

Tapio Salakoski
Filip Ginter
Sampo Pyysalo
Tapio Pahikkala (Eds.)

LNAI 4139

Advances in Natural Language Processing

5th International Conference on NLP, FinTAL 2006
Turku, Finland, August 2006
Proceedings



Springer

TP301.2-53

F516
2006

Tapio Salakoski Filip Ginter
Sampo Pyysalo Tapio Pahikkala (Eds.)

Advances in Natural Language Processing

5th International Conference on NLP, FinTAL 2006
Turku, Finland, August 23-25, 2006
Proceedings



Springer



E200604015

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Tapio Salakoski
Filip Ginter
Sampo Pyysalo
Tapio Pahikkala

University of Turku

Department of Information Technology

20014 Turku, Finland

E-mail: {Tapio.Salakoski, Filip.Ginter, Sampo.Pyysalo, Tapio.Pahikkala} @it.utu.fi

Library of Congress Control Number: 2006930503

CR Subject Classification (1998): I.2.7, F.4.2-3, I.2, H.3, I.7

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743

ISBN-10 3-540-37334-9 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-37334-6 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 11816508 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence

4139

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Lecture Notes in Artificial Intelligence (LNAI)

- Vol. 4139: T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), *Advances in Natural Language Processing*, XIV, 771 pages. 2006.
- Vol. 4114: D.-S. Huang, K. Li, G.W. Irwin (Eds.), *Computational Intelligence, Part II*, XXVII, 1337 pages. 2006.
- Vol. 4108: J.M. Borwein, W.M. Farmer (Eds.), *Mathematical Knowledge Management*, VIII, 295 pages. 2006.
- Vol. 4106: T.R. Roth-Berghofer, M.H. Göker, H. A. Güvenir (Eds.), *Advances in Case-Based Reasoning*, XIV, 566 pages. 2006.
- Vol. 4099: Q. Yang, G. Webb (Eds.), *PRICAI 2006: Trends in Artificial Intelligence*, XXVIII, 1263 pages. 2006.
- Vol. 4093: X. Li, O.R. Zaiane, Z. Li (Eds.), *Advanced Data Mining and Applications*, XXI, 1110 pages. 2006.
- Vol. 4092: J. Lang, F. Lin, J. Wang (Eds.), *Knowledge Science, Engineering and Management*, XV, 664 pages. 2006.
- Vol. 4088: Z.-Z. Shi, R. Sadananda (Eds.), *Agent Computing and Multi-Agent Systems*, XVII, 827 pages. 2006.
- Vol. 4068: H. Schärfe, P. Hitzler, P. Øhrstrøm (Eds.), *Conceptual Structures: Inspiration and Application*, XI, 455 pages. 2006.
- Vol. 4065: P. Perner (Ed.), *Advances in Data Mining*, XI, 592 pages. 2006.
- Vol. 4062: G. Wang, J.F. Peters, A. Skowron, Y. Yao (Eds.), *Rough Sets and Knowledge Technology*, XX, 810 pages. 2006.
- Vol. 4049: S. Parsons, N. Maudet, P. Moraitis, I. Rahwan (Eds.), *Argumentation in Multi-Agent Systems*, XIV, 313 pages. 2006.
- Vol. 4048: L. Goble, J.-J.C. Meyer (Eds.), *Deontic Logic and Artificial Normative Systems*, X, 273 pages. 2006.
- Vol. 4045: D. Barker-Plummer, R. Cox, N. Swoboda (Eds.), *Diagrammatic Representation and Inference*, XII, 301 pages. 2006.
- Vol. 4031: M. Ali, R. Dapoigny (Eds.), *Advances in Applied Artificial Intelligence*, XXIII, 1353 pages. 2006.
- Vol. 4029: L. Rutkowski, R. Tadeusiewicz, L.A. Zadeh, J.M. Zurada (Eds.), *Artificial Intelligence and Soft Computing – ICAISC 2006*, XXI, 1235 pages. 2006.
- Vol. 4027: H.L. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreassen, H. Christiansen (Eds.), *Flexible Query Answering Systems*, XVIII, 714 pages. 2006.
- Vol. 4021: E. André, L. Dybkjær, W. Minker, H. Neumann, M. Weber (Eds.), *Perception and Interactive Technologies*, XI, 217 pages. 2006.
- Vol. 4020: A. Bredenfeld, A. Jacoff, I. Noda, Y. Takahashi (Eds.), *RoboCup 2005: Robot Soccer World Cup IX*, XVII, 727 pages. 2006.
- Vol. 4013: L. Lamontagne, M. Marchand (Eds.), *Advances in Artificial Intelligence*, XIII, 564 pages. 2006.
- Vol. 4012: T. Washio, A. Sakurai, K. Nakajima, H. Takeda, S. Tojo, M. Yokoo (Eds.), *New Frontiers in Artificial Intelligence*, XIII, 484 pages. 2006.
- Vol. 4008: J.C. Augusto, C.D. Nugent (Eds.), *Designing Smart Homes*, XI, 183 pages. 2006.
- Vol. 4005: G. Lugosi, H.U. Simon (Eds.), *Learning Theory*, XI, 656 pages. 2006.
- Vol. 3978: B. Hnich, M. Carlsson, F. Fages, F. Rossi (Eds.), *Recent Advances in Constraints*, VIII, 179 pages. 2006.
- Vol. 3963: O. Dikenelli, M.-P. Gleizes, A. Ricci (Eds.), *Engineering Societies in the Agents World VI*, XII, 303 pages. 2006.
- Vol. 3960: R. Vieira, P. Quaresma, M.d.G.V. Nunes, N.J. Mamede, C. Oliveira, M.C. Dias (Eds.), *Computational Processing of the Portuguese Language*, XII, 274 pages. 2006.
- Vol. 3955: G. Antoniou, G. Potamias, C. Spyropoulos, D. Plexousakis (Eds.), *Advances in Artificial Intelligence*, XVII, 611 pages. 2006.
- Vol. 3949: F. A. Savacı (Ed.), *Artificial Intelligence and Neural Networks*, IX, 227 pages. 2006.
- Vol. 3946: T.R. Roth-Berghofer, S. Schulz, D.B. Leake (Eds.), *Modeling and Retrieval of Context*, XI, 149 pages. 2006.
- Vol. 3944: J. Quiñonero-Candela, I. Dagan, B. Magnini, F. d'Alché-Buc (Eds.), *Machine Learning Challenges*, XIII, 462 pages. 2006.
- Vol. 3930: D.S. Yeung, Z.-Q. Liu, X.-Z. Wang, H. Yan (Eds.), *Advances in Machine Learning and Cybernetics*, XXI, 1110 pages. 2006.
- Vol. 3918: W.K. Ng, M. Kitsuregawa, J. Li, K. Chang (Eds.), *Advances in Knowledge Discovery and Data Mining*, XXIV, 879 pages. 2006.
- Vol. 3913: O. Boissier, J. Padget, V. Dignum, G. Lindemann, E. Matsun, S. Ossowski, J.S. Sichman, J. Vázquez-Salceda (Eds.), *Coordination, Organizations, Institutions, and Norms in Multi-Agent Systems*, XII, 259 pages. 2006.
- Vol. 3910: S.A. Brueckner, G.D.M. Serugendo, D. Hales, F. Zambonelli (Eds.), *Engineering Self-Organising Systems*, XII, 245 pages. 2006.
- Vol. 3904: M. Baldoni, U. Endriss, A. Omicini, P. Torrioni (Eds.), *Declarative Agent Languages and Technologies III*, XII, 245 pages. 2006.

- Vol. 3900: F. Toni, P. Torroni (Eds.), *Computational Logic in Multi-Agent Systems*. XVII, 427 pages. 2006.
- Vol. 3899: S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*. XIV, 216 pages. 2006.
- Vol. 3898: K. Tuyls, P.J. 't Hoen, K. Verbeeck, S. Sen (Eds.), *Learning and Adaption in Multi-Agent Systems*. X, 217 pages. 2006.
- Vol. 3891: J.S. Sichman, L. Antunes (Eds.), *Multi-Agent-Based Simulation VI*. X, 191 pages. 2006.
- Vol. 3890: S.G. Thompson, R. Ghanea-Hercock (Eds.), *Defence Applications of Multi-Agent Systems*. XII, 141 pages. 2006.
- Vol. 3885: V. Torra, Y. Narukawa, A. Valls, J. Domingo-Ferrer (Eds.), *Modeling Decisions for Artificial Intelligence*. XII, 374 pages. 2006.
- Vol. 3881: S. Gibet, N. Courty, J.-F. Kamp (Eds.), *Gesture in Human-Computer Interaction and Simulation*. XIII, 344 pages. 2006.
- Vol. 3874: R. Missaoui, J. Schmidt (Eds.), *Formal Concept Analysis*. X, 309 pages. 2006.
- Vol. 3873: L. Maicher, J. Park (Eds.), *Charting the Topic Maps Research and Applications Landscape*. VIII, 281 pages. 2006.
- Vol. 3863: M. Kohlhase (Ed.), *Mathematical Knowledge Management*. XI, 405 pages. 2006.
- Vol. 3862: R.H. Bordini, M. Dastani, J. Dix, A.E.F. Seghrouchni (Eds.), *Programming Multi-Agent Systems*. XIV, 267 pages. 2006.
- Vol. 3849: I. Bloch, A. Petrosino, A.G.B. Tettamanzi (Eds.), *Fuzzy Logic and Applications*. XIV, 438 pages. 2006.
- Vol. 3848: J.-F. Boulicaut, L. De Raedt, H. Mannila (Eds.), *Constraint-Based Mining and Inductive Databases*. X, 401 pages. 2006.
- Vol. 3847: K.P. Jantke, A. Lunzer, N. Spyrtatos, Y. Tanaka (Eds.), *Federation over the Web*. X, 215 pages. 2006.
- Vol. 3835: G. Sutcliffe, A. Voronkov (Eds.), *Logic for Programming, Artificial Intelligence, and Reasoning*. XIV, 744 pages. 2005.
- Vol. 3830: D. Weyns, H. V.D. Parunak, F. Michel (Eds.), *Environments for Multi-Agent Systems II*. VIII, 291 pages. 2006.
- Vol. 3817: M. Faundez-Zanuy, L. Janer, A. Esposito, A. Satue-Villar, J. Roure, V. Espinosa-Duro (Eds.), *Nonlinear Analyses and Algorithms for Speech Processing*. XII, 380 pages. 2006.
- Vol. 3814: M. Maybury, O. Stock, W. Wahlster (Eds.), *Intelligent Technologies for Interactive Entertainment*. XV, 342 pages. 2005.
- Vol. 3809: S. Zhang, R. Jarvis (Eds.), *AI 2005: Advances in Artificial Intelligence*. XXVII, 1344 pages. 2005.
- Vol. 3808: C. Bento, A. Cardoso, G. Dias (Eds.), *Progress in Artificial Intelligence*. XVIII, 704 pages. 2005.
- Vol. 3802: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), *Computational Intelligence and Security, Part II*. XLII, 1166 pages. 2005.
- Vol. 3801: Y. Hao, J. Liu, Y.-P. Wang, Y.-m. Cheung, H. Yin, L. Jiao, J. Ma, Y.-C. Jiao (Eds.), *Computational Intelligence and Security, Part I*. XLI, 1122 pages. 2005.
- Vol. 3789: A. Gelbukh, Á. de Albornoz, H. Terashima-Marín (Eds.), *MICAI 2005: Advances in Artificial Intelligence*. XXVI, 1198 pages. 2005.
- Vol. 3782: K.-D. Althoff, A. Dengel, R. Bergmann, M. Nick, T.R. Roth-Berghofer (Eds.), *Professional Knowledge Management*. XXIII, 739 pages. 2005.
- Vol. 3763: H. Hong, D. Wang (Eds.), *Automated Deduction in Geometry*. X, 213 pages. 2006.
- Vol. 3755: G.J. Williams, S.J. Simoff (Eds.), *Data Mining*. XI, 331 pages. 2006.
- Vol. 3735: A. Hoffmann, H. Motoda, T. Scheffer (Eds.), *Discovery Science*. XVI, 400 pages. 2005.
- Vol. 3734: S. Jain, H.U. Simon, E. Tomita (Eds.), *Algorithmic Learning Theory*. XII, 490 pages. 2005.
- Vol. 3721: A.M. Jorge, L. Torgo, P.B. Brazdil, R. Camacho, J. Gama (Eds.), *Knowledge Discovery in Databases: PKDD 2005*. XXIII, 719 pages. 2005.
- Vol. 3720: J. Gama, R. Camacho, P.B. Brazdil, A.M. Jorge, L. Torgo (Eds.), *Machine Learning: ECML 2005*. XXIII, 769 pages. 2005.
- Vol. 3717: B. Gramlich (Ed.), *Frontiers of Combining Systems*. X, 321 pages. 2005.
- Vol. 3702: B. Beckert (Ed.), *Automated Reasoning with Analytic Tableaux and Related Methods*. XIII, 343 pages. 2005.
- Vol. 3698: U. Furbach (Ed.), *KI 2005: Advances in Artificial Intelligence*. XIII, 409 pages. 2005.
- Vol. 3690: M. Pěchouček, P. Petta, L.Z. Varga (Eds.), *Multi-Agent Systems and Applications IV*. XVII, 667 pages. 2005.
- Vol. 3684: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part IV*. LXXIX, 933 pages. 2005.
- Vol. 3683: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part III*. LXXX, 1397 pages. 2005.
- Vol. 3682: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part II*. LXXIX, 1371 pages. 2005.
- Vol. 3681: R. Khosla, R.J. Howlett, L.C. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems, Part I*. LXXX, 1319 pages. 2005.
- Vol. 3673: S. Bandini, S. Manzoni (Eds.), *AI*IA 2005: Advances in Artificial Intelligence*. XIV, 614 pages. 2005.
- Vol. 3662: C. Baral, G. Greco, N. Leone, G. Terracina (Eds.), *Logic Programming and Nonmonotonic Reasoning*. XIII, 454 pages. 2005.
- Vol. 3661: T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, T. Rist (Eds.), *Intelligent Virtual Agents*. XIII, 506 pages. 2005.
- Vol. 3658: V. Matoušek, P. Mautner, T. Pavelka (Eds.), *Text, Speech and Dialogue*. XV, 460 pages. 2005.

Preface

The research papers in this volume comprise the proceedings of FinTAL 2006, a Natural Language Processing conference continuing the TAL series of events: FracTAL 1997 at Université de Franche-Comté in Besançon, France; VexTAL 1999 at Università Ca' Foscari di Venezia in Venice, Italy; PorTAL 2002 at Universidade do Algarve in Faro, Portugal; and EsTAL 2004 at Universitat d'Alacant in Alicante, Spain. The main goals of the TAL conferences have been to bring together the international NLP community, to strengthen local NLP research, and to provide a forum for discussion of new NLP research and applications.

FinTAL 2006, organized by Turku Centre for Computer Science (TUCS) in Turku, Finland, also contributed to the goals mentioned above, further increasing the high international standing of the TAL conference series. We called for submissions both from academia and industry on any topic that is of interest to the NLP community, particularly encouraging research emphasizing multidisciplinary aspects of NLP and the interplay between linguistics, computer science and application domains such as biomedicine, communication systems, public services, and educational technology.

As a response, we received as many as 150 submissions from 38 countries in Europe, Asia, Africa, and the Americas. The manuscripts were reviewed by three members of FinTAL Program Committee, composed of researchers in the field, or external reviewers designated by the PC members. The PC members as well as the external reviewers are gratefully acknowledged in the following pages for their valuable contribution.

We would like to express here our gratitude to Turku Centre for Computer Science (TUCS), University of Turku, and Åbo Akademi University, as well as to our sponsors the city of Turku, Nokia, Lingsoft, PARC, and Sanako. We also thank our keynote speakers, the highly esteemed scholars Fred Karlsson, Lauri Karttunen, and Igor Mel'čuk. Last but obviously not least, we would like to thank all the individuals involved in organizing the event; without you it would not have been possible at all.

The careful evaluation of the submissions finally led to the selection of 72 papers to be presented at the conference and published in this volume. It is our firm belief that the accepted papers provide a significant contribution to the advance of science and technology. We hope that you will enjoy reading the articles and find them inspiring for your work, whether in basic NLP research or in the development of human language technology applications.

Organization

FinTAL 2006 was organized by Turku Centre for Computer Science (TUCS) in conjunction with the University of Turku and Åbo Akademi University.



Program Committee

Tapio Salakoski	University of Turku, Finland (Chair)
Olli Aaltonen	University of Turku, Finland
Walid El Abed	Nestlé Corp., Switzerland
Jorge Baptista	University of Faro, Portugal
Florence Beaujard	Airbus Corp., France
Krzysztof Bogacki	University of Warsaw, Poland
Caroline Brun	Xerox Corp., France
Sylviane Cardey	University of Franche-Comte, France
Nigel Collier	National Institute of Informatics, Japan
Walter Daelemans	University of Antwerp, Belgium
Rodolfo Delmonte	University of Venice, Italy
Pasi Fränti	University of Joensuu, Finland
Peter Greenfield	University of Franche-Comté, France
Jari Kangas	Nokia Research Center, Finland
Kimmo Koskenniemi	University of Helsinki, Finland
Kyoko Kuroda	Shimane College, Japan
Hsiang-I Lin	National Taiwan University (NTU), Taiwan
Nuno Mamede	University of Lisbon, Portugal
Patricio Martínez-Barco	University of Alicante, Spain
Einar Meister	Tallinn University of Technology, Estonia
Rada Mihalcea	University of North Texas, USA
Leonel Ruiz Miyares	University of Santiago de Cuba, Cuba
Adeline Nazarenko	University Paris-Nord, France
Elisabete Ranchhod	University of Lisbon, Portugal
Karl-Michael Schneider	Textkernel BV, Amsterdam, The Netherlands
Rolf Schwitter	Macquarie University, Australia
John Tait	University of Sunderland, UK
José Luis Vicedo	University of Alicante, Spain
Simo Vihjanen	Lingsoft Ltd., Finland
Roman Yangarber	University of Helsinki, Finland

Local Organizers

Filip Ginter
Sampo Pyysalo
Hanna Suominen
Tapio Pahikkala
Tomi 'bgt' Mäntylä
Irmeli Laine
Christel Donner

Reviewers

Adeline Nazarenko	Hugo Meinedo	Paloma Moreda
Agnes Sandor	Ismo Kärkkäinen	Pasi Fränti
Amanda Bouffier	Jari Kangas	Patricio Martínez-Barco
Annu Paganus	Jean-Sébastien Tisserand	Paula Carvalho
Anssi Yli-Jyrä	Joana L. Paulo	Peter Greenfield
Antoine Doucet	Joao Cabral	Rada Mihalcea
Antoine Rozenknop	John Tait	Rafael Carrasco
Armando Suárez-Cueto	Jorge Baptista	Rafael Muñoz
Borja Navarro	Jorma Boberg	Ricardo Ribeiro
Caroline Brun	José Luis Vicedo	Rodolfo Delmonte
Caroline Hagege	Jouni Järvinen	Rolf Schwitter
Christopher Stokoe	Juhani Saastamoinen	Roman Yangarber
Cristina Mota	Jussi Hakokari	Sampo Pyysalo
David Martins de Matos	Jussi Piitulainen	Shao Fen Liang
David Tomás	Jyri Paakkulainen	Simo Vihjanen
Davy Weissenbacher	Karl-Michael Schneider	Siobhan Devlin
Diamantino Caseiro	Kimmo Koskenniemi	Sylviane Cardey
Duygu Can	Krister Lindén	Séverine Vienney
Einar Meister	Krzysztof Bogacki	Tanja Kavander
Elisabete Ranchhod	Kyoko Kuroda	Tapio Pahikkala
Evgeni Tsvitshivadze	Leonel Ruiz Miyares	Tapio Salakoski
Evgenia Chernenko	Luisa Coheur	Thierry Poibeau
Felipe Sánchez-Martínez	Manuel Célio Conceição	Timo Honkela
Filip Ginter	Marc Dymetman	Timo Knuutila
Florence Beaujard	Marketta Hiissa	Tomi 'bgt' Mäntylä
Graham Wilcock	Maud Ehrmann	Tony Mullen
Guillaume Bouchard	Michael Oakes	Tuomo Saarni
Guillaume Jacquet	Navid Atar Sharghi	Ville Hautamäki
Hanna Suominen	Nigel Collier	Walid El Abed
Hervé Déjean	Nuno Mamede	Walter Daelemans
Hsiang-I Lin	Olli Aaltonen	Xiaohong Wu

Conference Sponsors



NOKIA
Connecting People



parc[®]
Palo Alto Research Center



Table of Contents

Keynote Addresses

Recursion in Natural Languages	1
<i>Fred Karlsson</i>	
The Explanatory Combinatorial Dictionary as the Key Tool in Machine Translation	2
<i>Igor Mel'čuk</i>	
A Finite-State Approximation of Optimality Theory: The Case of Finnish Prosody	4
<i>Lauri Karttunen</i>	

Research Papers

A Bilingual Corpus of Novels Aligned at Paragraph Level	16
<i>Alexander Gelbukh, Grigori Sidorov, José Ángel Vera-Félix</i>	
A Computational Implementation of Internally Headed Relative Clause Constructions	24
<i>Jong-Bok Kim, Peter Sells, Jaehyung Yang</i>	
A Corpus-Based Empirical Account of Adverbial Clauses Across Speech and Writing in Contemporary British English	32
<i>Alex Chengyu Fang</i>	
A Korean Syntactic Parser Customized for Korean-English Patent MT System	44
<i>Chang-Hyun Kim, Munpyo Hong</i>	
A Scalable and Distributed NLP Architecture for Web Document Annotation	56
<i>Julien Deriviere, Thierry Hamon, Adeline Nazarenko</i>	
A Straightforward Method for Automatic Identification of Marginalized Languages	68
<i>Ana Lilia Reyes-Herrera, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez</i>	
A Text Mining Approach for Definition Question Answering	76
<i>Claudia Denicia-Carral, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, René García-Hernández</i>	

Accommodating Multiword Expressions in an Arabic LFG Grammar	87
<i>Mohammed A. Attia</i>	
Analysis of EU Languages Through Text Compression	99
<i>Kimmo Kettunen, Markus Sadeniemi, Tiina Lindh-Knuutila, Timo Honkela</i>	
Applying Latent Dirichlet Allocation to Automatic Essay Grading	110
<i>Tuomo Kakkonen, Niko Myller, Erkki Sutinen</i>	
Automatic Acquisition of Semantic Relationships from Morphological Relatedness	121
<i>Delphine Bernhard</i>	
Automatic Feature Extraction for Question Classification Based on Dissimilarity of Probability Distributions	133
<i>David Tomás, José L. Vicedo, Empar Bisbal, Lidia Moreno</i>	
Cat3LB and Cast3LB: From Constituents to Dependencies	141
<i>Montserrat Civit, Ma. Antònia Martí, Núria Bufí</i>	
Classification of News Web Documents Based on Structural Features	153
<i>Shisanu Tongchim, Virach Sornlertlamvanich, Hitoshi Isahara</i>	
Cognition and Physio-acoustic Correlates—Audio and Audio-Visual Effects of a Short English Emotional Statement: On JL2, FL2 and EL1	161
<i>Toshiko Isei-Jaakkola</i>	
Compiling Generalized Two-Level Rules and Grammars	174
<i>Anssi Yli-Jyrä, Kimmo Koskenniemi</i>	
Computer Analysis of the Turkmen Language Morphology	186
<i>A. Cüneyd Tantuğ, Eşref Adalı, Kemal Oflazer</i>	
Coordination Structures in a Typed Feature Structure Grammar: Formalization and Implementation	194
<i>Jong-Bok Kim, Jaehyung Yang</i>	
Cue-Based Interpretation of Customer's Requests: Analysis of Estonian Dialogue Corpus	206
<i>Tiit Hennoste, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo</i>	
Czech-English Phrase-Based Machine Translation	214
<i>Ondřej Bojar, Evgeny Matusov, Hermann Ney</i>	
Deep vs. Shallow Semantic Analysis Applied to Textual Entailment Recognition	225
<i>Óscar Ferrández, Rafael Muñoz Terol, Rafael Muñoz, Patricio Martínez-Barco, Manuel Palomar</i>	

Dictionary-Free Morphological Classifier of Russian Nouns	237
<i>Igor A. Bolshakov, Elena I. Bolshakova</i>	
Discourse Segmentation of German Written Texts	245
<i>Harald Lüngen, Csilla Puskás, Maja Bärenfänger, Mirco Hilbert, Henning Lobin</i>	
Document Clustering Based on Maximal Frequent Sequences	257
<i>Edith Hernández-Reyes, René García-Hernández, Jesus A. Carrasco-Ochoa, Jose Fco. Martínez-Trinidad</i>	
Enriching Thesauri with Hierarchical Relationships by Pattern Matching in Dictionaries	268
<i>Lourdes Araujo, José R. Pérez-Agüera</i>	
Evaluation of Alignment Methods for HTML Parallel Text	280
<i>Enrique Sánchez-Villamil, Susana Santos-Antón, Sergio Ortiz-Rojas, Mikel L. Forcada</i>	
Experiments in Passage Selection and Answer Identification for Question Answering	291
<i>Horacio Saggion, Robert Gaizauskas</i>	
Extracting Idiomatic Hungarian Verb Frames	303
<i>Bálint Sass</i>	
Extracting Term Collocations for Directing Users to Informative Web Pages	310
<i>Eiko Yamamoto, Hitoshi Isahara</i>	
Feasibility of Enriching a Chinese Synonym Dictionary with a Synchronous Chinese Corpus	322
<i>Oi Yee Kwong, Benjamin K. Tsou</i>	
Finding Spanish Syllabification Rules with Decision Trees	333
<i>John Goddard, René MacKinney-Romero</i>	
Identifying Text Discourse Structure of the Narratives Describing Psychiatric Patients' Defense Mechanisms	341
<i>Eunmi Ham, Woojin Paik</i>	
Implementing a Rule-Based Speech Synthesizer on a Mobile Platform....	349
<i>Tuomo Saarni, Jyri Paakkulainen, Tuomas Mäkilä, Jussi Hakokari, Olli Aaltonen, Jouni Isoaho, Tapio Salakoski</i>	
Improving Phrase-Based Statistical Translation Through Combination of Word Alignments	356
<i>Boxing Chen, Marcello Federico</i>	

Improving Statistical Word Alignments with Morpho-syntactic Transformations	368
<i>Adrià de Gispert, Deepa Gupta, Maja Popović, Patrik Lambert, Jose B. Mariño, Marcello Federico, Hermann Ney, Rafael Banchs</i>	
Improving Term Extraction with Terminological Resources	380
<i>Sophie Aubin, Thierry Hamon</i>	
Improving Thai Spelling Recognition with Tone Features	388
<i>Chutima Pisarn, Thanaruk Theeramunkong</i>	
Incorporating External Information in Bayesian Classifiers Via Linear Feature Transformations	399
<i>Tapio Pahikkala, Jorma Boberg, Aleksandr Mylläri, Tapio Salakoski</i>	
Is a Morphologically Complex Language Really That Complex in Full-Text Retrieval?	411
<i>Kimmo Kettunen, Eija Airio</i>	
Language Independent Answer Prediction from the Web	423
<i>Alejandro Figueroa, Günter Neumann</i>	
Language Model Mixtures for Contextual Ad Placement in Personal Blogs	435
<i>Gilad Mishne, Maarten de Rijke</i>	
Local Constraints on Arabic Word Order	447
<i>Allan Ramsay, Hanady Mansour</i>	
MEDITE: A Unilingual Textual Aligner	458
<i>Julien Bourdaillet, Jean-Gabriel Ganascia</i>	
Maximum Likelihood Alignment of Translation Equivalents	470
<i>Saba Amsalu</i>	
Measuring Intelligibility of Japanese Learner English	476
<i>Emi Izumi, Kiyotaka Uchimoto, Hitoshi Isahara</i>	
Morphological Lexicon Extraction from Raw Text Data	488
<i>Markus Forsberg, Harald Hammarström, Aarne Ranta</i>	
On the Use of Topic Models for Word Completion	500
<i>Elisabeth Wolf, Shankar Vembu, Tristan Miller</i>	
Ord i Dag: Mining Norwegian Daily Newswire	512
<i>Unni Cathrine Eiken, Anja Therese Liseth, Hans Friedrich Witschel, Matthias Richter, Chris Biemann</i>	

Paraphrase Identification on the Basis of Supervised Machine Learning Techniques.....	524
<i>Zornitsa Kozareva, Andrés Montoyo</i>	
Passage Filtering for Open-Domain Question Answering	534
<i>Elisa Noguera, Fernando Llopis, Antonio Ferrández</i>	
Persian in MULTTEXT-East Framework	541
<i>Behrang QasemiZadeh, Saeed Rahimi</i>	
Prerequisites for a Comprehensive Dictionary of Serbian Compounds	552
<i>Cvetana Krstev, Duško Vitas, Agata Savary</i>	
Regular Approximation of Link Grammar	564
<i>Filip Ginter, Sampo Pyysalo, Jorma Boberg, Tapio Salakoski</i>	
Segmental Duration in Utterance-Initial Environment: Evidence from Finnish Speech Corpora	576
<i>Tuomo Saarni, Jussi Hakokari, Jouni Isoaho, Olli Aaltonen, Tapio Salakoski</i>	
Selection Strategies for Multi-label Text Categorization	585
<i>Arturo Montejo-Ráez, Luis Alfonso Ureña-López</i>	
Some Problems of Prepositional Phrases in Machine Translation	593
<i>Xiaohong Wu, Sylviane Cardey, Peter Greenfield</i>	
Speech Confusion Index ($\emptyset$): A Recognition Rate Indicator for Dysarthric Speakers	604
<i>Prakasith Kayasith, Thanaruk Theeramunkong, Nuttakorn Thubthong</i>	
Statistical Machine Translation of German Compound Words.....	616
<i>Maja Popović, Daniel Stein, Hermann Ney</i>	
Summarizing Documents in Context: Modeling the User's Information Need	625
<i>Yllias Chali</i>	
Supervised TextRank	632
<i>Fermín Cruz, José A. Troyano, Fernando Enríquez</i>	
Tagging a Morphologically Complex Language Using Heuristics	640
<i>Hrafn Loftsson</i>	
Terminology Structuring Through the Derivational Morphology	652
<i>Natalia Grabar, Thierry Hamon</i>	

Text Segmentation Criteria for Statistical Machine Translation 664
Mauro Cettolo, Marcello Federico

The Classification Sense-Mining System 674
*Sylviane Cardey, Peter Greenfield, Mounira Bioud,
Aleksandra Dziadkiewicz, Kyoko Kuroda, Izabel Marcelino,
Ciprian Melian, Helena Morgadinho, Guillaume Robardet,
S  verine Vienney*

The Role of Verb Sense Disambiguation in Semantic Role Labeling 684
Paloma Moreda, Manuel Palomar

The Vowel Game: Continuous Real-Time Visualization
for Pronunciation Learning with Vowel Charts 696
*Annu Paganus, Vesa-Petteri Mikkonen, Tomi M  ntyl  ,
Sami Nuuttila, Jouni Isoaho, Olli Aaltonen, Tapio Salakoski*

Towards a Framework for Evaluating Syntactic Parsers 704
Tuomo Kakkonen, Erkki Sutinen

Towards the Improvement of Statistical Translation Models Using
Linguistic Features 716
Alicia P  rez, In  s Torres, Francisco Casacuberta

Treating Unknown Light Verb Construction in Korean-to-English
Patent MT 726
Munpyo Hong, Chang-Hyun Kim, Sang-Kyu Park

Trees as Contexts in Formal Language Generation 738
Adrian-Horia Dediu, Gabriela Mart  n

Two String-Based Finite-State Models of the Semantics of Calendar
Expressions 748
Jyrki Niemi, Lauri Carlson, Kimmo Koskenniemi

Using Alignment Templates to Infer Shallow-Transfer Machine
Translation Rules 756
Felipe S  nchez-Mart  nez, Hermann Ney

Author Index 769

Recursion in Natural Languages

Fred Karlsson

Department of General Linguistics
P.O. Box 9, FI-00014 University of Helsinki, Finland
fgk@ling.helsinki.fi

Abstract. The received view is that there are no grammatical constraints on clausal embedding complexity in sentences in languages of the ‘Standard Average European’ (SAE) type like English, Finnish, and Russian. The foremost proponent of this thesis is Noam Chomsky. This hypothesis of unbounded clausal embedding complexity is closely related to the hypothesis of unbounded syntactic recursion.

Psycholinguistic experimentation in the 1960’s established that there are clear performance-related preferences especially regarding center-embedding. The acceptability of repeated center-embeddings (nesting) below depth 1 steeply decreases with each successive level of embedding.

Not much corpus-based work has been done to find out what the empirical ‘facts’ of clausal embedding complexity are. I have conducted extensive corpus studies of English, Finnish, German, Latin, and Swedish, with the aim of determining the most complex clausal embedding patterns actually used. The basic constraint on nested center-embedding in written language turns out to be two (with a marginal cline to three), in spoken language one. There are further specific restrictions on which types of clauses may be nested. The practical limit of final embedding (right-branching) is five. Repeated initial embedding (left-branching) of clauses below depth two is not possible.

These written language constraints were reached already in Sumerian, Akkadian, and Latin along with the advent of written language and have remained the same ever since.

The constraints on center-embedding imply that SAE syntax is finite-state, type 3 in the Chomsky hierarchy. Clause-level recursion is thus not unbounded. The special case of right-branching relative clauses is rather an instance of depth-preserving iteration.