to "not applicable". Each of these settings gave slightly different results. The results are discussed in Section 4 in more details.

3.2. Learning algorithms

We experimented with three different classification algorithms: the Naive Bayes classifier, CN2 rule-induction algorithm, and C4.5 decision tree building algorithm. In this section, we briefly describe each of them.

The **Naive Bayes (NB)** classifier is a probabilistic classifier based on Bayes' theorem.[3] It naively assumes a strong independence of features. Furthermore, it is a black box classifier in the sense that its decisions are not easily explainable.

CN2 is an if-then rule-induction algorithm (Clark & Niblett, 1989). It is a covering algorithm meaning that each new rule covers a set of examples which are thus removed from the dataset. Unlike the Naive Bayes classifier, the trained model (i.e. a set of induced rules) provides an explanation for a decision (i.e. an if-then rule

[3] c.f. http://en.wikipedia.org/wiki/Naive_Bayes_classifier

that was taken into account when classifying the example). Looking at the induced rules, it is also possible to read, understand, and also verify the knowledge that was discovered in the training set.

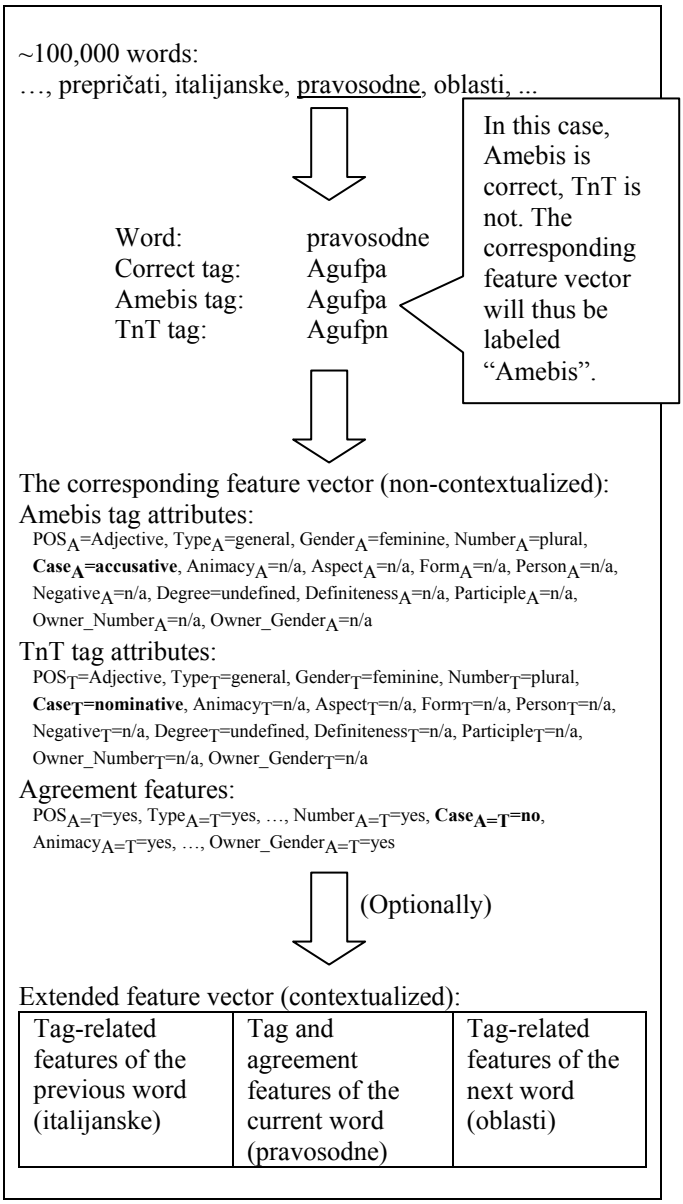


Figure 2: The feature construction process.

C4.5 is an algorithm for building decision trees; it is based on information entropy[4] (Quinlan, 1993). C4.5 uses the fact that each attribute of the data can be used to make a decision that splits the data into smaller subsets. It examines the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision. This process is repeated several times on smaller and smaller subsets of data. Similarly to CN2 (the rule-induction algorithm), C4.5 builds glass box models. Unlike its predecessor, the ID3 algorithm, C4.5 knows how to handle data with missing values (i.e. sparse data) and prunes the tree by cutting off branches that do not contribute to the classification accuracy.

4. Experiments

This section is focused on testing tagging accuracies of the meta-tagger for different combinations of feature sets and underlying classification models. The size of the set of examples for training and testing is 16,072 and consists of 8,518 cases where TnT tagger predicted the correct tag and Amebis tagger did not and 7,554 cases where Amebis was correct and TnT was not. All experiments were conducted with the Orange data mining tool (Demšar et al., 2004). 5-fold cross validation method was used to evaluate the tagging accuracy of the meta-tagger in all experimental scenarios. We first discuss two baseline models for the meta-tagger, after that we define the different feature sets, then continue with the description of non-contextualized models and end the section with models that incorporate context features.

4.1. Baselines

The first baseline is the majority classifier which always predicts that TnT tagger is correct. This classifier achieves the accuracy of 53%.

The second baseline model is a Naive Bayes model trained on only one feature: Amebis MSD. This is a very simple model, since to classify a new example (with only one feature *f* , that is the Amebis MSD), all one needs to do is count the number of cases with MSD equal to *f* where Amebis was correct and the number of cases with MSD equal to *f* where Amebis was incorrect (*P(x = f, y = amebis-correct)* and *P(x = f, y = amebis-incorrect)*) and predict the class (amebis-correct or amebis-incorrect) with the higher count. This is the simplest nontrivial model and it achieves the accuracy of 70.95% (28% higher than the first baseline).

Let us consider two examples. Assume that there were 200 cases where Amebis predicted the tag Pd-nsg, and it was correct in 150 of these cases (this means the TnT was correct in the remaining 50 cases). This means that P(Amebis-predicts: Pd-nsg, Amebis-correct) = 0.75. In this case the meta tagger would always predict the tag Pd-nsg if Amebis predicted it as well.

Now, if we assumed that Amebis was correct in 80 of 200 cases where it predicted, P(Amebis-predicts: Pd-nsg, Amebis-correct) = 0.4), then the meta-tagger would always predict the tag that the TnT tagger predicted, given that Amebis predicted Pd-nsg (the evidence in the training data tells us not to trust the Amebis tagger, since the probability of it being correct was less than 0.5).

4.2. Feature sets

We will now describe the features for the non-contextualized models. The first set of features for the non-contextualized models are called FULL features; they only include full Amebis MSD and full TnT MSD (two features). The second set of features called DEC is a decomposition of the FULL features as described in Section 3.1 (45 features: 15 Amebis features, 15 TnT features, 15 Agreement features). The third set of features, BASIC, is a subset of DEC features, where we only take the features corresponding to Category, Type, Gender, Number and Case (10 features: 5 for Amebis and 5 for

[4] c.f. http://en.wikipedia.org/wiki/C4.5_algorithm

TnT). The final set of features, ALL, is a union of FULL and DEC (47 features).

Feature sets for contextualized models (with and without punctuation) are extensions of non-contextualized feature sets, where the features of examples surrounding our training example are added (see Section 3.1). The context features (i.e. the features of the previous and next word) are the same ones as that of the current word except for the Agreement features which are only computed for the current word (in the DEC feature set we thus keep only 15 Agreement features: the ones of the current word).

Features ALL now include six features for MDS tags (Amebis-Prev, Amebis, Amebis-Next, TnT-Prev, TnT, TnT-Next), 45 for Amebis tag features (3 × 15 features), 45 for TnT tag features and 15 Agreement features, which sums up to 111 features.

4.3. Non-contextualized models

Experiments with features that do not take context into account (Table 2) show that C4.5 is the most robust classifier with respect to different feature sets and that it can achieve the highest accuracy. We can also observe that tag features are not very suitable for the Naive Bayes classifier, because the conditional independence assumptions are too strongly violated. Even though the CN2 algorithm results in slightly lower accuracy it can prove useful since the rules that it produces are easy to interpret and thus discover the strengths and weaknesses of the TnT and Amebis classifiers (see Figure 3).

Length	Quality	Coverage	Class	Distribution	Rule
2	0.999	760.0	Tnt	<0.0,760.0>	IF Amebis_POS=['Residual'] AND Tnt_Form=['0.000'] THEN Correct=Tnt
3	0.991	109.0	Amebis	<109.0,0.0>	IF Amebis_Case=['locative'] AND Tnt_Type=['common'] AND Agreement_in_Case=['no'] THEN Correct=Amebis
3	0.990	192.0	Tnt	<1.0,191.0>	IF Amebis_Aspect=['imperfective'] AND Amebis_Number=['dual'] AND Amebis_Gender=['neuter'] THEN Correct=Tnt
4	0.988	81.0	Amebis	<81.0,0.0>	IF Amebis_Type=['general'] AND Tnt_Number=['plural'] AND Tnt_Case=['genitive'] AND Agreement_in_Case=['no'] THEN Correct=Amebis
3	0.987	77.0	Tnt	<0.0,77.0>	IF Tnt_Type=['subordinating'] AND Amebis_Person=['0.000'] AND Amebis_Animacy=['0.000'] THEN Correct=Tnt
3	0.986	69.0	Amebis	<69.0,0.0>	IF Tnt_Definiteness=['definite'] AND Amebis_Gender=['feminine'] AND Agreement_in_Type=['yes'] THEN Correct=Amebis
3	0.982	55.0	Amebis	<55.0,0.0>	IF Tnt_Definiteness=['definite'] AND Amebis_Number=['plural'] AND Agreement_in_Type=['yes'] THEN Correct=Amebis
3	0.982	53.0	Amebis	<53.0,0.0>	IF Amebis_Gender=['feminine'] AND Amebis_Form=['participle'] AND Tnt_Number=['dual'] THEN Correct=Amebis

Figure 3: List of rules discovered by CN2 in Orange. Rules are ordered by their quality which is a function of rule coverage and rule accuracy. The second rule, for example, tells us that if Amebis predicted locative case and TnT predicted some other case and TnT predicted common type, then the meta-tagger should predict the same tag as Amebis. The first rule, IF Amebis_POS=[‘Residual’] AND TnT_Form=[‘0.000’] THEN Correct = TnT, covers the examples mentioned in Section 2.1, where Amebis predicts POS tag “unknown” (by default incorrect). The rule says that in such case, TnT is always correct, which is what is expected.

Table 2: Non-contextualized models (accuracy in %). Feature sets FULL, DEC, BASIC and ALL are explained in Section 4.2.

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.90	67.55	67.50	69.65
C4.5	73.51	**74.70**	74.23	73.59
CN2	60.61	72.57	71.68	70.90

4.4. Context and punctuation

When comparing the results of experiments with context, we notice that taking punctuation into account (see Section 3.1) is beneficial in almost all cases (see Tables 3 and 4). This can be explained by the fact that ignoring punctuation can yield unintuitive context tags, for instance the sequence of tags T1, T2, T3, where T1 is the last word of a sentence, T2 the first word and T3 the second word of the next sentence.

We notice that C4.5 can best benefit from extra contextual features, whereas the performance of the other algorithms does not change notably.

Table 3: Context without punctuation (accuracy in %).

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.10	68.29	67.96	70.55
C4.5	73.10	78.51	**79.23**	76.72
CN2	62.16	73.26	72.75	72.29

Table 4: Context with punctuation (accuracy in %).

Feature set / Classifier	FULL	DEC	BASIC	ALL
NB	73.44	68.32	68.14	70.53
C4.5	74.18	78.91	**79.73**	77.68
CN2	62.23	74.27	72.82	73.01

5. Conclusions

The paper presents a meta-tagger built on top of two taggers, namely the TnT HMM-based tagger and the Amebis rule-based tagger. The purpose of the meta-tagger is to decide which tag to take into account if the two taggers disagree in a particular case.

The experimental results show that the two taggers are quite orthogonal since very little information is needed to get a significant increase in performance from the first baseline.

Furthermore, using context can improve the performance of some models and taking punctuation into account when constructing context features is better than ignoring it. C4.5 with BASIC context features with punctuation achieved the highest accuracy, 79.73%, which resulted in a meta-tagger with significantly lower error rate than Amebis tagger or TnT tagger. The overall error rates are given in Figure 4. Note that the TnT overall error is equal to the first baseline error.

For the next step we will train and test the meta-tagger on the jos1M hand-validated corpus. With this we will avoid doing the 10-fold cross-tagging mentioned in Section 2.2 and clearly separate the dataset for training TnT and that for training the meta-tagger.

There are roughly 5% cases in which both taggers assign an incorrect tag. By using the technique, discussed in this paper (i.e. rule inference), it would be possible to learn under which conditions the two taggers are both mistaken and thus alert the user about such tags.

Furthermore, it would be possible to apply our technique on a per-attribute base. We would be able to predict incomplete tags, i.e. tags with some attributes missing, where the missing attributes would be those most likely predicted falsely by both taggers. This would be very useful as guidance for human taggers preparing the JOS corpus. The missing attributes would have to be entered manually; the rest would only need to be validated.

Tagging on a per-attribute base and looking at cases in which both taggers predict an incorrect tag will be the focus of our future research. In addition, we will consider including more taggers into the system. The main idea is to develop taggers, specialized to handle cases in which the two currently used taggers are not successful.

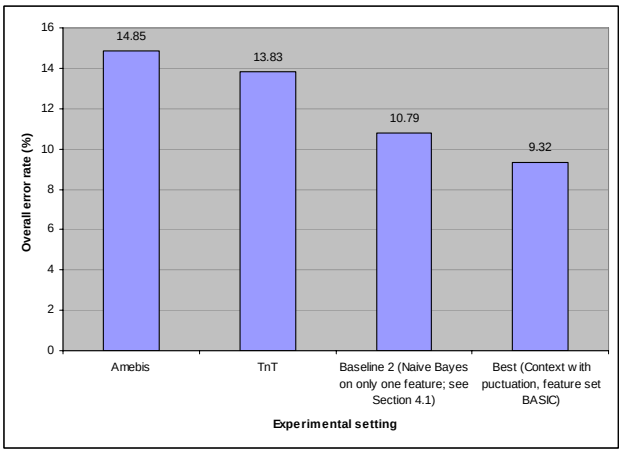


Figure 4: The overall error rates (%). We can see that our meta-tagger exhibits around 4% improvement over the two underlying taggers (i.e. TnT and Amebis).

Acknowledgements

The authors wish to thank the anonymous reviewers for their useful comments. The work described in this paper was supported in part by grant ARRS J2-9180 "Jezikoslovno označevanje slovenskega jezika: metode in viri" and EU 6FP-033917 SMART "Statistical Multilingual Analysis for Retrieval and Translation".

References

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

Brants T. (2000). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, 224–231.

Clark, P. and Niblett, T. (1989). *The CN2 Induction Algorithm.* Machine Learning, 3(4): 261–283.

Demšar J., Zupan B. and Leban G. (2004). *Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper* (www.ailab.si/orange), Faculty of Computer and Information Science, University of Ljubljana.

Erjavec, T., Džeroski, S., Zavrel, J. (2000). Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'2000)*. ELRA, Paris.

Erjavec, T. (2004). MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*, 1535–1538.

Erjavec, T. and Krek S. (2008). The JOS morphosyntactically tagged corpus of Slovene. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*.

Giménez, J. and Márquez, L. (2004). SVMTool: A General POS Tagger Generator Based on Support Vector Machines. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Hajič, J., Hladka, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. *COLING-ACL'98.* ACL.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers, Inc.

Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text. In *NoDaLiDa 2003, 14th Nordic Conference on Computational Linguistics.* Reykjavik.

Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P. (2007). The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing. June 2007. Prague, Czech Republic. Association for Computational Linguistics.

Zribi, C.B.O., Torjmen, A. and Ahmed, M.B. (2006). An Efficient Multi-agent System Combining POS-Taggers for Arabic Texts. In *Computational Linguistics and Intelligent Text Processing. LNCS Volume 3878/2006*, Springer.

Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian

Željko Agić*, Marko Tadić, Zdravko Dovedan***

*Department of Information Sciences
**Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{zeljko.agic, marko.tadic, zdravko.dovedan}@ffzg.hr

Abstract

This paper investigates several methods of combining output of a second order hidden Markov model part-of-speech/morphosyntactic tagger and a high-coverage inflectional lexicon for Croatian. Our primary motivation was to improve overall tagging accuracy of Croatian texts by using our newly-developed tagger. We also wanted to compare its tagging results – both standalone and utilizing the morphological lexicon – to the ones previously described in (Agić and Tadić, 2006), provided by the TnT statistical tagger which we used as a reference point having in mind that both implement the same tagging procedure. At the beginning we explain the basic idea behind the experiment, its motivation and importance from the perspective of processing the Croatian language. We also describe all the tools and language resources used in the experiment, including their operating paradigms and input and output format details that were of importance. With the basics presented, we describe in theory all the possible methods of combining these resources and tools with respect to their paradigm, input and production capabilities and then put these ideas to test using the F-measure evaluation framework. Results are then discussed in detail and conclusions and future work plans are presented.

1. Introduction

After obtaining satisfactory results of the preliminary experiment with applying a second order hidden Markov model part-of-speech/morphosyntactic tagging paradigm by using TnT tagger on Croatian texts – detailed description of the experiment given in (Agić and Tadić, 2006) and TnT tagger described in (Brants, 2000) – we decided to give it a try to reach a higher level of accuracy based on these results. (Please note that abbreviation HMM is used instead hidden Markov model and PoS/MSD tagging instead part-of-speech/morphosyntactic tagging further in the text).

In the section about our future work plans in (Agić and Tadić, 2006), we provided two main directions for further enhancements:

(i) producing new, larger and more comprehensive language resources, i.e. larger, more precisely annotated and systematically compiled corpora of Croatian texts, maybe with special emphasis on genre diversity and

(ii) developing our very own PoS/MSD tagger based on HMMs – being that TnT is available to public only as a black-box module – and then altering it by adding morphological cues about Croatian language or other rule-based modules.

We considered both courses of action as being equally important; HMM PoS/MSD trigram taggers make very few mistakes when trained on large and diverse corpora encompassing most of morphosyntactic descriptions for a language and, on the other hand, these taggers rarely seem to achieve 97-98% accuracy on PoS/MSD (excluding the tiered tagging approach by (Tufiş, 1999.) and (Tufiş, Dragomirescu, 2004)) without the help of rule-based modules, morphological cues or other enhancements which in fact turn statistical tagging systems into hybrid ones. Therefore, we have reasonably chosen to undertake both courses of action in order to create a robust version of Croatian PoS/MSD tagger that would be able to provide us automatically with new and well-annotated Croatian language resources.

However, knowing that manual production of MSD-tagged corpora takes time and human resources, we put an emphasis on developing and fine-tuning the trigram tagger in this experiment. Here we describe what is probably the most straightforward of fine-tuning options – combining the tagger and the Croatian morphological lexicon (HML), described in (Tadić and Fulgosi, 2003) and implemented in form of Croatian lemmatization server, described in (Tadić, 2006) and available online at hml.ffzg.hr.

Section 2 of the paper describes all the tools, language resources, annotation standards, input and output formats used in the experiment, while section 3 deals in theory with various methods of pairing trigram tagger and the before-mentioned morphological lexicon. Section 4 defines the evaluation framework that would finally provide us with results. Discussion and conclusions along with future plans are given in sections 5 and 6.

2. Tools, resources and standards

In this section, we give detailed insight on tools and resources used in the experiment, along with other facts of interest – basic characteristics of available annotated corpora and input-output file format standard used.

2.1. Lemmatization

At the first stage of the experiment, we had available the Croatian morphological lexicon in two forms – one was the generator of Croatian inflectional word-forms described in (Tadić, 1994) and another was the Croatian lemmatization server (Tadić, 2006). As it can be verified at hml.ffzg.hr, the server takes as input a UTF-8 encoded verticalized file. File verticalization is required because the server reads each file line as a single token which is used as a query in lemma and MSD lookup. Output is

provided in form of a text file and an equivalent HTML browser output.

Therefore, a text document could be obtained from the server containing all (lemma, token, MSD) triples and a computer program or a programming library could be created and used in our experiment as a black-box module. In order to do so, we obtained the TMT library, described in (Šilić et al., 2007) that had implemented a very fast and efficient dictionary module based on finite state automata, storing triples of wordforms, lemmas and tags into an incrementally constructed data structure. The TMT dictionary module has therefore provided us with the needed C++ object-oriented interface that we could use to get e.g. all lemmas and MSDs for a token, all MSDs for a (token, lemma) pair etc. A working lemmatization interface was now at our disposal and it could be used both as an input-output black-box and as a rule-based module to be integrated with the second-order HMM tagging paradigm at runtime.

2.2. PoS/MSD-tagging

The Croatian statistical PoS/MSD-tagger (CPT from this point on) was developed and made available as an early beta-version for purposes of validation in this experiment. Although many statistical taggers are now available in the community for scientific purposes – for example, the TnT tagger (Brants, 2000) and the HunPos tagger (Halácsy et al., 2007), a completely open-source reimplementation of TnT – and could be utilized in our research scheme, we chose to produce our own tagger so that we could alter its modus operandi by request and also be able to integrate it at will within other, larger natural language processing systems that are currently under development. CPT is written in standard C++ although some help from the HunPos development team and additional interpretation of the HunPos OCaml source itself was necessary.

At this moment, the tagger implements only a second order hidden Markov model tagging paradigm (trigram tagger), utilizing a modified version of the Viterbi algorithm (Thede and Harper, 1999), linear interpolation, successive abstraction and deleted interpolation as smoothing and default unknown word handling paradigms which are de facto standard methods, found in both TnT and HunPos. CPT presumes token emission upon reached state and is trained as a visible Markov model (VMM), i.e. on pre-tagged corpora, from which it acquires transition and emission probability matrices, as described in e.g. (Manning and Schütze, 1999).

Input and output formats of CPT are once again virtually identical to ones of TnT and HunPos; the training procedure takes a verticalized, sentence delimited corpus and creates the language model – i.e. the probability matrices – and the tagging procedure itself takes as input a verticalized, sentence delimited, non-tagged text and before mentioned language model matrices, providing an output formatted as is required of the training input: a verticalized text containing a token and MSD per line.

Since our tagger is still in beta-version, these procedures do not offer any possibility of setting the parameters to the user although the implementation of these options is taken into account. The further planned work on CPT beta is discussed in section 6 together with other possible research directions.

2.3. Annotated corpora

The Croatia Weekly 100 Kw newspaper corpus (CW100 corpus further in the text) consists of articles extracted from seven issues of the Croatia Weekly newspaper, which has been published from 1998 to 2000 by the Croatian Institute for Information and Culture. This 100 Kw corpus is a part of Croatian side of the Croatian-English parallel corpus described in detail in (Tadić, 2000). The CW100 corpus was manually tagged using the MULTEXT-East version 3 morphosyntactic specifications detailed in (Erjavec, 2004) and encoded using XCES standard (Ide et al., 2000). The corpus consists of 118529 tokens, 103161 of them being actual wordforms in 4626 sentences, tagged by 896 different MSD tags. Nouns make for a majority of corpus wordforms (30.45%), followed by verbs (14.53%) and adjectives (12.06%), which is in fact a predictable distribution for a newspaper corpus.

PoS	% corpus	Diff. MSD
Noun	30.45%	119
Verb	14.53%	62
Adjective	12.06%	284
Adposition	9.55%	9
Conjunction	6.98%	3
Pronoun	6.16%	312
Other	20.27%	107

Table 2.1 PoS distribution on the corpus

Details are provided in Table 2.1. Please note that the PoS category Other includes acronyms, punctuation, numerals etc. A more detailed insight on the CW100 corpus facts and preprocessing can be found in (Agić and Tadić, 2006).

3. Combining tagger and lemmatizer

Four different methods were considered while planning this experiment. They all shared the same preconditions for input and output file processing, as described in the previous section. We now describe in theory these methods of pairing our trigram tagger and morphological lexicon.

3.1. Tagger as a disambiguation module

The first idea is based on very high text coverage displayed by HML (more than 96.5% for newspaper texts) and the derived lemmatization interface: the text, consisting of one token per line to be tagged, could serve as lemmatizer input, the module providing as output all known MSDs given a wordform in each output line. The tagger would then be used only in context of its trained knowledge of tag sequence probabilities. A program module should therefore be derived from basic tagger function set – a module using its tag transition probabilities matrix to find the optimal tag sequence in the search space, now narrowed by using lemmatizer output instead of a generally poor lexical base acquired at training.

3.2. Lemmatizer as an unknown word handler

A second-order HMM tagger is largely dependent on its matrix of transition and emission probabilities, both of which are in our case obtained from previously annotated

corpus by a training procedure. As mentioned before, both our tagger and TnT use visible Markov model training procedures. It is well-known that a large gap occurs when comparing PoS/MSD tagging accuracies on tokens known and unknown to the tagger in terms of the training procedure. If the training procedure encounters wordforms and discovers their respective tag distributions at training, error rates for tagging these words decrease substantially compared to tagging words that were not encountered at training. Improving trigram tagger accuracy therefore often means implementing an advanced method of guessing distributions of tags for unknown wordforms based on transition probabilities and other statistical methods, e.g. deleted interpolation, suffix tries and successive abstraction. Namely, TnT tagger implements all the methods listed above. However, TnT can never link a wordform to an unknown tag, i.e. a tag that was not previously acquired by the training procedure. We based our second method of pairing HML and CPT on that fact alone: it should be investigated whether HML – as a large base of wordforms and associated lemmas and MSDs – could serve as unknown word handling module for the tagger at runtime.

In more detail, the idea builds on (Halacsy et al., 2006) and (Halacsy et al., 2007) and is basically a simple logical extension of the unknown word handling paradigm using suffix tries and successive abstraction (Samuelsson, 1993). Trigram tagger such as TnT uses algorithms to disambiguate between tags in tag lists provided by emission probability matrix for a known wordform. Upon encountering an unseen wordform, such a list cannot be found in the matrix and must be constructed from another distribution, e.g. based on wordform suffixes and implemented in the suffix trie. Successive abstraction contributes by iteratively choosing a more general distribution, i.e. distribution for shorter suffixes until a distribution of tags is finally assigned. This results in large and consequently low-quality distributions for unknown wordforms, resulting in lower tagging accuracy. Taking high coverage of HML into consideration, idea was to choose from the suffix trie distribution only those MSDs on which both HML and trie intersect, falling back to suffix tries and successive abstraction alone when both lemmatizer and tagger fail to recognize the wordform. By this proposition, we utilize (wordform, tag) probabilities as given by the suffix trie and yet choose only meaningful (wordform, tag) pairs, i.e. pairs confirmed by applying the morphological lexicon.

3.3. Lemmatizer as a preprocessing module

In this method, we train a trigram tagger using the VMM training method and obtain matrices of transition and emission probabilities. The latter one, emission probability matrix, links each of the tokens found in the training corpus to its associated tags and counts, as is shown in Figure 3.1. The figure provides an insight on similarities and differences of storing language specific knowledge of tagger and lemmatizer.

It was obvious that lemmatizer and lexicons acquired by training share properties and therefore it was possible to create a lemmatizer-derived module for error detection and correction on the acquired lexicon used internally by the tagger. From another perspective, lemmatizer and acquired lexicon could also be merged into a single resource by a well-defined merging procedure.

```
%% ...
ime 26 Ncnsa 24 Ncnsn 2
imena 8 Ncnpa 1 Ncnpg 1 Ncnpn 3 Ncnsg 3
imenima 2 Ncnpd 1 Ncnpi 1
imenom 3 Ncnsi 3
imenovan 2 Vmps-smp 2
imenovana 1 Vmps-sfp 1
imenovanja 3 Ncnpg 2 Ncnsg 1
imenovanje 1 Ncnsv 1
imenovanjem 1 Ncnsi 1
imenovanju 4 Ncnsl 4
%% ...
```

```
%% ...
ime ime Ncnsa ime Ncnsn ime Ncnsv
imenima ime Ncnpd ime Ncnpi ime Ncnpl
imenom ime Ncnsi
%% ...
```

Figure 3.1 Emission probability matrix file and lemmatizer output comparison

3.4. Lemmatizer as a postprocessing module

Similar to using lemmatizer's language knowledge before tagging, it could also be used after tagging. Output of the tagger could then be examined in the following manner:

1. Input is provided both to tagger and lemmatizer, each of them giving an output.
2. The two outputs are then compared, leading to several possibilities and corresponding actions:
   a. Both tagger and lemmatizer give an answer. Lemmatizer gives an unambiguous answer identical to the one provided by the tagger. No action is required.
   b. Both tagger and lemmatizer give an answer. Lemmatizer gives an unambiguous answer and it is different from the one provided by the tagger. Action is required and we choose to believe the lemmatizer as a manually assembled and therefore preferred source of language specific knowledge.
   c. Both tagger and lemmatizer give an answer. Lemmatizer gives an ambiguous answer, i.e. a sequence of tags. One of the tags in the sequence is identical to taggers answer. We keep the tagger's answer, being now confirmed by the lemmatizer.
   d. Both tagger and lemmatizer give an answer. Lemmatizer gives an ambiguous answer and none of the tags in the sequence matches the one provided by the tagger. A module should be written that takes into account the sequence provided by the lemmatizer and does re-tagging in a limited window of tokens in order to provide the correct answer. Basically, we define a window sized 3 tokens/tags and centered on the ambiguous token, lookup the most frequent of various trigram combinations available for the window (these are given by lemmatizer!) in transition probability matrix of the tagger and assign this trigram to the

window, disambiguating the output. By this we bypass tagger knowledge and once again choose to prefer lemmatizer output.

e. Tagger provides an answer, but token is unknown to the lemmatizer. We keep the tagger's answer, this being the only possible course of action.
f. Tagger does not provide an answer and lemmatizer does. If its answer is unambiguous, we assign it to the token. If it is ambiguous, we apply the procedure described in option 2d.

3. Final output produced by the merge is then investigated by the evaluation framework.

It should by all means be noted that each of the presented paradigms had to undergo a theoretical debate and possibly – if considered to be a reasonable course of action – a full sequence of tests described in section 4 in order to be accepted or rejected for introducing overall improvement of tagging accuracy or creating additional noise, respectively. Details are given in the following sections.

4. Evaluation method

As a testing paradigm, we chose the F-measure framework for evaluation on specific PoS and general accuracy for overall tagging performance. Firstly, we provide a comparison of CPT and TnT: overall PoS vs. MSD accuracy and also F-measures on nouns, pronouns and adjectives, proven to be the most difficult categories in (Agić and Tadić, 2006). We then discuss the proposed tagger-lexicon combinations and provide the measures – overall accuracy and F1-scores for those methods judged as suitable and meaningful at the time of conducting the investigation.

Each test consists of two parts: the worst-case scenario and the default scenario. Worst-case is a standard tagging accuracy measure scenario created by taking 90% of the CW100 corpus sentences for training and leaving the other 10% for testing; therefore, in a way, this scenario guarantees the highest number of unknown words to be found at runtime. The default scenario chooses 90% of sentences from the CW100 pool for training and then 10% for testing from the same pool, making it possible for sentences to overlap in these sets. The default scenario is by definition not a standard measure scenario and was introduced in order to respect the nature of random occurrences in languages, leaving a possibility (highly improbable) of tagger encountering identical sentences at training and at runtime.

Note that we do not include testing scenarios debating on training set size as a variable: in this test, we consider improving overall tagging accuracy and not investigating HMM tagging paradigm specifics as in (Agić and Tadić, 2006), being that conclusions on this specific topic were already provided by that test environment.

5. Results

The first set of results we present is from the set of tests evaluating overall tagging accuracy of CPT on full MULTEXT East v3 MSD and on PoS information (the first letter of the MSD tag, not comparable to English PoS of e.g. English Penn Treebank) only. Acquired results are displayed in Table 5.1.

		TnT		CPT	
		MSD	**PoS**	**MSD**	**PoS**
Worst case	Overall	86.05%	96.53%	86.05%	**96.84%**
	Known	89.05%	98.29%	**89.26%**	**98.42%**
	Unknown	**66.04%**	86.02%	65.95%	**87.29%**
	Corp. unk.	**13.07%**	14.40%	**13.77%**	14.11%
Default case	Overall	**97.54%**	98.51%	97.51%	**99.31%**
	Known	98.04%	98.74%	**98.05%**	**99.43%**
	Unknown	62.21%	83.11%	**63.75%**	**88.39%**
	Corp. unk.	**1.42%**	1.51%	1.59%	**1.13%**

Table 5.1 Overall tagging accuracy on MSD and PoS

It could be stated from this table that results on TnT and CPT are virtually identical and the differences exist merely because testing environment – mainly the number of unknown words – was variable. It is however quite apparent that CPT outperformed TnT on part-of-speech, especially regarding unknown tokens, but this should be taken with caution as well, being that CPT dealt with fewer unknowns in that specific test.

Second testing case considers combining CPT and the inflectional lexicon. Before presenting the results and in order to interpret them correctly, it should be stated that only two of the four initially proposed merging methods were chosen to proceed to the practical testing session: method (3.2) using the lemmatizer as an unknown world handler (3.4) using the lemmatizer as a postprocessing module to resolve potential errors produced by the tagger. We rejected applying (3.1) tagger as a disambiguation module for lemmatizer output because it would be costly to develop yet another tagger-derived procedure to handle transition probabilities only and because this procedure would, in fact, do nothing different than a common HMM-based tagger does with its own acquired lexicon: disambiguates its ambiguous entries upon encountering them in the text and applying the transition matrix and handling procedures on unknown words.

The idea of lemmatizer as preprocessing module (3.3) was also rejected, mainly because we were unable to define precisely how to merge its database to the one acquired by tagger at training procedure. Being that tagger assigns each entry with a number of its occurrences overall and number of occurrences under various MSDs and, in order to apply the lemmatizer, we would have to assign these numbers so the tagger could understand the new entries – if we assign all to 1, it does not contribute and is redundant and if we assign any other number, we are in fact altering the tagging procedure outcome in such a manner that is not in any way bound by the language model, i.e. the training corpus. Therefore, we proceed with considering proposed cases (3.2) and (3.4) only.

We have also omitted PoS results from this testing case because TnT and CPT are both able to achieve an accuracy over 95% without additional modules so we were focused in improving MSD accuracy, keeping in mind that most errors do not occur on PoS but on sub-PoS levels resolvable by the lexicon. Details are provided by Table 5.2.

The first apparent conclusion is that method (3.4) that cleans up the errors on tagger output has failed and that it has failed on unknown words – where we could have expected it (or hoped for it) to perform better. The reason

is, on the other hand, quite obvious: the tagger applies a tag to an unknown word using transition probabilities and smoothing procedures that are proven to operate quite satisfactory in TnT, HunPos and now CPT; when the postprocessing lemmatizer-based module encounters a word tagged as unknown – this word is rarely unambiguous on HML – therefore, a resolution module using transition probabilities had to be applied quite frequently and this module clearly and expectedly does not outperform default procedures (suffix tries, successive abstraction, deleted interpolation).

		TnT	CPT & 3.2	CPT & 3.4
Worst case	Overall	**86.05%**	85.58%	83.94%
	Known	**89.05%**	88.84%	88.18%
	Unknown	**66.04%**	65.13%	57.38%
	Corp. unk.	13.07%	13.77%	13.77%
Default case	Overall	97.54%	**97.97%**	97.88%
	Known	98.04%	**98.53%**	98.51%
	Unknown	62.21%	**63.49%**	59.40%
	Corp. unk.	01.42%	01.59%	01.59%

Table 5.2 Tagging accuracy with (3.2) unknown word handler and (3.4) postprocessing

Based on other, positive part of Table 5.2, we could end the section by stating that CPT, when combined with the HML-based lemmatizer in such a manner that the lemmatizer provides morphological cues to the tagger upon encountering unknown words, outperforms TnT by a narrow margin on the default MSD test case. However, a more sincere and exact statement – taking in regard all section 5 tables – would be that both TnT and CPT share the same functional dependency regarding the number of unknown words they encounter in the tagging procedure. That is, CPT outperforms TnT when less unknown tokens occur for him at runtime and vice versa, the lemmatizer contributing for around 1.3% improvement on unknown words. We could therefore argue that our beta-version of CPT tagger performs as well as TnT tagger – and that we have succeeded in implementing a state-of-the-art solution for tagging large-scale corpora of Croatian – given the test environment we had at hands, its drawbacks noted and hereby included.

In Table 5.3 we present results of evaluation broken down by three most difficult PoS categories: adjectives, nouns and pronouns. Data and analysis is given for PoS information only, as mentioned before.

		Adjective	Noun	Pronoun
TnT	Worst case	64.61%	82.10%	76.62%
	Default case	94.73%	96.89%	**97.11%**
CPT	Worst case	65.31%	80.85%	74.62%
	Default case	**95.86%**	**97.40%**	95.88%
CPT & 3.2	Worst case	**66.15%**	**81.71%**	**75.07%**
	Default case	95.06%	96.79%	95.82%

Table 5.3 Tagging accuracy with adjectives, nouns and pronouns

It can be clearly noticed that suggested combination mode (3.2) outperforms both TnT and CPT in the worst case scenario on all parts of speech since it has the support of HML when handling unknown words, that do occur frequently in this scenario. In the default case scenario, results are as expected, more even and inconclusive – default CPT actually outperforms lemmatizer combination because a unknown tokens were found in small numbers in the test sets, much too small for the lemmatizer to contribute significantly to overall tagging accuracy.

6. Conclusions and future work

In this contribution we have presented a beta-version of statistical PoS/MSD tagger for Croatian and proposed combining it with a large scale inflectional lexicon of Croatian, thus deriving a hybrid system for high-precision tagging of Croatian corpora. We have presented several possible types of combinations, tested and evaluated them using the F-measure evaluation framework. CPT provided results virtually identical to TnT – they differ only in hundredths of percentage in both directions in different evaluating conditions. This way we have shown that CPT functions at the level of state-of-the-art regarding HMM-based trigram tagging.

Our future directions for improvement of this system could and probably will fall into several different research pathways.

The first of them could be analyzing tagging accuracy on morphological (sub-part-of-speech) features in more detail and fine-tuning the tagger accordingly.

Various parameterization options could also be provided at tagger runtime. Such options could include parameters for unigram, bigram and trigram preference or implementing token emissions depending on previously encountered sequences (multiword unit dependencies).

Fine-tuned rule-based modules for Croatian language specifics could also be considered and applied before or after the statistical procedure. Another option would be the integration of lemmatizer into tagger as they have been programmed as separate modules.

The next direction would be to build a full lemmatizer which, unlike solution presented in this paper, gives a fully disambiguated output relying on the results of the tagger. Selection of proper lemmas from sets of possible ones would be done on the basis of tagger output, once again fine-tuning levels of confidence between tagger and lemmatizer similar to section 3 of the paper.

It should also be noted that (Agić and Tadić, 2008) takes into account an entirely different approach, putting an emphasis on corpora development. Namely, all the methods presented in previous sections are made exclusively for handling unknown word occurrences and they all required lots of time and human effort to be implemented. On the other hand, manual corpora development – although also requiring time and effort – is by definition less demanding and at the same time quite reasonable course of action: larger, better and more diverse corpora are always a necessity – a necessity that implicitly resolves unknown word issues as well. Courses of action could therefore be argued; we decided to take both throughout our future work in order to additionally improve tagging accuracy.

7. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under

the grants No. 130-1300646-1776, 130-1300646-0645 and 036-1300646-1986.

8. References

Agić, Ž., Tadić, M. (2006). Evaluating Morphosyntactic Tagging of Croatian Texts. In Proceedings of the Fifth International Conference on Language Resources and Evaluation. ELRA, Genoa – Paris 2006.

Agić, Ž., Tadić, M. (2008). Investigating Language Independence in HMM PoS/MSD-Tagging. In Proceedings of the 30th International Conference on Information Technology Interfaces. Cavtat, Croatia, 2008, pp. 657-662.

Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing. Seattle, Washington 2000.

Erjavec, T. (2004). Multext-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In Proceedings of the Fourth International Conference on Language Resources and Evaluation. ELRA, Lisbon-Paris 2004, pp. 1535-1538.

Halácsy, P., Kornai, A., Oravecz, C. (2007). HunPos - an open source trigram tagger. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. Association for Computational Linguistics, Prague, Czech Republic, pp. 209-212.

Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D. (2006). Using a morphological analyzer in high precision POS tagging of Hungarian. In Proceedings of 5th Conference on Language Resources and Evaluation (LREC). ELRA, pp. 2245-2248.

Ide, N., Bonhomme, P., Romary, L., (2000). An XML-based Encoding Standard for Linguistic Corpora. In Proceedings of the Second International Conference on Language Resources and Evaluation, pp. 825-830. (see also at http://www.xces.org).

Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing, The MIT Press, 1999.

Samuelsson, C. (1993). Morphological tagging based entirely on Bayesian inference. 9th Nordic Conference on Computational Linguistics NODALIDA-93. Stockholm University, Stockholm, Sweden.

Šilić, A., Šarić, F., Dalbelo Bašić, B., Šnajder, J. (2007). TMT: Object-Oriented Text Classification Library. Proceedings of the 29th International Conference on Information Technology Interfaces. SRCE, Zagreb, 2007. pp. 559-566.

Tadić, M. (1994). Računalna obrada morfologije hrvatskoga književnog jezika. Doctoral thesis. Faculty of Humanities and Social Sciences, University of Zagreb, 1994.

Tadić, M. (2000). Building the Croatian-English Parallel Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation. ELRA, Paris-Athens 2000, pp. 523-530.

Tadić, M., Fulgosi, S. (2003). Building the Croatian Morphological Lexicon. In Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages. Budapest 2003, ACL, pp. 41-46.

Tadić, M. (2006). Croatian Lemmatization Server. Formal Approaches to south Slavic and Balkan Languages. Bulgarian Academy of Sciences, Sofia, 2006. pp. 140-146.

Thede, S., Harper, M. (1999). A second-order Hidden Markov Model for part-of-speech tagging. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics, pp. 175-182.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (Eds.) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33.

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. In Proceedings of the 4th LREC Conference. Lisbon, Portugal, pp. 39-42.

Indeks avtorjev / Author index

Agić Željko....116
Arhar Špela....54
Bajec Boštjan....70
Baroni Marco....7
Beguš Ana....60
Bucik Valentin....70
Čavar Damir....80
Dorofeeva Uliana....10
Dovedan Zdravko....116
Dukič Davorin....60
Erjavec Tomaž....33, 37, 49, 65, 110
Fišer Darja....37
Gajšek Rok....70
Grašič Matej....20
Grčar Miha....110
Gros Žganec Jerneja....10
Hartrumpf Sven....92
Helbig Hermann....92
Holozan Peter....43
Jakopin Primož....104
Jazbec Ivo-Pavao....80
Kačič Zdravko....16, 20
Koderman Miha....60
Komidar Luka....70
Končar Bizjak Aleksandra....104
Kos Marko....16, 20
Kotnik Bojan....16
Krek Simon....49
Ledinek Nina....54
Maučec Sepesy Mirjam....16
Mihelič Aleš....10
Mihelič France....70
Mikolič Vesna....60
Nikolov Nicolas....75
Pavešić Nikola....10
Pavlović-Lažetić Gordana....86
Peterlin Pisanski Agnes....29
Podlesek Anja....70
Rotovnik Tomaž....16
Runjaić Siniša....80
Rupnik Jan....110
Sangawa Hmeljak Kristina....33
Satev Vesna....75
Schultz Tanja....9
Sočan Gregor....70
Štruc Vitomir....70
Tadić Marko....116
Tomašević Jelena....86
Verdonik Darinka....25, 29
Vičič Jernej....98
Vintar Špela....65
vor der Br¨uck Tim....92
Žganec Mario....10
Žgank Andrej....16, 29