

---

**IS 2000**  
**Information Society multi-conference**

Zbornik konference / Proceedings of the Conference

**JEZIKOVNE TEHNOLOGIJE**  
**Language Technologies**

---

Uredila / Edited by  
Tomaž Erjavec, Jerneja Gros

17. – 18. oktober 2000  
Ljubljana, Slovenija

# Information Society Multi-Conference 2000

## Information Society and Language Technologies

The Information Society (IS) multi-conference 2000 consists of the following conferences:

- Cognitive science
- Information society and governmental services
- Media in information society
- Education in information society
- Data warehouses and data mining
- Development and reengineering of information systems
- Language technologies.

Although there is an abundance of various conferences in Slovenia and worldwide, there seems to be a need for a multi-conference promoting the introduction of information society and the accompanying scientific, engineering and human activities. Language technologies are an essential part of information society in any country. Although information society merges people into one global society speaking English, we must devote special care and energy not to let languages and cultures die out. This is especially true for small languages like Slovenian. One single world without the richness of human diversities would be a boring place to live.

As the proceedings of this conference show, there is in Slovenia lively research into various aspects of this field. Still, there should be a governmental initiative or at least better support for such endeavours. Among major projects in future is surely a system for translation between English and Slovene. Translation is just one of many activities in the area of language-dependent information technologies which are of major importance for Slovenia. Lots of interesting projects are waiting ahead of us!

In the name of the whole conference, the IS organizing committee would like to thank Dr. Tomaž Erjavec and Dr. Jerneja Gros for organising the Language Technologies conference, and the contributors and participants for making it a success.

Matjaž Gams

President of the IS Organizing Committee

# Konferenca JEZIKOVNE TEHNOLOGIJE

## Language Technologies Conference

### **PREDGOVOR**

V pričujočem zborniku so objavljeni prispevki s konference “Jezikovne tehnologije”, ki je potekala 17. in 18. oktobra 2000 v Ljubljani, v okviru multikonference “Informacijska družba”, IS’2000. 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 prve slovenske konference o jezikovnih tehnologijah.

Zbornik vsebuje 25 prispevkov, ki obravnavajo široko paleto raziskav in aplikacij; posebej izstopa veliko število prispevkov s področja govornih tehnologij, dobro zastopani pa so tudi prispevki o izdelavi in uporabi korpusov in drugih jezikovnih virov. Članki podajajo večinoma raziskovalne rezultate, nekaj pa je tudi zamisli za razvoj področja v prihodnje.

Organizatorja bi se rada zahvalila vsem, ki so prispevali k uspehu konference: vabljenim predavateljem in avtorjem prispevkov, programskemu odboru za recenzentsko delo ter organizatorjem IS’2000.

### **Preface**

These proceedings contain the contributions to the conference “Language Technologies”, which took place on the 17th and 18th October 2000 in Ljubljana, in the scope of the Information Society Multi-conference, IS’2000. The conference was aimed at the members of the Slovenian Language Technology Society and others interested in the field, as a forum where they can present their work in the last two years which have passed since the first Slovenian conference on language technologies.

These proceedings contain 25 contributions, which present a wide variety of research and application topics; especially numerous are contributions dealing with speech technologies and papers about the creation and usage of corpora and other language resources. The papers present mainly research results although some also deal with plans or visions for future developments of the field.

The organisers would like to thank the many people who contributed to the success of the conference: the invited speakers and the authors of contributions, the programme committee of the conference and the organising committee of IS 2000.

Tomaž Erjavec, Jerneja Gros

Ljubljana, October 2000.

**Konferenca JEZIKOVNE TEHNOLOGIJE**  
**Language Technologies Conference**

**Programski odbor / Programme Committee**

<i>dr. Aleš Dobnikar</i>	Center Vlade za Informatiko
<i>dr. Tomaž Erjavec</i>	Odsek za inteligentne sisteme, Institut "Jožef Stefan"
<i>Vojko Gorjanc</i>	Oddelek za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
<i>dr. Jerneja Gros</i>	Laboratorij za umetno zaznavanje, Fakulteta za elektrotehniko, Univerza v Ljubljani
<i>dr. Primož Jakopin</i>	Inštitut za slovenski jezik Frana Ramovša ZRC SAZU
<i>prof.dr. Zdravko Kačič</i>	Inštitut za elektroniko, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
<i>Aleš Košir</i>	Hermes SoftLab
<i>Jaro Lajovic</i>	Slovensko društvo za jezikovne tehnologije
<i>prof.dr. France Mihelič</i>	Laboratorij za umetno zaznavanje, Fakulteta za elektrotehniko, Univerza v Ljubljani
<i>Primož Peterlin</i>	Institut za biofiziko, Medicinska fakulteta, Univerza v Ljubljani
<i>doc.dr. Marko Stabej</i>	Oddelek za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
<i>prof.dr. Rastislav Šuštaršič</i>	Oddelek za anglistiko, Filozofska fakulteta, Univerza v Ljubljani
<i>Hotimir Tivadar</i>	Oddelek za slovanske jezike in književnosti, Filozofska fakulteta, Univerza v Ljubljani
<i>Špela Vintar</i>	Oddelek za prevajanje in tolmačenje, Filozofska fakulteta, Univerza v Ljubljani

**Organizacijski odbor / Organising Committee**

<i>dr. Tomaž Erjavec</i>	Institut "Jožef Stefan"
<i>dr. Jerneja Gros</i>	Fakulteta za elektrotehniko, Univerza v Ljubljani

# Information Society Multi-Conference 2000

## **Program Committee Chairman**

dr. Cene Bavec

## **Program Committee Vice chairman**

prof. dr. Matjaž Gams

## **Program Committee**

prof. dr. Tadej Bajd, mag. Jaroslav Berce, prof. dr. Ivan Bratko, dr. Dušan Caf, prof. dr. Saša Divjak, dr. Tomaž Erjavec, Marko Grobelnik, prof. dr. Nikola Guid, prof. dr. Borka Jerman Blažič Džonova, doc. dr. Gorazd Kandus, doc. dr. Marjan Krisper, mag. Andrej Kuščer, prof. dr. Jadran Lenarčič, dr. Dunja Mladenič, dr. Franc Novak, prof. dr. Marjan Pivka, prof. dr. Vladislav Rajkovič, prof. dr. Ivan Rozman, Niko Schlamberger, dipl. ing., prof. dr. Franc Solina, prof. dr. Stanko Strmčnik, prof. dr. Jurij Tasič, prof. dr. Andrej Ule, dr. Tanja Urbančič, dr. Blaž Zupan

## **Organizing Committee Chairman**

prof. dr. Matjaž Gams

## **Organizing Committee Vice chairman**

dr. Tanja Urbančič

## **Organizational Committee**

Damjan Demšar, Msc, dr. Viljem Križman, dr. Dunja Mladenič, Aleksander Pivk, BSc, Mili Remetič, Maja Škrjanc, BSc

## **International Program Committee**

Vladimir Bajic, South Africa, Heiner Benking, Germany, Se Woo Cheon, Korea, Howie Firth, Scotland, Vladimir Fomichov, Russia, Alfred Inselberg, Izrael, 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, Japan

The contributions to the IS 2000 multi-conference are collected in three proceedings:

- Cognitive Science
- Language Technologies
- IS Joint Proceedings:
  - Information society and governmental services
  - Media in information society
  - Education in information society
  - Data warehouses and data mining
  - Development and reengineering of information systems

# Kazalo / Table of Contents

<i>Tamás Váradi:</i> Corpus Linguistics — Linguistics or Language Engineering? .....	1
<i>Simon Dobrišek, Janez Žibert, France Mihelič, Nikola Pavešić:</i> Ocenjevanja govornih modelov: razdalja med zaporedji govornih odsekov .....	6
<i>Tomaž Erjavec, Sašo Džeroski:</i> Strojno učenje geslenja neznanih slovenskih besed .....	14
<i>Vojko Gorjanc, Špela Vintar:</i> Iskanja po korpusu slovenskega jezika FIDA .....	20
<i>Jerneja Gros, France Mihelič, Simon Dobrišek, Tomaž Erjavec, Mario Žganec:</i> A Phonetically and Prosodically Annotated Slovene Speech Corpus .....	27
<i>Vladimir Hozjan, Zdravko Kačič, Daniel Ambruš Čeh:</i> Analiza prozodičnih značilk emocionalnega govora .....	31
<i>Zdravko Kačič, Bogomir Horvat, Matej Rojc, Aleksandra Zögling Markuš:</i> K samodejnemu pridobivanju jezikovnih virov s pomočjo interneta .....	35
<i>Janez Kaiser, Mirjam Sepesy Maučec, Zdravko Kačič, Bogomir Horvat:</i> Razpoznavanje tekočega slovenskega govora z velikim slovarjem .....	39
<i>France Mihelič, Jerneja Gros, Elmar Noth, Simon Dobrišek, Janez Žibert:</i> Recognition of Selected Prosodic Events in Slovenian Speech .....	45
<i>Primož Peterlin, Agnes Pisanski, Tomaž Erjavec:</i> “Poišči in zamenjaj” v pregibnih jezikih .....	49
<i>Primož Peterlin, Andraž Tori, Marko Samastur, Roman Maurer, Aleš Košir:</i> Uporaba pomnilnika prevodov pri prevajanju uporabniških vmesnikov v prostih programih .....	57
<i>Agnes Pisanski, Nataša Hirci:</i> Vloga jezikovnih korpusov pri izboljševanju jezikovne kompetence .....	64
<i>Vesna Požgaj Hadži, Marko Tadić:</i> Slovensko-hrvatski paralelni korpus .....	70
<i>Robert Rozman, Dušan Kodek:</i> Govorna baza ”Števk”in raziskave robustnosti sistemov za razpoznavanje govora .....	75
<i>Marko Stabej, Primož Vitez:</i> KGB (korpus govornih besedil) v slovenščini .....	79
<i>Janez Stergar, Vladimir Hozjan:</i> Steps Towards Preparation of Text Corpora for Data Driven Symbolic Prosody Labelling .....	82

<i>Robert Veronik, Zdravko Kačič, Klara Vicsi, Ferenc Csatari, Peter Roach, Anne-Marie Ester, Peter Barczikay:</i> Razvoj multimedijskega večjezičnega sistema za učenje in vadbo pravilnega izgovora govorno in slušno prizadetih otrok .....	86
<i>Damjan Vlaj, Zdravko Kačič, Bogomir Horvat:</i> Detektiranje pogrešanih značilk v šumnem okolju .....	90
<i>Aleksandra Zögling Markuš, Zdravko Kačič, Bogomir Horvat:</i> Razvoj slovenske baze izgovorjav "POLIDAT" .....	95
<i>Andrej Žgank, Zdravko Kačič, Bogomir Horvat:</i> Razpoznavanje tekočega angleškega govora z bazo "Broadcast News" .....	99
<i>Andrej Žgank, Zdravko Kačič, Bogomir Horvat:</i> SEDLIS — sistem za avtomatsko identifikacijo jezika iz besedila .....	104
<i>Janez Žibert, France Mihelič:</i> Govorna zbirka vremenskih napovedi .....	108
<i>Damir Čavar, Alexander Geyken, Gerald Neumann:</i> Digital Dictionary of the 20th Century German Language .....	112
<i>Primož Jakopin:</i> Elektronski in spletni naslovi v časopisu DELO 1998-2000 .....	116
<i>Laurent Romary, Jean-Luc Husson, Marc Van Campenhoudt, Nadia Viscogliosi:</i> A Standard Based Environment for the Edition and Distribution of Multilingual Terminological Data: the DHYDRO Platform .....	120
Kazalo po avtorjih / Author Index .....	127