

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Germany*

Madhu Sudan

*Microsoft Research, Cambridge, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbruecken, Germany*

Alexander Gelbukh (Ed.)

# Computational Linguistics and Intelligent Text Processing

14th International Conference, CICLing 2013  
Samos, Greece, March 24-30, 2013  
Proceedings, Part I



Springer

Volume Editor

Alexander Gelbukh  
National Polytechnic Institute  
Center for Computing Research  
Av. Juan Dios Bátiz, Col. Nueva Industrial Vallejo  
07738 Mexico D.F., Mexico

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-37246-9

e-ISBN 978-3-642-37247-6

DOI 10.1007/978-3-642-37247-6

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2013933372

CR Subject Classification (1998): H.3, H.4, F.1, I.2, H.5, H.2.8, I.5

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

© Springer-Verlag Berlin Heidelberg 2013

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

# Preface

CICLing 2013 was the 14<sup>th</sup> Annual Conference on Intelligent Text Processing and Computational Linguistics. The CICLing conferences provide a wide-scope forum for discussion of the art and craft of natural language processing research as well as the best practices in its applications.

This set of two books contains four invited papers and a selection of regular papers accepted for presentation at the conference. Since 2001, the proceedings of the CICLing conferences have been published in Springer's *Lecture Notes in Computer Science* series as volume numbers 2004, 2276, 2588, 2945, 3406, 3878, 4394, 4919, 5449, 6008, 6608, 6609, 7181, and 7182.

The set has been structured into 12 sections:

- General Techniques
- Lexical Resources
- Morphology and Tokenization
- Syntax and Named Entity Recognition
- Word Sense Disambiguation and Coreference Resolution
- Semantics and Discourse
- Sentiment, Polarity, Emotion, Subjectivity, and Opinion
- Machine Translation and Multilingualism
- Text Mining, Information Extraction, and Information Retrieval
- Text Summarization
- Stylometry and Text Simplification
- Applications

The 2013 event received a record high number of submissions in the 14-year history of the CICLing series. A total of 354 papers by 788 authors from 55 countries were submitted for evaluation by the International Program Committee; see Figure 1 and Tables 1 and 2. This two-volume set contains revised versions of 87 regular papers selected for presentation; thus the acceptance rate for this set was 24.6%.

The book features invited papers by

- Sophia Ananiadou, University of Manchester, UK
- Walter Daelemans, University of Antwerp, Belgium
- Roberto Navigli, Sapienza University of Rome, Italy
- Michael Thelwall, University of Wolverhampton, UK

who presented excellent keynote lectures at the conference. Publication of full-text invited papers in the proceedings is a distinctive feature of the CICLing conferences. Furthermore, in addition to presentation of their invited papers, the keynote speakers organized separate vivid informal events; this is also a distinctive feature of this conference series.

**Table 1.** Number of submissions and accepted papers by topic<sup>1</sup>

Accepted	Submitted	% accepted	Topic
18	75	24	Text mining
18	64	28	Semantics, pragmatics, discourse
17	80	21	Information extraction
17	67	25	Lexical resources
14	44	32	Other
14	35	40	Emotions, sentiment analysis, opinion mining
13	40	33	Practical applications
11	52	21	Information retrieval
11	51	22	Machine translation and multilingualism
8	30	27	Syntax and chunking
7	40	17	Underresourced languages
7	39	18	Clustering and categorization
6	23	26	Summarization
5	32	16	Morphology
5	24	21	Word sense disambiguation
5	19	26	Named entity recognition
4	20	20	Noisy text processing and cleaning
4	17	24	Social networks and microblogging
4	13	31	Natural language generation
3	11	27	Coreference resolution
3	9	33	Natural language interfaces
3	8	38	Question answering
2	23	9	Formalisms and knowledge representation
2	18	11	POS tagging
2	2	100	Computational humor
1	11	9	Speech processing
1	11	9	Computational terminology
1	8	12	Spelling and grammar checking
1	3	33	Textual entailment

<sup>1</sup> As indicated by the authors. A paper may belong to several topics.

With this event we continued with our policy of giving preference to papers with verifiable and reproducible results: in addition to the verbal description of their findings given in the paper, we encouraged the authors to provide a proof of their claims in electronic form. If the paper claimed experimental results, we asked the authors to make available to the community all the input data necessary to verify and reproduce these results; if it claimed to introduce an algorithm, we encouraged the authors to make the algorithm itself, in a programming language, available to the public. This additional electronic material will be permanently stored on the CICLing’s server, [www.CICLing.org](http://www.CICLing.org), and will be available to the readers of the corresponding paper for download under a license that permits its free use for research purposes.

In the long run we expect that computational linguistics will have verifiability and clarity standards similar to those of mathematics: in mathematics, each

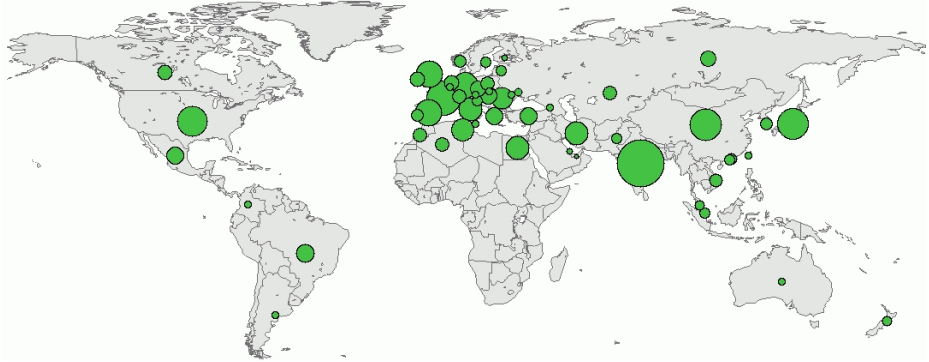
**Table 2.** Number of submitted and accepted papers by country or region

Country or region	Authors		Papers <sup>2</sup>	Country or region	Authors		Papers <sup>2</sup>
	Subm.	Subm.	Accp.		Subm.	Subm.	Accp.
Algeria	4	4	–	Malaysia	7	1.67	1
Argentina	3	1	–	Malta	1	1	–
Australia	3	1	–	Mexico	14	6.25	3.25
Austria	1	1	–	Moldova	3	1	–
Belgium	3	1	1	Morocco	7	4	1
Brazil	13	6.83	2	Netherlands	8	4.50	1
Canada	11	4.53	1.2	New Zealand	5	1.67	–
China	57	21.72	3.55	Norway	6	2.92	0.92
Colombia	2	1	1	Pakistan	5	2	–
Croatia	5	2	2	Poland	8	3.75	0.75
Czech Rep.	10	5	2	Portugal	9	3	–
Egypt	22	11.67	1	Qatar	2	0.67	–
Finland	2	0.67	–	Romania	14	9.67	2
France	64	25.9	5.65	Russia	15	4.75	1
Georgia	1	1	0.5	Singapore	5	2.25	0.25
Germany	32	13.92	6.08	Slovakia	2	1	–
Greece	21	6.12	2.12	Spain	39	15.50	8.75
Hong Kong	9	2.53	0.2	Sweden	2	2	–
Hungary	12	6	–	Switzerland	8	3.83	1.33
India	98	49.2	5.6	Taiwan	1	1	–
Iran	14	11.33	–	Tunisia	24	11	2
Ireland	6	4.5	1.5	Turkey	11	6.25	3.25
Italy	22	11.37	4.5	Ukraine	2	1.25	0.50
Japan	48	20.5	5	UAE	1	0.33	–
Kazakhstan	10	3.75	–	UK	35	15.73	5.20
Korea, South	7	3	–	USA	54	18.98	8.90
Latvia	6	2	1	Viet Nam	8	3.50	–
Macao	6	2	–	<i>Total:</i>	788	354	87

<sup>2</sup> By the number of authors: e.g., a paper by two authors from the USA and one from UK is counted as 0.67 for the USA and 0.33 for UK.

claim is accompanied by a complete and verifiable proof (usually much longer than the claim itself); each theorem’s complete and precise proof—and not just a vague description of its general idea—is made available to the reader. Electronic media allow computational linguists to provide material analogous to the proofs and formulas in mathematics in full length—which can amount to megabytes or gigabytes of data—separately from a 12-page description published in the book. More information can be found on [www.CICLing.org/why\\_verify.htm](http://www.CICLing.org/why_verify.htm).

To encourage providing algorithms and data along with the published papers, we selected a winner of our Verifiability, Reproducibility, and Working Description Award. The main factors in choosing the awarded submission were technical correctness and completeness, readability of the code and documentation, simplicity of installation and use, and exact correspondence to the claims of the



**Fig. 1.** Submissions by country or region. The area of a circle represents the number of submitted papers.

paper. Unnecessary sophistication of the user interface was discouraged; novelty and usefulness of the results were not evaluated—instead, they were evaluated for the paper itself and not for the data. This year’s winning paper was published in a separate proceedings volume and is not included in this set.

The following papers received the Best Paper Awards, the Best Student Paper Award, as well as the Verifiability, Reproducibility, and Working Description Award, correspondingly (the best student paper was selected among papers of which the first author was a full-time student, excluding the papers that received a Best Paper Award):

- 1<sup>st</sup> Place: *Automatic Detection of Idiomatic Clauses*, by Anna Feldman and Jing Peng, USA;
- 2<sup>nd</sup> Place: *Topic-Oriented Words as Features for Named Entity Recognition*, by Ziqi Zhang, Trevor Cohn, and Fabio Ciravegna, UK;
- 3<sup>rd</sup> Place: *Five Languages are Better than One: An Attempt to Bypass the Data Acquisition Bottleneck for WSD*, by Els Lefever, Veronique Hoste, and Martine De Cock, Belgium;
- Student: *Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English-Latvian IT Localisation*, by Mārcis Pinnis, Inguna Skadiņa, and Andrejs Vasiljevs, Latvia;
- Verifiability: *Linguistically-Driven Selection of Correct Arcs for Dependency Parsing*, by Felice Dell’Orletta, Giulia Venturi, and Simonetta Montemagni, Italy.

The authors of the awarded papers (except for the Verifiability Award) were given extended time for their presentations. In addition, the Best Presentation Award and the Best Poster Award winners were selected by a ballot among the attendees of the conference.

Besides its high scientific level, one of the success factors of CICLing conferences is their excellent cultural program. The attendees of the conference had a chance to visit unique historical places: the Greek island of Samos, the birthplace

of Pythagoras (Pythagorean theorem!), Aristarchus (who first realized that the Earth rotates around the Sun and not vice versa), and Epicurus (one of the founders of the scientific method); the Greek island of Patmos, where John the Apostle received his visions of the Apocalypse; and the huge and magnificent archeological site of Ephesus in Turkey, where stood the Temple of Artemis, one of the Seven Wonders of the World (destroyed by Herostratus), and where the Virgin Mary is believed to have spent the last years of her life.

I would like to thank all those involved in the organization of this conference. In the first place these are the authors of the papers that constitute this book: it is the excellence of their research work that gives value to the book and sense to the work of all other people. I thank all those who served on the Program Committee, Software Reviewing Committee, Award Selection Committee, as well as additional reviewers, for their hard and very professional work. Special thanks go to Ted Pedersen, Adam Kilgariff, Viktor Pekar, Ken Church, Horacio Rodriguez, Grigori Sidorov, and Tamar Solorio for their invaluable support in the reviewing process.

I would like to thank the conference staff, volunteers, and the members of the local organization committee headed by Dr. Efstathios Stamatatos. In particular, we are grateful to Dr. Ergina Kavallieratou for her great effort in planning the cultural program and Mrs. Manto Katsiani for her invaluable secretarial and logistics support. We are deeply grateful to the Department of Information and Communication Systems Engineering of the University of the Aegean for its generous support and sponsorship. Special thanks go to the Union of Vinicultural Cooperatives of Samos (EOSS), A. Giannoulis Ltd., and the Municipality of Samos for their kind sponsorship. We also acknowledge the support received from the project WIQ-EI (FP7-PEOPLE-2010-IRSES: Web Information Quality Evaluation Initiative).

The entire submission and reviewing process was supported for free by the EasyChair system ([www.EasyChair.org](http://www.EasyChair.org)). Last but not least, I deeply appreciate the Springer staff's patience and help in editing these volumes and getting them printed in record short time—it is always a great pleasure to work with Springer.

February 2013

Alexander Gelbukh



# Organization

CICLing 2013 is hosted by the University of the Aegean and is organized by the CICLing 2013 Organizing Committee in conjunction with the Natural Language and Text Processing Laboratory of the CIC (Centro de Investigación en Computación) of the IPN (Instituto Politécnico Nacional), Mexico.

## Organizing Chair

Efstathios Stamatatos

## Organizing Committee

Efstathios Stamatatos (chair)

Ergina Kavallieratou

Manolis Maragoudakis

## Program Chair

Alexander Gelbukh

## Program Committee

Ajith Abraham

Marianna Apidianaki

Bogdan Babych

Ricardo Baeza-Yates

Kalika Bali

Sivaji Bandyopadhyay

Srinivas Bangalore

Leslie Barrett

Roberto Basili

Anja Belz

Pushpak Bhattacharyya

Igor Boguslavsky

António Branco

Nicoletta Calzolari

Nick Campbell

Michael Carl

Ken Church

Dan Cristea

Walter Daelemans

Anna Feldman

Alexander Gelbukh (chair)

Gregory Grefenstette

Eva Hajicova

Yasunari Harada

Koiti Hasida

Iris Hendrickx

Ales Horak

Veronique Hoste

Nancy Ide

Diana Inkpen

Hitoshi Isahara

Sylvain Kahane

Alma Kharrat

Adam Kilgariff

Philipp Koehn  
Valia Kordoni  
Leila Kosseim  
Mathieu Lafourcade  
Kristen Lindén  
Elena Lloret  
Bente Maegaard  
Bernardo Magnini  
Cerstin Mahlow  
Sun Maosong  
Katja Markert  
Diana Mccarthy  
Rada Mihalcea  
Jean-Luc Minel  
Ruslan Mitkov  
Dunja Mladenic  
Marie-Francine Moens  
Masaki Murata  
Preslav Nakov  
Vivi Nastase  
Costanza Navarretta  
Roberto Navigli  
Vincent Ng  
Kjetil Nørvåg  
Constantin Orasan  
Ekaterina Ovchinnikova  
Ted Pedersen  
Viktor Pekar  
Anselmo Peñas  
Maria Pinango  
Octavian Popescu

Irina Prodanof  
James Pustejovsky  
German Rigau  
Fabio Rinaldi  
Horacio Rodriguez  
Paolo Rosso  
Vasile Rus  
Horacio Saggion  
Franco Salvetti  
Roser Sauri  
Hinrich Schütze  
Satoshi Sekine  
Serge Sharoff  
Grigori Sidorov  
Kiril Simov  
Vaclav Snasel  
Thamar Solorio  
Lucia Specia  
Efstathios Stamatatos  
Josef Steinberger  
Ralf Steinberger  
Vera Lúcia Strube de Lima  
Mike Thelwall  
George Tsatsaronis  
Dan Tufis  
Olga Uryupina  
Karin Verspoor  
Manuel Vilares Ferro  
Aline Villavicencio  
Piotr W. Fuglewicz  
Annie Zaenen

## Software Reviewing Committee

Ted Pedersen  
Florian Holz  
Miloš Jakubíček

Sergio Jiménez Vargas  
Miikka Silfverberg  
Ronald Winnemöller

## Award Committee

Alexander Gelbukh  
Eduard Hovy  
Rada Mihalcea

Ted Pedersen  
Yorick Wiks

## Additional Referees

Rodrigo Agerri  
 Katsiaryna Aharodnik  
 Ahmed Ali  
 Tanveer Ali  
 Alexandre Allauzen  
 Maya Ando  
 Javier Artiles  
 Wilker Aziz  
 Vt Baisa  
 Alexandra Balahur  
 Somnath Banerjee  
 Liliana Barrio-Alvers  
 Adrián Blanco  
 Francis Bond  
 Dave Carter  
 Chen Chen  
 Jae-Woong Choe  
 Simon Clematide  
 Geert Coorman  
 Victor Darriba  
 Dipankar Das  
 Orphee De Clercq  
 Ariani Di Felippo  
 Maud Ehrmann  
 Daniel Eisinger  
 Ismail El Maarouf  
 Tilia Ellendorff  
 Milagros Fernández Gavilanes  
 Santiago Fernández Lanza  
 Daniel Fernández-González  
 Karèn Fort  
 Koldo Gojenola  
 Gintare Grigonyte  
 Masato Hagiwara  
 Kazi Saidul Hasan  
 Eva Hasler  
 Stefan Höfler  
 Chris Hokamp  
 Adrian Iftene  
 Iustina Ilisei  
 Leonid Iomdin  
 Milos Jakubicek  
 Francisco Javier Guzman

Nattiya Kanhabua  
 Aharodnik Katya  
 Kurt Keena  
 Natalia Konstantinova  
 Vojtech Kovar  
 Kow Kuroda  
 Gorka Labaka  
 Shibamouli Lahiri  
 Egoitz Laparra  
 Els Lefever  
 Lucelene Lopes  
 Oier López de La Calle  
 John Lowe  
 Shamima Mithun  
 Tapabrata Mondal  
 Silvia Moraes  
 Mihai Alex Moruz  
 Koji Murakami  
 Sofia N. Galicia-Haro  
 Vasek Nemcik  
 Zuzana Neverilova  
 Anthony Nguyen  
 Inna Novalija  
 Neil O'Hare  
 John Osborne  
 Santanu Pal  
 Feng Pan  
 Thiago Pardo  
 Veronica Perez Rosas  
 Michael Piotrowski  
 Ionut Cristian Pistol  
 Soujanya Poria  
 Luz Rello  
 Noushin Rezapour Asheghi  
 Francisco Ribadas-Pena  
 Alexandra Roshchina  
 Tobias Roth  
 Jan Rupnik  
 Upendra Sapkota  
 Gerold Schneider  
 Djamé Seddah  
 Keiji Shinzato  
 João Silva

Sara Silveira  
Sen Soori  
Sanja Štajner  
Tadej Štajner  
Zofia Stankiewicz  
Hristo Tanev  
Irina Temnikova  
Mitja Trampus  
Diana Trandabat  
Yasushi Tsubota

Srinivas Vadrevu  
Francisco Viveros Jiménez  
Josh Weissbock  
Clarissa Xavier  
Victoria Yaneva  
Manuela Yapomo  
Hikaru Yokono  
Taras Zagibalov  
Vanni Zavarella  
Alisa Zhila

## Website and Contact

The webpage of the CICLing conference series is [www.CICLing.org](http://www.CICLing.org). It contains information about past CICLing conferences and their satellite events, including published papers or their abstracts, photos, video recordings of keynote talks, as well as information about the forthcoming CICLing conferences and contact options.

# Table of Contents – Part I

## General Techniques

Unsupervised Feature Adaptation for Cross-Domain NLP with an Application to Compositionality Grading . . . . .	1
<i>Lukas Michelbacher, Qi Han, and Hinrich Schütze</i>	
Syntactic Dependency-Based N-grams: More Evidence of Usefulness in Classification . . . . .	13
<i>Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández</i>	

## Lexical Resources

A Quick Tour of BabelNet 1.1 (Invited paper) . . . . .	25
<i>Roberto Navigli</i>	
Automatic Pipeline Construction for Real-Time Annotation . . . . .	38
<i>Henning Wachsmuth, Mirko Rose, and Gregor Engels</i>	
A Multilingual GRUG Treebank for Underresourced Languages . . . . .	50
<i>Oleg Kapanadze and Alla Mishchenko</i>	
Creating an Annotated Corpus for Extracting Canonical Citations from Classics-Related Texts by Using Active Annotation . . . . .	60
<i>Matteo Romanello</i>	
Approaches of Anonymisation of an SMS Corpus . . . . .	77
<i>Namrata Patel, Pierre Accorsi, Diana Inkpen, Cédric Lopez, and Mathieu Roche</i>	
A Corpus Based Approach for the Automatic Creation of Arabic Broken Plural Dictionaries . . . . .	89
<i>Samhaa R. El-Beltagy and Ahmed Rafea</i>	
Temporal Classifiers for Predicting the Expansion of Medical Subject Headings . . . . .	98
<i>George Tsatsaronis, Iraklis Varlamis, Nattiya Kanhabua, and Kjetil Nørvåg</i>	
Knowledge Discovery on Incompatibility of Medical Concepts . . . . .	114
<i>Adam Grycner, Patrick Ernst, Amy Siu, and Gerhard Weikum</i>	
Extraction of Part-Whole Relations from Turkish Corpora . . . . .	126
<i>Tuğba Yıldız, Savaş Yıldırım, and Banu Diri</i>	

Chinese Terminology Extraction Using EM-Based Transfer Learning  
Method . . . . . 139  
*Yanxia Qin, Dequan Zheng, Tiejun Zhao, and Min Zhang*

Orthographic Transcription for Spoken Tunisian Arabic . . . . . 153  
*Inès Zribi, Marwa Graja, Mariem Ellouze Khmekhem,  
Maher Jaoua, and Lamia Hadrich Belguith*

**Morphology and Tokenization**

An Improved Stemming Approach Using HMM for a Highly Inflectional  
Language . . . . . 164  
*Navanath Saharia, Kishori M. Konwar, Utpal Sharma, and  
Jugal K. Kalita*

Semi-automatic Acquisition of Two-Level Morphological Rules for Iban  
Language . . . . . 174  
*Suhaila Sae, Lay-Ki Soon, Tek Yong Lim,  
Bali Ranaivo-Malançon, and Enya Kong Tang*

Finite State Morphology for Amazigh Language . . . . . 189  
*Fatima Zahra Nejme, Siham Boulaknadel, and Driss Aboutajdine*

New Perspectives in Sinographic Language Processing through the Use  
of Character Structure . . . . . 201  
*Yannis Haralambous*

The Application of Kalman Filter Based Human-Computer Learning  
Model to Chinese Word Segmentation . . . . . 218  
*Weimeng Zhu, Ni Sun, Xiaojun Zou, and Junfeng Hu*

Machine Learning for High-Quality Tokenization Replicating Variable  
Tokenization Schemes . . . . . 231  
*Murhaf Fares, Stephan Oepen, and Yi Zhang*

**Syntax and Named Entity Recognition**

Structural Prediction in Incremental Dependency Parsing . . . . . 245  
*Niels Beuck and Wolfgang Menzel*

Semi-supervised Constituent Grammar Induction Based on Text  
Chunking Information . . . . . 258  
*Jesús Santamaría and Lourdes Araujo*

Turkish Constituent Chunking with Morphological and Contextual  
Features . . . . . 270  
*İlknur Durgar El-Kahlout and Ahmet Afşın Akın*

Enhancing Czech Parsing with Verb Valency Frames . . . . .	282
<i>Miloš Jakubíček and Vojtěch Kovář</i>	
An Automatic Approach to Treebank Error Detection Using a Dependency Parser . . . . .	294
<i>Bhasha Agrawal, Rahul Agarwal, Samar Husain, and Dipti M. Sharma</i>	
Topic-Oriented Words as Features for Named Entity Recognition (Best Paper Award, Second Place) . . . . .	304
<i>Ziqi Zhang, Trevor Cohn, and Fabio Ciravegna</i>	
Named Entities in Judicial Transcriptions: Extended Conditional Random Fields . . . . .	317
<i>Elisabetta Fersini and Enza Messina</i>	
Introducing Baselines for Russian Named Entity Recognition . . . . .	329
<i>Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov</i>	

## Word Sense Disambiguation and Coreference Resolution

Five Languages Are Better Than One: An Attempt to Bypass the Data Acquisition Bottleneck for WSD (Best Paper Award, Third Place) . . . . .	343
<i>Els Lefever, Véronique Hoste, and Martine De Cock</i>	
Analyzing the Sense Distribution of Concordances Obtained by Web as Corpus Approach . . . . .	355
<i>Xabier Saralegi and Pablo Gamallo</i>	
MaxMax: A Graph-Based Soft Clustering Algorithm Applied to Word Sense Induction . . . . .	368
<i>David Hope and Bill Keller</i>	
A Model of Word Similarity Based on Structural Alignment of Subject-Verb-Object Triples . . . . .	382
<i>Dervla O’Keeffe and Fintan Costello</i>	
Coreference Annotation Schema for an Inflectional Language . . . . .	394
<i>Maciej Ogrodniczuk, Magdalena Zawislawska, Katarzyna Głowińska, and Agata Savary</i>	
Exploring Coreference Uncertainty of Generically Extracted Event Mentions . . . . .	408
<i>Goran Glavaš and Jan Šnajder</i>	

## Semantics and Discourse

LIAR <sub>C</sub> : Labeling Implicit ARGuments in Spanish Deverbal Nominalizations .....	423
<i>Aina Peris, Mariona Taulé, Horacio Rodríguez, and Manuel Bertran Ibarz</i>	
Automatic Detection of Idiomatic Clauses (Best Paper Award, First Place) .....	435
<i>Anna Feldman and Jing Peng</i>	
Evaluating the Results of Methods for Computing Semantic Relatedness .....	447
<i>Felice Ferrara and Carlo Tasso</i>	
Similarity Measures Based on Latent Dirichlet Allocation .....	459
<i>Vasile Rus, Nobal Niraula, and Rajendra Banjade</i>	
Evaluating the Premises and Results of Four Metaphor Identification Systems .....	471
<i>Jonathan Dunn</i>	
Determining the Conceptual Space of Metaphoric Expressions .....	487
<i>David B. Bracewell, Marc T. Tomlinson, and Michael Mohler</i>	
What is being Measured in an Information Graphic? .....	501
<i>Seniz Demir, Stephanie Elzer Schwartz, Richard Burns, and Sandra Carberry</i>	
Comparing Discourse Tree Structures .....	513
<i>Elena Mitocariu, Daniel Alexandru Anechitei, and Dan Cristea</i>	
Assessment of Different Workflow Strategies for Annotating Discourse Relations: A Case Study with HDRB .....	523
<i>Himanshu Sharma, Praveen Dakwale, Dipti M. Sharma, Rashmi Prasad, and Aravind Joshi</i>	
Building a Discourse Parser for Informal Mathematical Discourse in the Context of a Controlled Natural Language .....	533
<i>Raúl Ernesto Gutiérrez de Piñerez Reyes and Juan Francisco Díaz Frias</i>	
Discriminative Learning of First-Order Weighted Abduction from Partial Discourse Explanations .....	545
<i>Kazeto Yamamoto, Naoya Inoue, Yotaro Watanabe, Naoaki Okazaki, and Kentaro Inui</i>	



Facilitating the Analysis of Discourse Phenomena in an Interoperable  
 NLP Platform ..... 559  
*Riza Theresa Batista-Navarro, Georgios Kontonatsios,  
 Claudiu Mihăilă, Paul Thompson, Rafal Rak, Raheel Nawaz,  
 Ioannis Korkontzelos, and Sophia Ananiadou*

**Author Index** ..... 573

## Table of Contents – Part II

### Sentiment, Polarity, Emotion, Subjectivity, and Opinion

Damping Sentiment Analysis in Online Communication: Discussions, Monologs and Dialogs (Invited paper) . . . . .	1
<i>Mike Thelwall, Kevan Buckley, George Paltoglou, Marcin Skowron, David Garcia, Stephane Gobron, Junghyun Ahn, Arvid Kappas, Dennis Küster, and Janusz A. Holyst</i>	
Optimal Feature Selection for Sentiment Analysis . . . . .	13
<i>Basant Agarwal and Namita Mittal</i>	
Measuring the Effect of Discourse Structure on Sentiment Analysis . . . . .	25
<i>Baptiste Chardon, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher</i>	
Lost in Translation: Viability of Machine Translation for Cross Language Sentiment Analysis . . . . .	38
<i>Balamurali A.R., Mitesh M. Khapra, and Pushpak Bhattacharyya</i>	
An Enhanced Semantic Tree Kernel for Sentiment Polarity Classification . . . . .	50
<i>Luis A. Trindade, Hui Wang, William Blackburn, and Niall Rooney</i>	
Combining Supervised and Unsupervised Polarity Classification for non-English Reviews . . . . .	63
<i>José M. Perea-Ortega, Eugenio Martínez-Cámara, María-Teresa Martín-Valdivia, and L. Alfonso Ureña-López</i>	
Word Polarity Detection Using a Multilingual Approach . . . . .	75
<i>Cüneyd Murad Özsert and Arzucan Özgür</i>	
Mining Automatic Speech Transcripts for the Retrieval of Problematic Calls . . . . .	83
<i>Frederik Cailliau and Ariane Cavet</i>	
Cross-Lingual Projections vs. Corpora Extracted Subjectivity Lexicons for Less-Resourced Languages . . . . .	96
<i>Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu</i>	
Predicting Subjectivity Orientation of Online Forum Threads . . . . .	109
<i>Prakhar Biyani, Cornelia Caragea, and Prasenjit Mitra</i>	

Distant Supervision for Emotion Classification with Discrete Binary Values . . . . .	121
<i>Jared Suttles and Nancy Ide</i>	
Using Google n-grams to Expand Word-Emotion Association Lexicon . . . . .	137
<i>Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj</i>	
A Joint Prediction Model for Multiple Emotions Analysis in Sentences . . . . .	149
<i>Yunong Wu, Kenji Kita, Kazuyuki Matsumoto, and Xin Kang</i>	
Evaluating the Impact of Syntax and Semantics on Emotion Recognition from Text . . . . .	161
<i>Gözde Özbal and Daniele Pighin</i>	
Chinese Emotion Lexicon Developing via Multi-lingual Lexical Resources Integration . . . . .	174
<i>Jun Xu, Rui Feng Xu, Yanzhen Zheng, Qin Lu, Kai-Fai Wong, and Xiaolong Wang</i>	
N-Gram-Based Recognition of Threatening Tweets . . . . .	183
<i>Nelleke Oostdijk and Hans van Halteren</i>	
Distinguishing the Popularity between Topics: A System for Up-to-Date Opinion Retrieval and Mining in the Web . . . . .	197
<i>Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos</i>	
<b>Machine Translation and Multilingualism</b>	
No Free Lunch in Factored Phrase-Based Machine Translation . . . . .	210
<i>Aleš Tamchyna and Ondřej Bojar</i>	
Domain Adaptation in Statistical Machine Translation Using Comparable Corpora: Case Study for English Latvian IT Localisation (Best Student Paper Award) . . . . .	224
<i>Mārcis Pinnis, Inguna Skadiņa, and Andrejs Vasiļjevs</i>	
Assessing the Accuracy of Discourse Connective Translations: Validation of an Automatic Metric . . . . .	236
<i>Najeh Hajlaoui and Andrei Popescu-Belis</i>	
An Empirical Study on Word Segmentation for Chinese Machine Translation . . . . .	248
<i>Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu</i>	
Class-Based Language Models for Chinese-English Parallel Corpus . . . . .	264
<i>Junfei Guo, Juan Liu, Michael Walsh, and Helmut Schmid</i>	

Building a Bilingual Dictionary from a Japanese-Chinese Patent Corpus .....	276
<i>Keiji Yasuda and Eiichiro Sumita</i>	
A Diagnostic Evaluation Approach for English to Hindi MT Using Linguistic Checkpoints and Error Rates .....	285
<i>Renu Balyan, Sudip Kumar Naskar, Antonio Toral, and Niladri Chatterjee</i>	
Leveraging Arabic-English Bilingual Corpora with Crowd Sourcing-Based Annotation for Arabic-Hebrew SMT .....	297
<i>Manish Gaurav, Guruprasad Saikumar, Amit Srivastava, Premkumar Natarajan, Shankar Ananthakrishnan, and Spyros Matsoukas</i>	
Automatic and Human Evaluation on English-Croatian Legislative Test Set .....	311
<i>Marija Brkić, Sanja Seljan, and Tomislav Vičić</i>	
<b>Text Mining, Information Extraction, and Information Retrieval</b>	
Enhancing Search: Events and Their Discourse Context (Invited paper) .....	318
<i>Sophia Ananiadou, Paul Thompson, and Raheel Nawaz</i>	
Distributional Term Representations for Short-Text Categorization .....	335
<i>Juan Manuel Cabrera, Hugo Jair Escalante, and Manuel Montes-y-Gómez</i>	
Learning Bayesian Network Using Parse Trees for Extraction of Protein-Protein Interaction .....	347
<i>Pedro Nelson Shiguihara-Juárez and Alneu de Andrade Lopes</i>	
A Model for Information Extraction in Portuguese Based on Text Patterns .....	359
<i>Tiago Luis Bonamigo and Renata Vieira</i>	
A Study on Query Expansion Based on Topic Distributions of Retrieved Documents .....	369
<i>Midori Serizawa and Ichiro Kobayashi</i>	
Link Analysis for Representing and Retrieving Legal Information .....	380
<i>Alfredo López Monroy, Hiram Calvo, Alexander Gelbukh, and Georgina García Pacheco</i>	

**Text Summarization**

Discursive Sentence Compression . . . . .	394
<i>Alejandro Molina, Juan-Manuel Torres-Moreno, Eric SanJuan, Iria da Cunha, and Gerardo Eugenio Sierra Martínez</i>	
A Knowledge Induced Graph-Theoretical Model for Extract and Abstract Single Document Summarization . . . . .	408
<i>Niraj Kumar, Kannan Srinathan, and Vasudeva Varma</i>	
Hierarchical Clustering in Improving Microblog Stream Summarization . . . . .	424
<i>Andrei Olariu</i>	
Summary Evaluation: Together We Stand NPower-ed . . . . .	436
<i>George Giannakopoulos and Vangelis Karkaletsis</i>	

**Stylometry and Text Simplification**

Explanation in Computational Stylometry (Invited paper) . . . . .	451
<i>Walter Daelemans</i>	
The Use of Orthogonal Similarity Relations in the Prediction of Authorship . . . . .	463
<i>Upendra Sapkota, Thamar Solorio, Manuel Montes-y-Gómez, and Paolo Rosso</i>	
ERNESTA: A Sentence Simplification Tool for Children’s Stories in Italian . . . . .	476
<i>Gianni Barlacchi and Sara Tonelli</i>	
Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules . . . . .	488
<i>Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion</i>	
The Impact of Lexical Simplification by Verbal Paraphrases for People with and without Dyslexia . . . . .	501
<i>Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion</i>	
Detecting Apposition for Text Simplification in Basque . . . . .	513
<i>Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Ander Soraluze</i>	

**Applications**

Automation of Linguistic Creativitas for Adslogia . . . . .	525
<i>Gözde Özbal and Carlo Strapparava</i>	

Allongos: Longitudinal Alignment for the Genetic Study of Writers’ Drafts . . . . .	537
<i>Adrien Lardilleux, Serge Fleury, and Georgeta Cislaru</i>	
A Combined Method Based on Stochastic and Linguistic Paradigm for the Understanding of Arabic Spontaneous Utterances . . . . .	549
<i>Chahira Lhioui, Anis Zouaghi, and Mounir Zrigui</i>	
Evidence in Automatic Error Correction Improves Learners’ English Skill . . . . .	559
<i>Jiro Umezawa, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui</i>	
<b>Author Index . . . . .</b>	<b>573</b>