CURRICULUM VITAE
of

Dr Tomaž Erjavec

Ph.D. Thesis: Unification, Inheritance and Paradigms in the Morphology of Natural Languages **(**PhD in Computer Science, University of Ljubljana)

Employment:
Dept. of Knowledge Technologies at the Jožef Stefan Institute in Ljubljana, Slovenia.

Project Manager:
International
CLARIN
Research infrastructure for language resources
Bilateral Slovene-Serbian project (2004-2005)
Development of Slovene and Serbian Language Resources for Machine Translation
GENIA JSPS Research for the Future program (2002)
Automatic extraction of information from biomedical texts
(local page)
CONCEDE Copernicus Joint Project (1998-2000):
Consortium for Central European Dictionary Encoding
(local page)
TELRI Copernicus Concerted Action (1999-2001, 1995-1997):
Trans-European Language Resources Infrastructure II
(local page)
ELAN MLIS EU Project (1998-1999):
European Language Activity Network
MULTEXT-EAST Copernicus Joint Project COP 106 (1996-1997):
Multilingual Texts and Corpora for Eastern and Central European Languages
LLL Informal SIG
Learning Language in Logic
(local page)
ILD UK SALT project (1994-1996)
The Integrated Language Database
(RA, Centre for Cognitive Science, 1994)

National
Knowledge Technologies (2004-2009)
Ministry of Sports, Science and Education Research Project
Cover financing for IJS Department of Knowledge Technologies
Mini projects (work in progress):
Slovene Dependency Treebank
Japanese-Slovene Learner's Dictionary
Concordancing the Slovenian Informatics Conference Corpus
Ministry of Education, Science and Sport Applied Project (2004-2006)
Digital Critical Editions of Slovene Literature
Slovene Cross-Ministry Targeted Research Projects: (2006-2008)
VoiceTRAN - a speech-to-speech communicator
Slovene Cross-Ministry Targeted Research Projects: (2005-2006)

Oblikovanje slovenskega korpusnega omrežja (Compilation of the Slovene Corpus Network)
Izdelava virov in sistema za simultano prevajanje slovenscina-anglescina (Producing resources and system for simultaneous translation Slovene-English)
Ministry of Information Society Project (2001)
Localisation of Open Source Spell Checkers ispell and aspell
MZT L2-0461-0106 (1998-2001)
Development of Digital Publishing with Distance Learning Support
(project leader)
MZT T2-0409 (1998-2000)
Speech Copora and Tools for the Slovenian Language
FIDA (1996-1999)
Corpus of the Slovene Language
(TEI/SGML consulting)
(local access)
GNUsl (1995--)
A GNU effort for the Slovene Language
(server maintenance, resource contribution)
(local access)
RR(S)J Slovene Ministry of Science & Technology funded project (1993-1996)
Ra"cunalni"sko razumevanje (slovenskega) jezika
(Computational Understanding of (Slovene) Language)
(researcher)

Research Interests
Language technologies and computational linguistics for Slovene
Development of textual corpora and other linguistic resources
(XML & TEI structure, linguistic annotation)
Machine learning methods for natural language
Computational morphology
Typed feature-structure formalisms and implementations
I am currently serving as president of SDJT, the Slovenian Language Technologies Society. I am a member of EACL and TEI subscriber. I am the Web master of the IJS Natural Language Server.

List of Publications

Tomaž Erjavec: **TEI and Microsoft: a marriage made in...**. In *Digital Historical Corpora- Architecture, Annotation, and Retrieval*. Dagstuhl Seminar Proceedings 06491, 2007.

Tomaž Erjavec, Sarossy Bence: **Morphosyntactic Tagging of Slovene Legal Language**. Informatica, 30, pp. 483-488, 2006.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, Daniel Varga. **The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages**. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, 2006.

Tomaž Erjavec. **The English-Slovene ACQUIS corpus**. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, 2006.

Saso Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Źabokrtsky, Andreja Źele. **Towards a Slovene Dependency Treebank**. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, 2006.

Tomaž Erjavec, Darja Fišer. **Building Slovene Wordnet**. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06, ELRA, Paris, 2006.

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

Jerneja Žganec Gros, France Mihelič, Tomaž Erjavec, Špela Vintar. **The VoiceTRAN speech-to-speech communicator**. In 8th International Conference, TDS 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Text, speech and dialogue : proceedings, (Lecture notes in computer science, Lecture notes in artificial intelligence, 3658). Berlin: Springer, 2005, pp. 379-384.

Tomaž Erjavec, Matija Ogrin. **Digitalisation of literary heritage using open standards**. In Paul Cunningham, Miriam Cunningham (eds.). Innovation and knowledge economy: issues, applications, case studies, (Information and communication technologies and the knowledge economy). Amsterdam [etc.]: IOS Press, 2005, str. 999-1006.

Tomaž Erjavec. **MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora**. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, pp. 1535 - 1538, ELRA, Paris, 2004. [c.f. also http://nl.ijs.si/ME/V3/]

Syd Bauman, Alejandro Bia, Lou Burnard, Tomaž Erjavec, Christine Ruotolo and Susan Schreibman: **Migrating Language Resources from SGML to XML: the Text Encoding Initiative Recommendations**. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, p. 139 - 142 ELRA, Paris, 2004. [c.f. also http://www.tei-c.org.uk/Activities/MI/]

Tomaž Erjavec and Sašo Džeroski: **Machine Learning of Morphosyntactic Structure: Lemmatising Unknown Slovene Words**. *Applied Artificial Intelligence* 18(1), pp. 17-40, 2004.

Tomaž Erjavec, Roger Evans, Nancy Ide, Adam Kilgarriff: **From Machine Readable Dictionaries to Lexical Databases: the Concede Experience**. In the Proceedings of the 7th International Conference on Computational Lexicography, COMPLEX'03, Budapest, 2003.

Tomaž Erjavec: **Compiling and Using the IJS-ELAN Parallel Corpus**. Informatica, 26(3), pp. 299-307, 2002.

Sašo Džeroski, Tomaž Erjavec and Jakub Zavrel: **Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets**. Second International Conference on Language Resources and Evaluation, LREC'00, pp. 1099-1104, 2000. (available also in PDF)

James Cussens, Sašo Džeroski, Tomaž Erjavec: **Morphosyntactic Tagging of Slovene using Progol**. Proceedings of the Ninth International Workshop on Inductive Logic Programming, ILP-99; volume 1634 of Lecture Notes in Artificial Intelligence; Springer-Verlag, pp. 68--79.

Suresh Manandhar, Saso Džeroski, Tomaž Erjavec (1998): **Learning Multilingual Morphology with CLOG**. In David Page (ed): *Inductive Logic Programming, 8th International Conference, ILP'98, Proceedings*. Lecture Notes in Artificial Intelligence 1446, Springer, pp. 135-144.