

Zbornik B 7. mednarodne multikonference
Proceedings B of the 7th International Multiconference

INFORMACIJSKA DRUŽBA IS 2004

INFORMATION SOCIETY IS 2004

Jezikovne tehnologije

Language Technologies

Uredila / Edited by

Tomaž Erjavec, Jerneja Žganec Gros

<http://nl.ijs.si/isjt04/>

<http://is.ijs.si>

9. do 15. oktober 2004 / October 9-15th 2004
Ljubljana, Slovenia

PREDGOVOR H KONFERENCI JEZIKOVNE TEHNOLOGIJE

Multikonferenco Informacijska družba (IS) sestavljajo naslednje samostojne konference:

- BIOMA – Biološko osnovane optimizacijske metode in aplikacije
- Inteligentni sistemi
- Izkopavanje znanja in podatkovna skladišča
- Jezikovne tehnologije
- Kognitivna znanost
- Kompleksni sistemi v e-poslovanju
- Razvoj in prenovitev informacijskih sistemov
- Sodelovanje in informacijska družba
- Teoretično računalništvo
- Upravljanje v informacijski družbi
- Vzgoja in izobraževanje v informacijski družbi

V okviru multikonference vsako drugo leto dr. Tomaž Erjavec in dr. Jerneja Žganec Gros organizirata samostojno konferenco Jezikovne tehnologije. Zato se jima in slovenski strokovni jezikovni skupnosti iskreno zahvaljujemo. Ob vsaki konferenci ugotavljamo, da slovenski jezik ne izumira pod pritiski globalizacije, ampak da živi v sožitju z največjimi svetovnimi jeziki. Ravno letos pa se je s priključitvijo Evropski uniji slovenski jezik pridružil drugim uradno priznanim jezikom EU.

Globalizacija poleg številnih prednosti vpeljuje tudi poenoteno kulturo in jezik. Internet je s svojo univerzalnostjo in uporabo angleščine močan vzvod poenotenja. Znanstveniki namenjamo posebno skrb namenjamo temu, da ne bi posamezni jeziki in kulturne izumirale, saj predstavljajo pomemben del dediščine celotnega človeštva. Naša naloga je, da slovenski jezik postavimo ob bok svetovnim posebej na internetu in posebej preko informacijskih storitev informacijske družbe. Konferenca Jezikovne tehnologije je eden izmed ključnih dogodkov, ki promovirajo in razvijajo slovenski jezik in kulturo v informacijski družbi.

Na vsaki konferenci ugotavljamo nove dosežke. Letos med najuspešnejšimi najdemo sistema za prevajanje med slovenščino in angleščino, razne jezikovne izdelke in celo simultano prevajanje. Marsikateri izmed omenjenih sistemov je več kot konkurenčen tujim, za slovenščino samo prilagojenim sistemom.

V imenu organizatorjev se zahvaljujemo vsem udeležencem za sodelovanje, podporo in zanimanje za omenjene tematike.

Matjaž Gams

Predsednik organizacijskega odbora IS'04

FOREWORD TO LANGUAGE TECHNOLOGIES

The Information Society (IS) multi-conference 2004 consists of the following conferences:

- BIOMA – Bioinspired Optimization Methods and their Applications
- Cognitive Sciences
- Collaboration and Information Society
- CSeB - Complex Systems in e-Business
- Data-Mining and Data-Warehouses
- Development and Reengineering of Information Systems
- Education in Information Society
- Intelligent Systems
- Language Technologies
- Management in Information Society
- Theoretical Computer Science

Recently, each odd year the “Language technologies” conference is organized by Dr. Tomaž Erjavec and Dr. Jerneja Žganec Gros. For that, not only they, but all the language community in Slovenia deserves all thanks. The significant progress reported in the last two years not only shows that the Slovenian language is not becoming overwhelmed by the globalization processes, it also shows that it can go along the developed world languages. Furthermore, from this year on Slovenia is one of the established and respected European Union languages.

Although information society merges people into one global society speaking English and excessively using the Internet, we must devote special care and energy not to let languages and cultures die out, but to preserve it as part of the heritage of human civilisation. Hence, language technologies promoting the Slovenian language into the Internet age are one of the key ingredients of Slovenian progress. Events such as this conference elaborate and promote Slovenian language and culture at the same time.

Each language technology conference presents some new achievements. In 2004 we find the most important progress in several areas of Slovenian-English translation and in advanced speech-technology products including symoultaneous translation. Several of our programs are more than competitive to the foreign ones, especially when the Slovenian language is concerned.

In the name of the whole conference we would like to thank all the participants for their cooperation, contribution, support or just for being interested in this event.

Matjaž Gams

President of the IS'04 Organizing Committee

Konferenca JEZIKOVNE TEHNOLOGIJE 2004

PREDGOVOR

V pričujočem zborniku so objavljeni prispevki s konference “Jezikovne tehnologije”, ki je potekala 13. in 14. oktobra 2004 v Ljubljani, v okviru multikonference “Informacijska družba”, IS’2004. 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 slovenske konference o jezikovnih tehnologijah.

Zbornik vsebuje 23 prispevkov, ki obravnavajo široko paleto raziskav in aplikacij. Prispevki so približno enakovredno razdeljeni med tiste, ki obravnavajo govorne tehnologije, in take, ki se ukvarjajo z besedilom. Prispevki so po obravnavani tematiki izrazito raznovrstni, je pa, glede na prejšnje konference opazen premik k večjezikovnim aplikacijam.

Organizatorji bi se radi zahvalili vsem, ki so prispevali k uspehu konference: vabljenim predavateljem, avtorjem prispevkov in predstavitev v demo sekciji, programskemu odboru za recenzentsko delo ter organizatorjem IS’2004.

Tomaž Erjavec, Jerneja Žganec Gros

Ljubljana, oktober 2004.

Language Technologies Conference 2004

Preface

These proceedings contain the contributions to the conference “Language Technologies”, which took place on October 13th and 14th 2004 in Ljubljana, in the scope of the Information Society Multiconference, IS’2004. The conference was aimed at the members of the Slovenian Language Technology Society and others interested in the field, as a forum where they could present their work in the last two years, which have passed since the previous Slovenian conference on Language Technologies.

These proceedings contain 23 contributions, which present a wide variety of research and application topics. The contributions are about equally divided into those that address speech technologies and those that deal with text. The papers are quite diverse as regards their subject matter but, in comparison to the previous Slovene LT conference, we can note a shift to multilingual applications.

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 2004.

Tomaž Erjavec, Jerneja Žganec Gros

Ljubljana, October 2004.

Konferenca JEZIKOVNE TEHNOLOGIJE

Language Technologies Conference

Programski odbor / Programme Committee

<i>dr. Aleš Dobnikar</i>	Center Vlade za Informatiko
<i>dr. Simon Dobrišek</i>	Fakulteta za elektrotehniko, Univerza v Ljubljani
<i>doc.dr. Tomaž Erjavec</i>	Institut "Jožef Stefan"
<i>doc.dr. Vojko Gorjanc</i>	Filozofska fakulteta, Univerza v Ljubljani
<i>doc.dr. Primož Jakopin</i>	Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU
<i>prof.dr. Zdravko Kačič</i>	Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
<i>Jaro Lajovic</i>	Slovensko društvo za jezikovne tehnologije
<i>prof.dr. France Mihelič</i>	Fakulteta za elektrotehniko, Univerza v Ljubljani
<i>dr. Primož Peterlin</i>	Medicinska fakulteta, Univerza v Ljubljani
<i>prof.dr. Marko Stabej</i>	Filozofska fakulteta, Univerza v Ljubljani
<i>prof.dr. Rastislav Šuštaršič</i>	Filozofska fakulteta, Univerza v Ljubljani
<i>doc.dr. Špela Vintar</i>	Filozofska fakulteta, Univerza v Ljubljani
<i>dr. Jerneja Žganec Gros</i>	Alpineon, d.o.o.

Organizacijski odbor / Organising Committee

<i>Tomaž Erjavec</i>	Institut "Jožef Stefan"
<i>Jerneja Žganec Gros</i>	Alpineon, d.o.o.
<i>Jaro Lajovic</i>	Slovensko društvo za jezikovne tehnologije
<i>Špela Vintar</i>	Filozofska fakulteta, Univerza v Ljubljani

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, Israel
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

Organizing Committee

prof. dr. Matjaž Gams, chair
mag. Aleksander Pivk
dr. Damjan Demšar
dr. Gregor Papa
Lili Lasič
Mili Bauer, dipl. ekon
Mitja Lasič
Mitja Luštrek, dipl. ing
Tea Robič, dipl. ing

Programme Committee

doc. dr. Cene Bavec, chair
dr. Tomaž Kalin,
dr. Jozsef Györkös,
prof. dr. Tadej Bajd
mag. Jaroslav Berce
prof. dr. Marko Bohanec
prof. dr. Ivan Bratko
dr. Andrej Brodnik
dr. Dušan Caf
prof. dr. Saša Divjak
dr. Tomaž Erjavec
doc. dr. Bogdan Filipič
prof. dr. Matjaž Gams
Marko Grobelnik
prof. dr. Nikola Guid
dr. Marjan Heričko
prof. dr. Borka Jerman Blažič Džonova
prof. dr. Gorazd Kandus
prof. dr. Marjan Krisper
mag. Andrej Kuščer
prof. dr. Jadran Lenarčič
doc. dr. Borut Likar
dr. Dunja Mladenič
prof. dr. Franc Novak
prof. dr. Marjan Pivka
prof. dr. Vladislav Rajkovič
asist. dr. Grega Repovš
prof. dr. Ivan Rozman
prof. dr. Franc Solina
prof. dr. Stanko Strmčnik
dr. Tomaž Šef
dr. Jurij Šilc
prof. dr. Jurij Tasič
prof. dr. Denis Trček
prof. dr. Andrej Ule
doc. dr. Tanja Urbančič
prof. dr. Boštjan Vilfan
prof. dr. David B. Vodusek
prof. dr. Baldomir Zajc
prof. dr. Blaž Zupan

There are 11 independent conferences and 5 separate proceedings:

- Proceedings A: Cognitive Science
- Proceedings B: Language Technologies
- Proceedings C: Complex Systems in e-Business, Data-Mining and Data-Warehouses, Intelligent Systems
- Proceedings D: Collaboration and Information Society, Development and Reengineering of Information Systems, Education in Information Society, Management in Information Society, Theoretical Computer Science
- Proceedings E: BIOMA - Bioinspired Optimization Methods and their Applications

Kazalo / Table of Contents

VABLJENI PRISPEVKI / INVITED CONTRIBUTIONS

Anton Batliner:

Automatic classification of emotional states: purpose, possibilities, prospects 1

Ralf Steinberger, Bruno Pouliquen, Camelia Ignat:

Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications 2

PRISPEVKI / CONTRIBUTIONS

Boštjan Jerko:

Samodejno indeksiranje posnetkov 13

Jure Leskovec, Marko Grobelnik, Nataša Milic-Frayling:

Learning Sub-structures of Document Semantic Graphs for Document Summarization 18

Nina Modrijan:

Korpusni pristop pri analizi naslavljanja 26

Peter Jurgec:

Natančnost odčitavanja formantov pri digitalnem spektrografiranju na podlagi LPC-analize 34

Darinka Verdonik, Matej Rojc:

Jezikovni viri projekta LC-STAR 42

Birte Lönneker, Primož Jakopin:

Checking POSBeseda, a Part-of-Speech tagged Slovenian corpus 48

Birte Lönneker, Katarina Rozman:

Online SLO-DE-SLO: Spletni slovensko-nemški in nemško-slovenski slovar 56

Darja Fišer, Špela Vintar:

Uvajanje prevajalskega namizja Trados v delovno okolje prevajalske agencije 64

Janez Žibert, France Mihelič:

Development, Evaluation and Automatic Segmentation of Slovenian Broadcast News Speech Database 72

Melita Hajdinjak, France Mihelič:

Eksperiment Čarovnik iz Oza 79

Tomaž Erjavec, Matija Ogrin:

E-Slomšek: elektronska znanstvenokritična izdaja retorske proze 19. stoletja po standardu XML TEI 87

Andrej Žgank, Tomaž Rotovnik, Darinka Verdonik, Zdravko Kačič:

Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora 94

Boštjan Vesnicer, France Mihelič, Nikola Pavešič:

Vrednotenje na prikritih Markovovih modelih temelječega sistema za umetno tvorjenje slovenskega govora 98

Cvetana Krstev, Duško Vitas:

Restructuring Lemma in a Dictionary of Serbian 103

<i>Cvetana Krstev, Duško Vitas, Tomaž Erjavec:</i> MULTEXT-East Resources for Serbian	108
<i>Tea Robič, Tomaž Šef:</i> Vpliv izbire atributov na točnost naglaševanja nepoznanih slovenskih besed	115
<i>Matjaž Rodman:</i> Problemi pri vključevanju ekspertnega lingvističnega znanja v akustično modeliranje	119
RAZŠIRJENI POVZETKI / EXTENDED ABSTRACTS	
<i>Peter Weiss:</i> Vnašalni sistem ZRCola	124
<i>Miro Romih:</i> Uporaba programskega jezika K2.0 v sistemih dialoga	126
<i>Peter Holozan:</i> Uporaba glagolskih predlog pri strojnem prevajanju	128
<i>Jerneja Žganec Gros, Aleš Mihelič, Mario Žganec:</i> AlpSynth: Unit Selection Slovenian Text-to-Speech Synthesis	129
Kazalo po avtorjih / Author Index	131