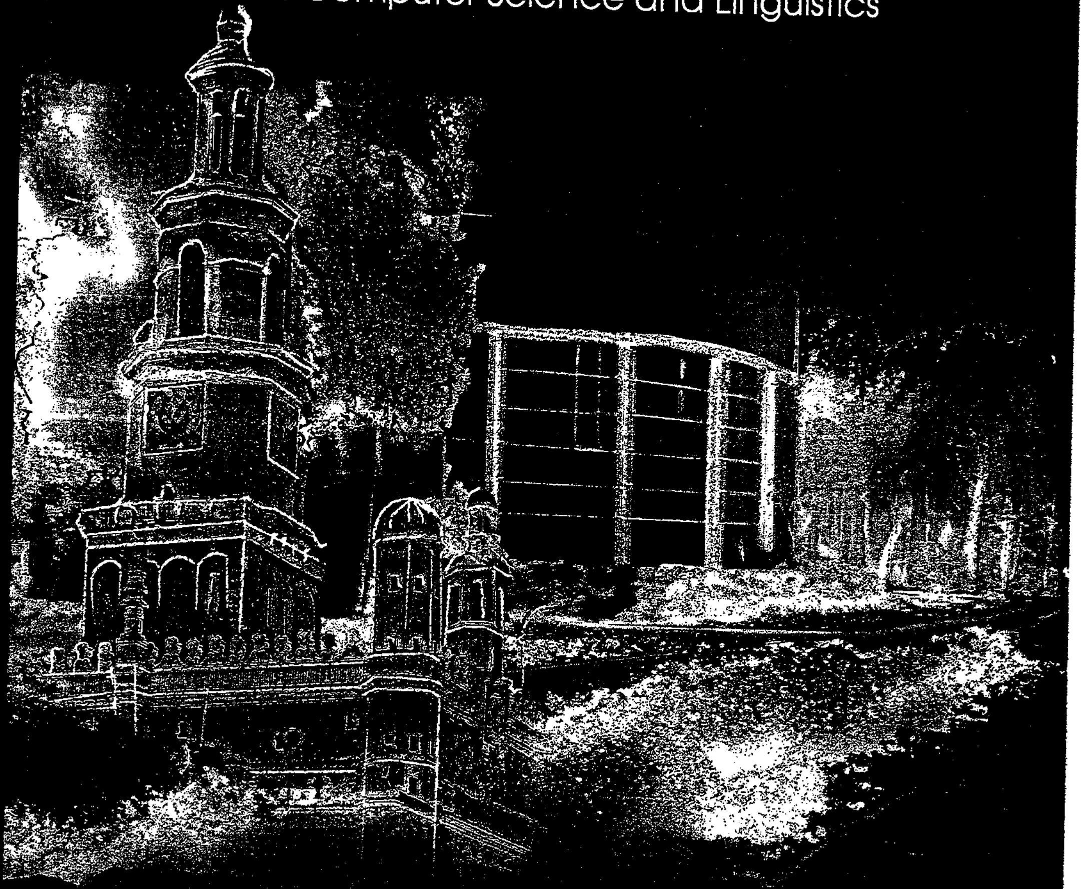
Human Language Technologies as a Challenge for Computer Science and Linguistics



Proceedings

of 3rd Language & Technology Conference
October 5-7, 2007, Poznań, Poland

UAM

FUNDACJA

Zygmunt Vetulani (ed.)

ORGANIZING COMMITTEE

Patronage
Mr Ryszard Grobelny
Mayor of Poznań

CO-OPERATING ORGANIZATIONS

ELRA
EUROPEAN LANGUAGE RESOURCES ASSOCIATION

MIĘDZYNARODOWE
TARGI
POZNAŃSKIE

PPBW

PEARSON
Education

ISBN 978-83-7177-407-2

Human Language Technologies as a Challenge for Computer Science and Linguistics

3rd Language & Technology Conference
October 5-7, 2007, Poznań, Poland

Proceedings

Zygmunt Vetulani (ed.)

Poznań 2007

LTC website: http://www.ltc.amu.edu.pl

Publisher:
Wydawnictwo Poznańskie Sp. z o.o.
with co-operation of
Fundacja Uniwersytetu im. A. Mickiewicza
ul. Rubież 46
61-612 Poznań, Poland
tel. +48 61 827 97 00
e-mail: fundacja@amu.edu.pl

Distributed by:
Wydawnictwo Poznańskie Sp. z o.o.
ul. Fredry 8
61-701 Poznań
tel: +48 61 852 66 05; dział handlowy/commercial department: +48 61 852 38 44
fax: +48 61 853 80 75
e-mail: wydawnictwo@wydawnictwo-poznanskie.pl
and
ELRA/ELDA
55-57 rue Brillat Savarin
75013 Paris, France
Tel.: +33 143 13 33 33
Fax: +33 143 13 33 30
e-mail: info@elda.fr
website: http://www.elra.info or http://www.elda.fr

Contact person for distribution:
Zygmunt Vetulani
e-mail: vetulani@amu.edu.pl
website: http://www.amu.edu.pl/~vetulani

ISBN 978-83-7177-407-2

Technical editors:
Paweł Konieczka
Justyna Walkowska

Cover design and technical preparation:
IMPRESJA Wydawnictwa Elektroniczne S.A.
http://www.impresja.com.pl

This volume has been compiled from the pdf files supplied by the Authors.

CONTENTS

LTC'07 Committees 4
List of Reviewers 5
Conference Program 6
Acknowledgments 15
Preface by Zygmunt Vetulani 16
Invited Lectures (Abstracts) 19
Technical Papers 23
Demo Session 513
  Full Papers 514
  Abstracts 524
Homeland Security Workshop 535
  Homeland Security Platform Presentation 536
  Panel Discussion 537
  Technical Papers 539
  Poster Session Abstracts 554
  Demos 573
Author Index 576

PROGRAM COMMITTEE

Zygmunt Vetulani (Adam Mickiewicz University, Poznań, Poland) – Chair
Victoria Arranz (ELRA, France)
Anja Belz (University of Brighton, UK)
Janusz Bień (Warsaw University, Poland)
Christian Boitet (IMAG, France)
Leonard Bolc (IPI PAN, Poland)
Nicoletta Calzolari (ILC/CNR, Italy)
Nick Campbell (ATR, Japan)
Julie Carson-Berndsen (University College Dublin, Ireland)
Khalid Choukri (ELRA, France)
Adam Dąbrowski (Poznań University of Technology, Poland)
Grażyna Demenko (Adam Mickiewicz University, Poznań, Poland)
Elżbieta Dura (University of Skovde, Sweden)
Katarzyna Dziubalska-Kołaczyk (Adam Mickiewicz University, Poznań, Poland)
Tomaz Erjavec (Josef Stefan Institute, Slovenia)
Cedrick Fairon (University of Louvain, Belgium)
Christiane Fellbaum (Princeton University, USA)
Maria Gavrilidou (ILSP, Greece)
Aleksander Gerd (St. Petersburg State University, Russia)
Dafydd Gibbon (University of Bielefeld, Germany)
Stefan Grocholewski (PTI/Poznań University of Technology, Poland)
Franz Guenthner (Ludwig-Maximilians-University München, Germany)
Roland Hausser (Erlangen, Germany)
Wacław Iszkowski (PIIT, Poland)
Margaret King (University of Geneva, Switzerland)
Orest Kossak (Technical University Lviv/Ericpol Telecom, Ukraine)
Eric Laporte (University Marne-la-Vallee, France)
Gerard Ligozat (LIMSI/CNRS, France)
Natalia Loukachevitch (Research Computing Center of Moscow State University, Russia)
Wiesław Lubaszewski (AGH/UJ, Poland)
Bente Maegaard (Centre for Language Technology, Denmark)
Joseph Mariani (LIMSI-CNRS, France)
Jacek Martinek (Poznań University of Technology, Poland)
Václav Matousek (University of West Bohemia in Pilsen, Czech Rep.)
Keith J. Miller (MITRE, USA)
Nicholas Ostler (Linguacubun Ltd., UK)
Karel Pala (Masaryk University, Czech Rep.)
Pavel S. Pankov (National Academy of Sciences, Kyrgyzstan)
Marcin Paprzycki (IBS PAN, Warsaw, Poland)
Patrick Paroubek (LIMSI-CNRS, France)
Emil Pływaczewski (University of Białystok, Poland)
Adam Przepiórkowski (IPI PAN, Poland)
Reinhard Rapp (University Mainz, Germany)
Mike Rosner (University of Malta)
Justus Roux (University of Stellenbosch, South Africa)
Vasile Rus (University of Memphis, Fedex Inst. of Technology, USA)
Frédérique Ségond (Xerox, France)
Włodzimierz Sobkowiak (Adam Mickiewicz University, Poznań, Poland)
Marek Świdziński (University of Warsaw, Poland)
Ryszard Tadeusiewicz (AGH, Poland)
Dan Tufiş (RCAI, Romania)
Hans Uszkoreit (DFKI, Germany)
Piek Vossen (University of Amsterdam, Netherlands)
Tom Wachtel (Independent Consultant, Italy)
Jan Węglarz (Poznań University of Technology, Poland)
Richard Zuber (CNRS, France)

ORGANIZING COMMITTEE

Zygmunt Vetulani – Chair
Filip Graliński
Paweł Konieczka
Maciej Lison
Jacek Marciniak
Tomasz Obrębski
Justyna Walkowska

LIST OF REVIEWERS

Victoria Arranz (ELRA, France)
Anja Belz (University of Brighton, UK)
Julie Berndsen (UCD School of Computer Science and Informatics, Ireland)
Janusz Bień (Warsaw University, Poland)
Busemann Stephan (DFKI GmbH Language Technology Lab)
Nicoletta Calzolari (ILC/CNR, Italy)
Nick Campbell (ATR, Japan)
Irena Chmielewska (Poznań University of Technology, Poland)
Berthold Crysmann (Universität des Saarlandes, Germany)
Jolanta Cybulka (Poznań University of Technology, Poland)
Adam Dąbrowski (Poznań University of Technology, Poland)
Grażyna Demenko (Adam Mickiewicz University, Poland)
Elżbieta Dura (University of Skovde, Sweden)
Katarzyna Dziubalska-Kołaczyk (Adam Mickiewicz University, Poland)
Tomaž Erjavec (Josef Stefan Institute, Slovenia)
Cédrick Fairon (University of Louvain, Belgium)
Christiane Fellbaum (Princeton University, USA)
Maria Gavrilidou (ILSP, Greece)
Dafydd Gibbon (University of Bielefeld, Germany)
Filip Graliński (Poleng Sp. z o.o., Poland)
Stefan Grocholewski (PTI/Poznań University of Technology, Poland)
Franz Guenthner (Ludwig-Maximilians-University München, Germany)
Elżbieta Hajnicz (Polish Academy of Sciences, Poland)
Roland Hausser (Erlangen, Germany)
Aleš Horák (Masaryk University, Czech Republic)
Maciej Kandulski (Adam Mickiewicz University, Poland)
Walter Kasper (DFKI GmbH Language Technology Lab, Germany)
Laurent Kevers (Cental, Belgium)
Valia Kordoni (Saarland University/DFKI GmbH, Germany)
Orest Kossak (Technical University Lviv/Ericpol Telecom, Ukraine)
Ilona Koutny (Adam Mickiewicz University, Poland)
Eric Laporte (University Marne-la-Vallee, France)
Gérard Ligozat (LIMSI/CNRS, France)
Maciej Lison (Adam Mickiewicz University, Poland)
Natalia Loukachevitch (Moscow State University, Russia)
Wiesław Lubaszewski (AGH/UJ, Poland)
Bente Maegaard (Centre for Language Technology, Denmark)
Jacek Marciniak (Adam Mickiewicz University, Poland)
Joseph Mariani (LIMSI-CNRS, France)
Domen Marincic (Jozef Stefan Institute, Slovenia)
Jacek Martinek (Poznań University of Technology, Poland)
Piet Mertens (K.U. Leuven, Belgium)
Keith J. Miller (MITRE, USA)
Guenter Neumann (DFKI GmbH Language Technology Lab, Germany)
Tomasz Obrębski (Adam Mickiewicz University, Poland)
Nicholas Ostler (Linguacubun Ltd., UK)
Karel Pala (Masaryk University, Czech Republic)
Pavel Pankov (National Academy of Sciences, Kyrgyzstan)
Tadeusz Pankowski (Adam Mickiewicz University, Poland)
Marcin Paprzycki (IBS PAN, Warsaw, Poland)
Patrick Paroubek (LIMSI-CNRS, France)
Adam Przepiórkowski (IPI PAN, Poland)
Reinhard Rapp (University Mainz, Germany)
Mike Rosner (University of Malta)
Justus Roux (University of Stellenbosch, South Africa)
Vasile Rus (University of Memphis, Fedex Inst. of Technology, USA)
Frédérique Ségond (Xerox Research Centre Europe, France)
Włodzimierz Sobkowiak (Adam Mickiewicz University, Poland)
Marek Świdziński (University of Warsaw, Poland)
Dan Tufiş (Alexandru Ioan Cuza University of Iaşi, Romania)
Hans Uszkoreit (DFKI, Germany)
Zygmunt Vetulani (Adam Mickiewicz University, Poland)
Grażyna Vetulani (Adam Mickiewicz University, Poland)
Piek Vossen (University of Amsterdam, Netherlands)
Tom Wachtel (Independent Consultant, Italy)
Marcin Woliński (Polish Academy of Sciences, Poland)
Maciej Wygralak (Adam Mickiewicz University, Poland)
Richard Zuber (CNRS, France)

CONFERENCE PROGRAM

Day 1: Friday, October 5			
07:00 - 08:30	Registration		
08:30 - 09:00	Opening		
09:00 - 09:15	Break (time to proceed to session rooms)		
09:15 - 10:45	**Technical Sessions:**		
	Information Retrieval/Extraction 1 (IR1)	Computational Morphology 1 (MR1)	Speech Processing 1 (SP1)
10:45 - 11:00	Coffee Break		
11:00 - 12:30	**Technical Sessions:**		
	Information Retrieval/Extraction 2 (IR2)	Computational Semantics 1 (SE1)	Speech Processing 2 (SP2)
12:30 - 13:00	Break (30-minute walk to the City Hall)		
13:00 - 14:00	Reception in the City Hall by Mr Ryszard Grobelny, Mayor of Poznan		
14:00 - 14:05	Official 3rd LTC participants photo in front of the City Hall		
14:05 - 15:30	Lunch Break		
15:30 - 17:00	**Technical Sessions:**		
	Communication 1 (CO1)	WordNet Special Track 1 (WN1)	Speech Processing 3 (SP3)
17:00 - 17:30	Coffee Break		
17:30 - 19:00	**Technical Sessions:**		
	Communication 2 (CO2)	WordNet Special Track 2 (WN2)	Speech Processing 4 (SP4)

Day 2: Saturday, October 6			
08:30 - 09:30	**Technical Sessions:**		
	Digital Language Resources 1 (RS1)	Computational Semantics 2 (SE2)	Parsing 1 (PR1)
09:30 - 10:00	Coffee Break		
10:00 - 10:45	Invited Talk (Piek Vossen, University of Amsterdam – The Global Wordnet Grid: anchoring languages to universal meaning)		
10:45 - 12:15	**Technical Sessions:**		
	Digital Language Resources 2 (RS2)	Computational Semantics 3 (SE3)	Parsing 2 (PR2)
12:15 - 14:00	Lunch Break		
14:00 - 14:15	Presentation of Polish Platform for Homeland Security by Emil Pływaczewski and Zbigniew Rau		
14:15 - 15:00	**Technical Sessions:**		
	Applications for Homeland Security		
15:00 - 15:45	**Wine + Coffee +** Polish Platform for Homeland Security Poster Presentations	**Wine + Coffee +** Demos + **Invited Posters**	
15:45 - 16:00	**Press Conference**		
16:00 - 17:00	Open Panel Discussion: Human Language Technologies in Application to Homeland Security – Vision and Prospects. Panelists: Dafydd Gibbon (moderator), Bernardo Magnini, Keith Miller, Emil Pływaczewski, Kimmo Rossi, Sebastian Serwiak, Zygmunt Vetulani		
17:00 - 20:00	Break		
20:00 - 00:00	Conference Banquet in Sala Lustrzana (the Mirror Hall) at Adam Mickiewicz University's Collegium Minus		

Day 3: Sunday, October 7			
09:00 - 10:30	**Technical Sessions:**		
	Digital Language Resources 3 (RS3)	Information Retrieval/Extraction 3 (IR3)	Language Formalisms 1 (FO1)
10:30 - 11:00	Coffee Break		
11:00 - 11:30	Invited Talk (Kimmo Rossi, European Commission – Language technology in the European programmes and policies)		
11:30 - 13:00	**Technical Sessions:**		
	Information Retrieval/Extraction 4 (IR4)	Computational Morphology 2 (MR2)	Parsing 3 (PR3)
13:00 - 14:15	Lunch Break		
14:15 - 16:00	**Technical Sessions:**		
	Machine Translation 1 (MT1)	Parsing 4 (PR4)	Information Retrieval/Extraction 5 (IR5)
16:00 - 16:30	Coffee Break		
16:30 - 17:00	Closure Ceremony		

Program of Technical Sessions, Invited Talks, Demos and Posters

Day 1: October 5, 2007

8:30 - 9:00 Opening

09:15 - 10:45 Technical Sessions

Session IR1: Information Retrieval/Extraction 1

On the Selection of Negative Examples for Hierarchical Text Categorization / Tiziano Fagni, Fabrizio Sebastiani / Italy 24

Classifying texts through natural language parsing and semantic filtering / Jacques Chauché, Violaine Prince / France 29

Detecting Subject-Shift based on Keygraph for Topic Tracking / Fumiyo Fukumoto, Yoshimi Suzuki / Japan 34

Exploring the Structure of Broadcast News for Topic Segmentation / Rui Amaral, Isabel Trancoso / Portugal 39

Session MR1: Computational Morphology 1

VerbTagGr, Web-based Tool for Statistical Morphosyntactic Recognition of One-word Modern Greek Standard Verb Forms / Penelope Lembessi, Stavroula Kroustalli, Maria Gregoriadou / Greece 44

Extraction of Macedonian Nominal Diminutives / Aleksandar Petrovski, Katerina Zdravkova / Macedonia 49

Containing overgeneration in Zulu computational morphology / Laurette Pretorius, Sonja E. Bosch / South Africa 54

A relational model of Polish inflection / Marcin Woliński / Poland 59

Session SP1: Speech Processing 1: resources

Annotation of Polish spoken dialogs in LUNA project / Agnieszka Mykowiecka, Krzysztof Marasek, Małgorzata Marciniak, Joanna Rabiega-Wiśniewska, Ryszard Gubrynowicz / Poland 64

Creating Large Speech Databases via the WWW – the System Architecture of the German Ph@ttSessionz Web Application / Christoph Draxler, Klaus Jänsch / Germany 69

Triphone Statistics for Polish Language / Bartosz Ziółko, Jakub Gałka, Suresh Manandhar, Richard C. Wilson, Mariusz Ziółko / United Kingdom 74

Digital Talking Books in Multiple Languages and Varieties / Isabel Trancoso, António Serralheiro, Diamantino Caseiro, Céu Viana, Isabel Mascarenhas / Portugal 79

11:00 - 12:30 Technical Sessions

Session IR2: Information Retrieval/Extraction 2

Iterative Person Coreference Using Name Frequency Estimates / Octavian Popescu, Bernardo Magnini / Italy 84

A Decision Tree and Rule-based Learning Model for Anaphora Resolution in Turkish / Savaş Yildirim, Yilmaz Kiliçaslan, Tugba Yildiz / Turkey 89

User Evaluation of a novel technique for word-order error correction / Theologos Athanaselis, Konstantinos Mamouras, Stelios Bakamidis, Ioannis Dologlou / Greece 94

Resources for Information Extraction from Polish texts / Agnieszka Mykowiecka, Anna Kupść, Małgorzata Marciniak, Jakub Piskorski / Poland 99

Session SE1: Computational Semantics 1

Extended Similarity Test for the Evaluation of Semantic Similarity Functions / Maciej Piasecki, Stanisław Szpakowicz, Bartosz Broda / Poland 104

An ontology based semantic library catalogue / Dariusz Daćko, Joanna Józefowska, Agnieszka Ławrynowicz / Poland 109

Mapping out a Semantic Network of Japanese Word Associations through a Combination of Recurrent Markov Clustering and Modularity / Maki Miyake, Terry Joyce / Japan 114

Enhancing SIMPLE Semantic Relations: A Proposal / Nilda Ruimy / Italy 119

Session SP2: Speech Processing 2: recognition

Fuzzy Recall and Precision for Speech Segmentation Evaluation / Bartosz Ziółko, Suresh Manandhar, Richard C. Wilson / United Kingdom 124

A Knowledge based Approach Using Fuzzy Inference Rules for Vowel Recognition / Hrudaya Ku. Tripathy, Pradip K. Das, B.K. Tripathy / India 129

Application of slope filtering to robust spectral envelope extraction for speech recognition / Szymon Drgas, Adam Dąbrowski / Poland 134

A study on bilingual speech recognition involving a minority language / Míriam Luján, Carlos D. Martínez, Vicent Alabau / Spain 138

15:30 - 17:00 Technical Sessions

Session CO1: Communication 1

Application of MCL in a dialog agent / Darsana P. Josyula, Scott Fults, Michael L. Anderson, Shomir Wilson, DonPerlis / United States 143

Dialog Management for Decision Processes / Paul Fodor / United States 148

An approach to processing of user's commands in human-machine interaction / Milan Gnjatović, Dietmar Rösner / Germany 152

Centering-Theory-Based Text Planning of a Conjunctive Query / Paolo Dongilli / Italy 157

Session WN1: WordNet Special Track 1

Leveraging parallel corpora and existing wordnets for automatic construction of the Slovene wordnet / Darja Fišer / Slovenia 162

WordNet Translator: First Steps in Automatic Translation of an English Semantic Database into Polish / Leszek Bajkowski / Poland 167

PolNet – Polish WordNet project algorithm / Zygmunt Vetulani, Justyna Walkowska, Tomasz Obrębski, Paweł Konieczka, Przemysław Rzepecki, Jacek Marciniak / Poland 172

Session SP3: Speech Processing 3: systems

Flexible Natural Language Generation in Multiple Contexts / Caroline Cullen, Ian O'Neill, Philip Hanna / United Kingdom 177

Spoken Language Interface for Mobile Devices / João Freitas, António Calado, Maria João Barros, Miguel Sales Dias / Portugal 182

Festival Based Maiden TTS System for Tamil Language / Sreekanth Majji, Ramakrishnan A. G. / India 187

Natural Language Processing for Tamil TTS / A. G. Ramakrishnan, Lakshmish N Kaushik, Laxmi Narayana M. / India 192

17:30 - 19:00 Technical Sessions

Session CO2: Communication 2

Main principles of authoring educational hypermedia: lessons learned from the construction of the Web-resource "Russian Dialectal Phonetics" / Galina Kedrova, Valentine Kolybasova / Russia 197

Recognition of signed Polish words using visually-oriented subunits / Tomasz Kapuścinski, Marian Wysocki / Poland 202

From language to pictorial representations / Gérard Ligozat, Jakub Nowak, Didier Schmitt / France, Poland 207

Context Identification: A Relational Database Approach / Zsolt T. Kardkovács, Katalin E. Lejtovicz, Gábor Kovács / Hungary 211

Session WN2: WordNet Special Track 2

Automatically converting and enriching a computational lexicon Ontology for NLP semantic tasks / Antonio Toral, Monica Monachini, Rafael Munoz / Italy 216

Random-Walk Models of Term Semantics: An Application to Opinion-Related Properties / Andrea Esuli, Fabrizio Sebastiani / Italy 221

Putting Semantics into WordNet's "Morphosemantic" Links / Christiane Fellbaum, Anne Osherson, Peter E. Clark / United States 226

Session SP4: Speech Processing 4: algorithms

Modelling acoustic parameters of prosody in Slovak using Classification and Regression trees / Milan Rusko, Marián Trnka, Sachia Darjaa, Richard Kováč / Slovakia 231

Perception Experiments for Effective Unit replacement for Tamil TTS / A. G. Ramakrishnan, Laxmi Narayana M. / India 236

MVDR Spectral Estimation for DCT based Pitch Modification / R. Muralishankar, M. Ravi Shanker, A. G. Ramakrishnan / India 241

The relation between the structure and prosody of complex sentences inHungarian / Ilona Koutny / Poland 246

Day 2: October 6, 2007

8:30 - 9:30 Technical Sessions

Session RS1: Digital Language Resources 1

PredicateDB: A tool for assisting the creation of a lexicon-grammar of Predicative Nouns / Mohamed Mahdi Malik, Jean Royauté / France 250

Extracting Collocations in Contexts / Amalia Todirascu, Christopher Gledhill, Dan Stefanescu / France 255

Argument co-occurrence matrix as a description of verb valence / Łukasz Dębowski, Marcin Woliński / Poland 260

Session SE2: Computational Semantics 2

ARDICO: an Arabic Termino-Ontological Resource Dedicated to Children Education / Aroua Torjmen, Mohamed Ben Ahmed, Elisabeth Métais / Tunisia 265

A Semantically Oriented Readability Checker for German / Tim vor der Brück, Sven Hartrumpf / Germany 270

Extracting coordinate terms from newspaper articles and patent documents / Yoshimi Suzuki, Fumiyo Fukumoto / Japan 275

Session PR1: Parsing 1

Efficient parsing using recursive transition networks with output / Javier M. Sastre, Mikel L. Forcada / France 280

Innovative use of parameters in DCG-like logic grammars / Janusz S. Bień / Poland 285

Robustness Evaluation of Two CCG, a PCFG and a Link Grammar Parsers / Tuomo Kakkonen / Finland 290

10:00 - 10:45 Invited Talk I

The Global Wordnet Grid: anchoring languages to universal meaning / Piek Vossen / Netherlands 20

10:45 - 12:15 Technical Sessions

Session RS2: Digital Language Resources 2

Toward a Standard Lexical Resource in the Bio Domain / Valeria Quochi, Riccardo Del Gratta, Eva Sassolini, Monica Monachini, Nicoletta Calzolari / Italy 295

LMF-QL: A Graphical Tool to Query LMF databases / Mehdi Ben Abderrahmen, Bilel Gargouri, Mohamed Jmaiel / Tunisia 300

Novelty extraction from special and parallel corpora / Elżbieta Dura, Barbara Gawrońska / Sweden 305

Session SE3: Computational Semantics 3

Automatic Acquiring of Semantic Relations From Text Collection / Wiktor Dernowicz / Poland 310

The DANTE Temporal Expression Tagger / Paweł Mazur, Robert Dale / Poland, Australia 315

Capturing The Meaning of Time Expressions – A Functional Approach / Petr Nemec / Czech Republic 320

The structure of verbal sequences analyzed with unsupervised learning techniques / Catherine Recanati, Nicoleta Rogovschi, Younes Bennani / France 325

Session PR2: Parsing 2

A Hybrid System for Named Entity Metonymy Resolution / Caroline Brun, Maud Ehrmann, Guillaume Jacquet / France 330

Heuristic Disambiguation of Deverbal Nominals in Greek / Vassiliki Rentoumi, Stasinos Konstantopoulos / Greece 335

SPADE: Shallow Parsing and Disambiguation Engine / Adam Przepiórkowski, Aleksander Buczyński / Poland 340

The Polsyn Parser / Nina Suszczańska, Krzysztof Simiński / Poland 345

14:00 - 14:15 Homeland Security Workshop – Presentation of PPHS by Emil Pływaczewski and Zbigniew Rau

Polish Platform for Homeland Security / Emil Pływaczewski, Zbigniew Rau / Poland 536

14:15 - 15:00 Applications for Homeland Security

Intelligent Awareness: Event Extraction, Information Evaluation & Risk Assessment / Philippe Capet, Thomas Delavallade, Julien Jacquelinet, Claude Martineau, Takuya Nakamura, Aude Rebotier, Agnes Sándor, Stavroula Voyatzi / France 539

On the Modelling of Money Laundering Techniques as Courses of Events / Jacek Martinek, Czesław Jędrzejek / Poland 544

Towards the Homeland Security Topical Ontology - a Core Layer Specification / Jolanta Cybulka / Poland 549

15:00 - 15:45 Demos & Posters

Polish Platform for Homeland Security Poster Presentations

Cryptographic algorithms for efficient security of data flow in information system / Tomasz Bilski, Krzysztof Bucholc, Krzysztof Chmiel, Tadeusz Gajewski, Anna Grocholewska-Czuryło, Ewa Idzikowska, Janusz Stokłosa / Poland 554

Multimedia System Supporting Identification and Combating Crime and Terrorism / Andrzej Czyżewski, Piotr Szczuko / Poland 556

Emergency telephone conversations database with signal analysis, classification, and identification facilities / Adam Dąbrowski, Szymon Drgas / Poland 558

Verbal information processing and speech recognition technologies in internal security systems / Grażyna Demenko, Stefan Grocholewski / Poland 560

Intelligent Search Engine for Court Records Based on Text Processing Experience / Andrzej Dziech, Tomasz Ruść, Remigiusz Baran, Andrzej Zeja / Poland 561

PPBW – Knowledge and information management in high security level services / Gerard Frankowski, Cezary Mazurek, Norbert Meyer, Maciej Stroiński, Marcin Wolski / Poland 563

A tool supporting investigation procedures by means of automatic reasoning / Czesław Jędrzejek, Jarosław Bąk, Maciej Falkowski, Jacek Martinek, Arkadiusz Ryś / Poland 565

Information Extraction Tools for Polish Text / Wiesław Lubaszewski / Poland 567

Application of Artificial Intelligence Methods and Agent-based Technologies to Support Investigatory Activities and Proof Procedures with Emphasis on Internet Crimes / Edward Nawarecki, Grzegorz Dobrowolski / Poland 568

A mobile, network-centric system of supporting police operational work / Mikołaj Sobczak / Poland 569

Natural language based communication between human users and emergency center in critical situations. POLINT-112 / Zygmunt Vetulani / Poland 571

Demo session

DEB Platform tools for effective development of WordNets in application to PolNet / Aleš Horák, Karel Pala, Tomasz Obrębski, Przemysław Rzepecki, Paweł Konieczka, Jacek Marciniak, Adam Rambousek, Zygmunt Vetulani, Justyna Walkowska / Poland 514

Environment Supporting Construction of the Polish Wordnet / Maciej Piasecki, Paweł Koczan / Poland 519

Intelligent Awareness: Event Extraction, Information Evaluation & Risk Assessment (article demonstration) / Philippe Capet, Thomas Delavallade, Julien Jacquelinet, Claude Martineau, Takuya Nakamura, Aude Rebotier, Ágnes Sándor, Stavroula Voyatzi / France 524

Novelty extraction from special and parallel corpora (article demonstration) / Elżbieta Dura, Barbara Gawrońska / Sweden 525

Translation applications under the SisHiTra framework (article demonstration) / J. González Rubio, J. González, A. L. Lagarda, A. Giménez, J. R. Navarro, F. Casacuberta / Spain 526

The DANTE Temporal Expression Tagger (article demonstration) / Paweł Mazur, Robert Dale / Poland, Australia 527

New Polish tests for measuring speech intelligibility / Edward Ozimek, Dariusz Kutzner, Aleksander Sęk, Andrzej Wicher / Poland 528

Software "Self-Tutor for Foreign Languages" / Pavel Pankov, Polina Dolmatova / Kyrgizstan 530

A DCG account of Polish relative constructions (article demonstration) / Marek Świdziński / Poland 531

PolNet (Polish WordNet) creation: methodology and results / Paweł Konieczka, Tomasz Obrębski, Zygmunt Vetulani, Justyna Walkowska / Poland 532

POLINT 112-SMS / Zygmunt Vetulani, Krzysztof Sielski, Mariusz Tański / Poland 533

16:00 - 17:00 Open Panel Discussion

Human Language Technologies in Application to Homeland Security – Vision and Prospects

Open Panel Discussion: Human Language Technologies in Application to Homeland Security – Vision and Prospects. Chairman's Statement / Dafydd Gibbon / Germany 537

Day 3: October 7, 2007

9:00 - 10:30 Technical Sessions

Session RS3: Digital Language Resources 3

"Die Fackel" and "Der Brenner": Online Digital Editions of Literary Journals in the "AAC – Austrian Academy Corpus" / Hanno Biber, Evelyn Breiteneder, Karlheinz Mörth / Austria 350

Construction of Text Corpus of Polish Using the Internet / Sławomir Kulików / Poland 354

A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language / Fco. Mario Barcala Rodríguez, Eva Domínguez Noya, Pablo Gamallo Otero, Marisol López Martínez, Eduardo Miguel Moscoso Mato, Guillermo Rojo, María Paula Santalla del Río, Susana Sotelo Docío / Spain 359

Session IR3: Information Retrieval/Extraction 3

A Survey on Human Language Technology Terminology / Gabriella Pardelli, Manuela Sassi, Sara Goggi / Italy 364

Automatically Determining Attitude Type and Force for Sentiment Analysis / Shlomo Argamon , Kenneth Bloom, Andrea Esuli, Fabrizio Sebastiani / United States, Italy 369

Developing a Definitional Knowledge Extraction System / Rodrigo Alarcón, Gerardo Sierra, Carme Bach / Mexico 374

Session FO1: Language Formalisms 1

Applying pregroups for natural language processing / Aleksandra Kiślak-Malinowska / Poland 379

From TY$_n$ to DRT: an implementation / Patrick Blackburn, Sebastien Hinderer / France 384

An algebraic approach of the Chinese sentence with Pregroups / Daniela Bargelli / Canada 389

11:00 - 11:30 Invited Talk II

Language technology in the European programmes and policies / Kimmo Rossi / Finland 21

11:30 - 13:00 Technical Sessions

Session IR4: Information Retrieval/Extraction 4

LSA as an associative text retrieval tool / Agnieszka Figiel / Poland 394

Result Aggregation for Knowledge-Intensive Multicultural Name Matching / Keith J. Miller, Mark Arehart / United States 399

Usability of String Distance Metrics for Name Matching Tasks in Polish /Jakub Piskorski, Marcin Sydow / Poland 403

Session MR2: Computational Morphology 2

ROG – A Paradigmatic Morphological Generator for Romanian / Elena Irimia / Romania 408

Grammar-Based Speech and Word splitting / Ildar Kagirov, Anastasia Leontyeva / Russia 413

Treatment of Numerals in Text Processing / Cvetana Krstev, Duško Vitas / Serbia 418

Building a morphosyntactic lexicon and a pre-syntactic processing chain for Polish / Benoît Sagot / France 423

Session PR3: Parsing 3

Syntactic spreadsheets. In search for a human-readable representation of parse tree forests / Janusz S. Bień / Poland 428

Reducing the Number of Resulting Parsing Trees for the Czech Language Using the Beautified Chart Method / Vojtech Kovář, Aleš Horák / Czech Republic 433

Efficient Encoding of Parsed Natural Language Text / Jakub Swacha / Poland 438

Using of the Constituent Context Model to Induce a Grammar for a Free Word Order Language: Persian / Seyed Abolghassem Mirroshandel, Gholamreza Ghassem-Sani, Mohamadali Honarpisheh / Islamic Republic of Iran 443

14:15 - 16:00 Technical Sessions

Session MT1: Machine Translation 1

Data Sparsity Reduction in Statistical Machine Translation From Highly Inflected Language to English / Mirjam Sepesy Maučec, Janez Brest / Slovenia 448

Translation applications under the SisHiTra framework / Jesús González, Jorge González, A. L. Lagarda, A. Giménez, J. R. Navarro, F. Casacuberta / Spain 453

The contribution of *hapax legomena* to word alignment / Adrien Lardilleux, Yves Lepage / France 458

Using a Treebank Grammar for the Syntactical Annotation of German Lexical Phrases / Marcin Junczys-Dowmunt, Filip Graliński / Poland 463

The Use of Syntactic and Semantics Valences of the Verb for Formal Delimitation of Verb Word Phrases / Ali Abbasov, Abulfat Fatullayev / Azerbaijan 468

Session PR4: Parsing 4

On the evaluation of Polish definition extraction grammars / Adam Przepiórkowski, Łukasz Degórski, Beata Wójtowicz / Poland 473

A DCG account of Polish relative constructions / Marek Świdziński / Poland 478

Disfluency Detection and Parsing of Transcribed Speech of Estonian / Kaili Müürisep, Helen Nigol / Estonia 483

Mining Parsing Results for Lexical Corrections / Lionel Nicolas, Jacques Farré, Éric de la Clergerie / France 488

A Model for Processing Procedural Texts / Estelle Delpech, Patrick Saint Dizier / France 493

Session IR5: Information Retrieval/Extraction 5

Web based Factoid type Question Answering System For Tamil Documents – A Term Based Approach / S. Lakshmana Pandian, T.V. Geetha / India 498

Analogies of form between chunks in Japanese are massive and far from being misleading / Yves Lepage, Julien Migeot, Erwan Guillerm / France 503

Semi-Automatic Creation of a Dictionary of Nominal Compounds / Tomasz Stępien, Bartosz Podlejski / Poland 508

ACKNOWLEDGEMENTS

The Program Committee of the 3rd Language and Technology Conference acknowledge with gratitude the support and assistance of several institutions and individuals who contributed to this event:

THE CITY OF POZNAŃ

represented by Mr Ryszard Grobelny, Mayor of Poznań

CO-ORGANIZING ORGANIZATIONS

ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
ADAM MICKIEWICZ UNIVERSITY FUNDATION

CO-OPERATING ORGANIZATIONS

ELRA/ELDA
INTERNATIONAL POZNAŃ FAIR (MTP)
KOSZALIN UNIVERSITY OF TECHNOLOGY
PEARSON EDUCATION
POLISH PLATFORM FOR HOMELAND SECURITY

EXTERNAL REVIEWERS

STUDENT VOLUNTEERS

MEDIA

Preface by Zygmunt Vetulani

Half a century ago not many people had realized that a new epoch in the history of *homo sapiens* had just started. The term **Information Society Age** seems an appropriate name for this epoch.

There is little doubt that the human race began when our predecessors started to communicate with each other using language. This highly abstract means of communication was probably one of the major factors contributing to the evolutionary success of the human race within the animal world. Physically weak and imperfect, humans started to dominate the rest of the world through the creation of communication-based societies where individuals communicated initially to satisfy immediate needs, and then to create, accumulate and process knowledge for future use.

We can confidently state that the next crucial step in the history of humanity was the invention of writing. It is worth noting that writing is a *human invention*, not a phenomenon resulting from natural selection. Humans invented writing as a technique for recording speech as well as storing and communicating knowledge. Writing was the invention which made it easier to disseminate knowledge across the world. Humans continue to be born illiterate, and therefore teaching and conscious supervised learning is necessary to maintain this basic social skill. The invention of writing and the resulting writing-based civilizations stimulated the development of ever more sophisticated technologies.

Humans began to produce artefacts and technologies created in tune with the laws of nature and based on a good understanding of these laws. It must be recognized, however, that many of these inventions were incidental and there is no evidence that they were in any way necessary (i.e. they might simply not have happened). The development of technologies and the production of artefacts now seems be beyond the control of any individual, group or organization.

The emerging Information Society is a new kind of social structure where humans will be surrounded by a new generation of information-rich artefacts and technologies. These artefacts and technologies are designed to be *collaborative* with human users. This means that their role is to assist humans at least as well as human assistants could (assuming their good intentions). Artefacts made by humans "in their own image" will need to communicate with humans.

This incipient Information Society will probably be characterized by the use of **Human Language Technologies** which contribute to a world where humans need to communicate not only with each other but also with the artificially created interactive and autonomous technological environment; collaborative, but possibly also hostile.

The history of human civilization tell us that humanity evolves in ways which are barely under control and which are difficult to anticipate. This development is stimulated by the human need to rise to ever greater challenges.

We can but hope that rising to the challenges that stimulate the development of Human Language Technologies will result in a better, cleverer and happier world.

*****

In this volume the Reader will find contributions of 275 authors from 32 countries. Among these contributions are 2 invited talks, 105 technical papers, 11 invited poster presentations, 11 invited demos of system prototypes, software and resources. Technical papers were selected anonymously by experienced reviewers (authors were asked to hide their identity when submitting papers for reviewing). Papers cover a wide range of topics and were grouped in the following thematic sessions (here in alphabetical order):
Information Retrieval/Extraction (17)
Computational Morphology (8)
Computational Semantics (11)
Speech Processing (16)
Communication (8)
WordNet Special Track (8)
Digital Language Resources (10)
Parsing(16)
Language Formalisms (3)
Machine Translation (5)
Applications for Homeland Security (3 technical papers + 11 invited posters)

The Workshop on Contribution of Language Technologies to Homeland Security (including the panel discussions) are an integral part of the LTC conference.

Invited talks will be given by Kimmo Rossi (EC) and Piek Vossen (Uni. Amsterdam). The Polish Platform for Homeland Security (PPBW) will be presented by Zbigniew Rau.

I would also like to thank all authors, invited lecturers and panelists who contributed to this conference with their expertise in the domain of Language Technologies. Special thanks are due to the invited reviewers for their scrupulous reviewing, as well as to the colleagues from the Program Committee, the members of the Organizing Committee and the volunteers for their hard and effective work.

I wish you all very fruitful deliberations

Poznań, October 2007

Zygmunt Vetulani
LTC'07 Chair

Suszczanska, N., & Szmal, P., & Kulików, S. (2005). Continuous Text Translation Using Text Modeling in the Thetos System. Ali Okatan, redaktor, International Conference on Computational Intelligence. International Computational Intelligence Society (pp. 156–160).

Szmal, P., & Suszczańska, N. (2001). Selected Problems of Translation from the Polish Written Language to the Sign Language. Archiwum Informatyki Teoretycznej i Stosowanej 13(1) (pp. 37–51).

A Corpus and Lexical Resources for Multi-word Terminology Extraction in the Field of Economy in a Minority Language

Fco. Mario Barcala Rodríguez*, Eva Domínguez Noya*, Pablo Gamallo Otero†, Marisol López Martínez†, Eduardo Miguel Moscoso Mato†, Guillermo Rojo†, María Paula Santalla del Río†, Susana Sotelo Docío†

*Centro Ramón Piñeiro para a Investigación en Humanidades
Estrada Santiago-Noia, Km. 1 - A Barcia
E-15896, Santiago de Compostela, A Coruña, Spain
{fbarcala, edomin}@cirp.es
†Universidade de Santiago de Compostela
Facultade de Filoloxía, Burgo das Nacions, s/n
E-15782, Santiago de Compostela, A Coruña, Spain
{pablogam, fgmarsol, fgmato, guillermo.rojo, fempsr, fesdocio}@usc.es

Abstract

In this paper, we describe the compilation and structure of two linguistic resources, a corpus and a dictionary of terms in the field of economy, developed for Galician. In addition to this, we describe the use of these resources for the automatic extraction of multi-word terms by means of a combination of linguistic and statistical techniques. While doing this, special attention will be paid to the problems posed by minority languages such as Galician for the achievement of these tasks.

1. Introduction

The work described in this paper is part of a general project, RICOTERM-2[1], aimed at the development of a multi-lingual system for the re-formulation of queries posed by users of Internet interested in the search of information about some specialized communicative field, in our case, economy. The system is currently being developed for English, Spanish, Catalan, Basque and Galician. Its general design can be found in (Lorente, 2005). However, for the purposes of this document it is enough to point out that, to improve the result of the information retrieval task involved, the system will make use of techniques of query expansion in which a combination of methods of both term-only and full-text expansion are employed. For term-only expansion, the project plans to make use of a domain specific ontology. For full-text expansion, we try to prove the benefits of using a domain-specific corpus, structurally and linguistically annotated, in order to detect, by means of various integrated tools such as a terminology extractor, collocations or recurrent contextual phraseology of the terms included in the query or obtained after consulting the ontology.

A specific part of this general project, GARI-COTER[2], is mainly devoted to the development of the resources needed by such a system for one of the languages involved: Galician. In Sections 2. and 3. of this article we briefly describe the current stage of these resources: a galician corpus in the field of economy and a lexical collection of terms compiled from previously existing resources.

Besides this, and in line with the general approach underlying the RICOTERM project, we describe here the exploitation of the resources themselves to improve them in what can be seen as a bootstrapping process. Specifically, in Section 4. we describe the use of the corpus and the lexical resources to automatically extract multi-word terms of the field of economy in Galician.

To do this, we make use of a method that combines specific linguistic and statistical techniques in a way that can be compared to the widely-used approaches in the research community to deal with the task of terminology extraction.

Finally, as regards the development of resources, as well as for their application in terminology extraction strategies, we found that the situation of minority languages such as Galician constitutes a non-negligible difficulty. All along this document we have wanted to highlight this fact, very frequently not taken into account when designing terminology extraction techniques and, more generally, information retrieval systems.

2. The corpus

The first problem to be solved when trying to do automatic terminological extraction in a minority language like Galician is to get domain-specific documents. As already mentioned above, our research focus is in the field of economy.

The task of the development of a domain-specific corpus was divided in the following way: the development of a more general one, containing economy journal news, and of a specific one, with specialized texts of economy. This decision was taken on the basis of two reasons: the first type of corpus was much easier to obtain, but the second one was expected to be much richer from the point of view of terminology.

On the one hand, we had no problems to obtain documents for the first corpus, given that, thanks to a special agreement with the Center for Humanities Research Ramón Piñeiro[3], we could include in our corpus news collected in the CORGA (Reference Corpus of

Present-day Galician Language) corpus[4]. These news were already available both in electronic format and very carefully XML (eXtensible Markup Language)[5] structured. On the other hand, nevertheless, we had great difficulties to obtain documents for the collection of the specialized corpus, given that there were indeed very few economy specialized texts in Galician. In electronic format, only several texts, whose appropriatenes can in certain cases be arguable, could be found. We also had to encode them according to the above-mentioned XML structure.

As a result of this work, we could compile a general corpus constituted by 609 newspaper news which include 206510 words in 7892 sencences, and a specific one constituted by 14 books and 2 specialized journals which include 801702 words in 34588 sentences.

Apart from being collected, every document in the specialized corpus (each book or article from a specialized journal) has been classified by an expert according to two different taxonomies of the field. As a result of this classification, we can at least ensure that, with respect to the documents taken from the specialized journals, the corpus is reasonably representative of the field. The same, however, cannot be ensured for the book texts, for reasons that, when dealing with minority languages as Galician, are obvious: as very few texts of this type are available, only in extremely particular circumstances, one can decide not to include an available electronic text in a specialized corpus of a minority language.

2.1. Corpus encoding

As we have already pointed out, documents are structured according to the XML standard. Each document has a header which includes bibliographical details, as well as the argument or arguments of the document, this being followed by the text of the document itself, structured up to the sentence level. For example the XML structure of a single news item is:

```
... preambles of XML standard ...
<noticia> (single news item)
  <cabeceira_noticia> (header)
    <nome_publicación>
      name of the publication
    </nome_publicación>
    <editorial>publisher</editorial>
... more bibliographic information ...
    <identificador>
      single news item identifier
    </identificador>
    <autor>author</autor>
    </área_temática>
      argument
    </área_temática>
  </cabeceira_noticia>
  <contido_noticia> (content)
    <titular> (title)
      <parágrafo> (paragraph)
        <oración>sentence</oración>
      </parágrafo>
    </titular>
    <resumen> (summary)
      <parágrafo>
        <oración>sentence</oración>
... more sentences ...
      </parágrafo>
    </resumen>
    <corpo> (content)
      <parágrafo>
        <oración>sentence</oración>
... more sentences ...
      </parágrafo>
... more paragraphs ...
  </contido_noticia>
</noticia>
```

2.2. Corpus annotation

In order to use morphosyntactic information to perform automatic terminological extraction in the way we describe below, Section 4., the corpus was annotated with *POS* (Part-of-Speech) information. The tagset used is based on the one developed by the XIADA (Tagger/Lemmatizer of Present-day Galician Language) project[6]. It consists of approximately 370 tags and is designed according to the guidelines EAGLES (Expert Advisory Group on Language Engineering Standards (EAGLES), 1996).

In the first step, this tagset identifies the morphological category, and in the second one, it identifies the grammatical attributes considered relevant for the corresponding category. In the development of this tagset, the completeness of morphological descriptions was given preference over the introduction of any syntactic information in its widest sense. The latter was, in fact, reduced to the specification for only certain elements of certain categories of their functional capabilities in terms of nucleus and modifiers.

To annotate the general corpus, we have made use of the Galician default trained tagger developed by the XIADA project (Barcala et al., 2006; Graña and M. A. Alonso, 2002; Graña et al., 2002). As this tagger can manage XML information, the result was a set of documents encoded in an intermediate XML format which integrates *POS* information.

After the automatic annotation of the corpus, we performed a manual revision of its results. To do this, we have used a simple generic XML editor (XMLMind Editor[7]) adapted with Cascade Stylesheets[8]. In this stage, we took a great advantage of the tagger's intermediate XML format, which allowed us to do this task much less cumbersome.

Once the manual check of the general corpus was accomplished, the tagger was first trained again with the data of the general corpus, and then used to tag the specific one. The result of this second automatic annotation process was not manually revised.

Finally, the tagger's XML intermediate format is automatically simplified. The final format is similar to the one previosly shown, but includes *POS* information within the sentence structure:

```
...
<oración> (sentence)
  <expresión>
    full text of the sentence
  </expresión>
  <análise> (analysis)
    <análise_unidade> (analysis unit)
      <unidade>
        lexical unit to be analysed
      </unidade>
      <constituínte> (constituent)
        <forma>word</forma>
        <etiqueta>POS tag</etiqueta>
        <lema>lemma</lema>
      </constituínte>
... more constiuents if necessary ...
    </análise_unidade>
... more analysis units
  </análise>
</oración>
...
```

Let's notice especially the presence of *constituents* in the format. Although in the great majority of cases lexical units have only one constituent, this element is needed, and mainly used, to handle verb forms with enclitic pronouns, which may, in fact, have a very complex compound structure in Galician. By using constituents, however, those compounds can be efficiently accounted for, on the basis of their segmentation into a verb part and as many additional parts as enclitic pronouns attached to the verb, each one, as the verb part, analyzed separately. This phenomenon is correctly managed by the tagger, so we could get rid of it (Barcala et al., 2006) before further processing of the corpus for terminology extraction itself, see Section 4.

3. Lexical resources

One of the needs, and a goal on itself, for terminology extraction as described below is the compilation of a database of terms in the field of economy[9]. Two techniques were used to obtain this database of terms: the automatic extraction from the domain corpus, as described in Section 4., and the manual compilation of terms from a wide range of sources which include electronic glossaries and dictionaries. In this section we are going to describe the lexical resources developed using the second technique, as well as the sources from which they could be obtained. Although we will not go into detail with respect to this for each of the sources examined, we want to remark here that Galician is a language which has recently undergone –it still undergoes– a process of normalisation, which means that in the collection of terms from different sources we had to handle the different forms in which same words can be transcribed.

The sources[10] considered were quite heterogeneous, as can be deduced from Table 1: two dictionaries (*Eiras* and *Formoso*, one of them trilingual), two electronic glossaries freely available from the web, and the section of economy of the terminological database built by the Linguistic Normalization Service of the University of Santiago de Compostela (a very large terminological database which tries to cover the terminology of several scientific fields).

The last one is the most reliable and accurate, since it was carefully collected from 26 different sources and includes very rich and varied information, such as the equivalence of terms in other languages, information about semantic relations such as synonymy or hiperonymy, and definitions. Dictionaries also must be considered good and reliable sources: they include definitions and translations, as well as a not too exhaustive information about synonyms and antonyms.

Not only with respect to quality (volume of information for each term), but also to quantity (number of terms supplied), these three resources are more important than the others: in addition, in effect, to the fact that more terms are indeed gathered in them, the percentage of unique terms in these resources is also higher (see Table 1).

The internet glossaries, then, must be considered minor resources, both in number of terms and with respect to the information included for each term.

Each source was encoded using XML and a common structure defined by a DTD, the one that is used for the Gari-Coter term database.

Source	Terms	Type	Unique lemmas
Eiras	3232	dictionary	2291 (70,88%)
SNL	2894	terminological DB	1746 (60,33%)
Formoso	1346	multiling. dictionary	839 (62,33%)
Panlatino	273	multiling. glossary	20 (7,32%)
galego.org	153	glossary	46 (30%)
Gari-Coterm	6046		

Table 1: Sources of the lexical resources

3.1. Dictionary encoding

The Gari-Coter list of terms was encoded according to the XML standard as a result of merging the different sources described above. Each term is enclosed within the tag `<term>`, and includes exhaustive information about lemma, part-of-speech and definition, and in most cases it includes also the equivalence in other languages, as well as some semantic information about synonyms or hyperonyms.

In future work, we plan to convert this XML-based resource into a relational database with a web interface. This will quite easily allow us to generate subsets of the list in accordance with specific restrictions, something which we expect that will be very useful to perform sub-domain terminology extraction.

4. Terminology extraction

Terms are seen here as useful indexing units in IR applications. So, they must be good from a semantic point of view, that is, they must capture as much as possible the meaning of a domain-specific corpus. Moreover, it has been recognized that single words are not always useful for the terminological representation of domain-specific texts.

terms	similar multi-words	Dice
forza de traballo (*labour force*)	man de obra (*labour force*)	0.15
	medio de produción (*production means*)	0.08
gasto público (*public spending*)	diñeiro en circulación (*money supply*)	0.12
	déficit comercial (*trade deficit*)	0.10
tecido industrial (*business network*)	Baixa Idade Media* (*Late Middle Ages**)	0.12
	explotación agraria (*land cultivation*)	0.11
taxa de crecemento (*growth rate*)	ritmo de crecemento (*rhythms of growth*)	0.11
	maior crecemento* (*bigger growth*)	0.11
	taxa de paro (*rate of unemployment*)	0.11
enerxía renovable (*renewable energy*)	enerxía solar (*solar energy*)	0.13

Table 2: 5 terms and their similar multi-words

For this purpose, multi-word expressions seem to be more appropriate. In this section, we describe an approach to automatically extract multi-word terms.

Our strategy consists of two steps. First, a list of seed terms is semi-automatically selected from the annotated corpus making use of available glossaries and resources. Then, we use that list as a set of positive examples to identify multi-word units with similar contextual distribution in the corpus. Similar multi-word units will be considered as new term candidates.

4.1. Term samples

The first objective is to build a starting list of positive term examples. For this purpose, we follow a very basic strategy. First, some morpho-syntactic patterns are used as endogenous constraints to select a generic list of multi-word units from the annotated corpus. Four nominal patterns are used:

noun – adj
adj – noun
noun – noun
noun – prep – noun

Then, a statistical filter is applied to identify those multi-word units in the generic list with a high degree of cohesion. The glue measure employed in the filtering process is *SCP*, defined in (Silva et al., 1999). Finally, the filtered list is revised by hand using as gold standard the available terminological resources described above, in Section 3.

4.2. Corpus-based similarity

The second objective is to learn new candidate terms by making use of both the annotated corpus and the list of positive examples selected in the previous step. For this purpose, we follow a method based on exogenous (i.e. contextual) information (Basili et al., 2001; Maynard and Ananiadou, 1999; Cimiano and Völker, 2005). The assumption the method is based on is the following: a multi-word unit that appears in the same local contexts as a given multi-word term should also be considered as a term. So, we implemented an algorithm calculating the similarity between terms and multi-word units on the basis of contextual features extracted from the corpus. The multi-word units compared to the list of term samples are all those selected using the 4 nominal patterns described above.

Lexico-syntactic contexts of multi-word units are extracted from the POS tagged corpus using pattern matching techniques (articles and pronouns are previously removed). For instance, given the expressions:

> "loss of *labour force*"
> "*labour force* of a country"

containing the compound noun "labour force", two contexts are extracted:

> < loss of [NOUN] >
> < [NOUN] of country >

where NOUN stands for the head category of the multi-word unit. To extract lexico-syntactic contexts, we follow the notion of *co-requirements* introduced in (Gamallo et al., 2005). According to this notion, two words (*head* and *dependent* words) related by a syntactic dependency are mutually constrained. They impose linguistic requirements on each other. A prefixed "Predicate-Argument" organization does not exist. The head imposes syntactic and semantic constraints on the words that fill the dependent position, as well as the dependent word imposes specific restrictions on the kind of head it depends on. Experimental tests showed that co-requirements permits a finer-grained characterization of "meaningful" syntactic contexts.

Once lexico-syntactic contexts have been extracted, they are associated to their co-occurring multi-word units in order to build a collocation database. Each multi-word unit (term or not) is defined as a vector where each lexico-syntactic context corresponds to a feature. Before starting to compute similarity between vectors, sparse contexts are filtered out. A context is sparse if it has high word dispersion. Dispersion is defined as the number of different multi-word units occurring with a lexico-syntactic context divided by the total number of different multi-word units in the training corpus. So, the vector space is only constituted by those lexico-syntactic contexts whose multi-word unit dispersion is lower than an empirically set threshold.

Each multi-word term of the starting list is compared to the rest of multi-word units in the corpus using Dice coefficient as similarity measure. Similarity between a multi-word term, t, and a multi-word unit, mw, which is not in the starting list of term samples, is computed as follows:

	Accuracy
Test list 1	.74
Test list 2	.70
Test size	160

Table 3: Evaluation of candidate terms

$$Dice(t, mw) = \frac{2 * \sum_i min(f(t, c_i), f(mw, c_i))}{f(t) + f(mw)}$$

where $f(t, c_i)$ represents the number of times t co-occurs with the context c_i. Likewise, $f(mw, c_i)$ represents the number of times mw co-occurs with the context c_i. For each term, we select the k most similar multi-word units (where $k = 5$) with a Dice score $>= 0.05$. Table 2 shows the most similar multi-word units associated to five terms of the starting list. Similar multi-word units are considered to be candidate terms. Those extracted multi-word units with asterisk are odd terms.

4.3. Experiments and evaluation

Experiments have been carried out over the annotated corpus described in Section 2. The starting glossary of terms contains 150 entries, while the final list of candidate terms we have extracted contains 740 multi-word units. To evaluate the accuracy of the system, we randomly selected 2 test lists of 160 multi-word units from the final list. A human evaluator decided if they are correct or incorrect terms. Table 3 depicts the accuracy scores, where *accuracy* is defined as the number of correct terms divided by the total number of test words.

The main problem of our strategy is that co-occurrences of multi-word units are still more sparse than those of simple words. Indeed, corpus-based algorithms to extract any information on multi-word units (for instance, information on *termhood*) require larger domain-specific corpus. This is a challenge for minority languages.

Notes

[1] *Terminological and Discursive Control for Information Retrieval in Specialized Communicative Fields, by means of Specific Linguistic Resources and a Re-Elaborator of Queries*, financed between 2004 and 2007 by the Ministry of Science and Technology of the Spanish Government.

[2] *Development and Multilingual Integration of Linguistic Resources in Galician for Information Retrieval by means of Strategies of Terminological and Discursive Control in Specialized Communicative Fields*, financed between 2004 and 2007 by the Ministry of Science and Technology of the Spanish Government.

[3] http://www.cirp.es. [Consulted: june, 2, 2007].

[4] http://corpus.cirp.es/corgaxml. [Consulted: june, 2, 2007].

[5] http://www.w3.org/XML/

[6] http://corpus.cirp.es/xiada, 0.1.0 version. [Consulted: june, 2, 2007].

[7] http://www.xmlmind.com

[8] http://www.w3.org/Style/CSS/

[9] In the course of the Gari-Coter project, this database is going to be integrated in an ontology of the field of economy.

[10] **Eiras**: Eiras Rey, A.: *Dicionario de economía*, to be published.
Formoso: Formoso Gosende, V. (coord.) (1997): *Diccionario de termos económicos e empresariais galego-castelán-inglés*. Santiago de Compostela: Confederación de Empresarios de Galicia.
Panlatin Electronic Commerce Glossary: http://fon.gs/panlatino
Glossary about commerce from galego.org: http://galego.org/vocabularios/ccomercial.html
SNL: http://www.usc.es/en/servizos/portadas/snl.jsp

5. References

Barcala, F. M., M. A. Molinero, and E. Domínguez, 2006. XML rules for enclitic segmentation. In A. Quesada-Arencibia, J. C. Rodríguez-Rodríguez, R. Moreno-Díaz (jr.), and R. Moreno-Díaz (eds.), *Computer Aided Systems Theory (Extendeds Abstracts)*. Las Palmas de Gran Canaria.

Basili, R., M. Pazienza, and F. M. Zanzotto, 2001. Modelling syntactic context in automatic term extraction. In *3th Conference on Recent Advances in Natural Language Processing, RANLP2001*.

Cimiano, P. and J. Völker, 2005. Text2Onto - A framework for ontology learning and data-driven change discovery. In *10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005)*.

Expert Advisory Group on Language Engineering Standards (EAGLES), 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to european languages. EAGLES Document EAG-CLWG-MORPHSYN/R. Technical report.

Gamallo, P., A. Agustini, and G. Lopes, 2005. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146.

Graña, J., F. M. Barcala, and J. Vilares, 2002. Formal methods of tokenization for part-of-speech tagging. In *Computational Linguistics and Intelligent Text Processing*. LNAI, Springer-Verlag, pages 240–249.

Graña, J. and M. Vilares M. A. Alonso, 2002. A common solution for tokenization and part-of-speech tagging: One-pass Viterbi algorithm vs. iterative approaches. In *Text, Speech and Dialogue*. LNAI, Springer-Verlag, pages 3–10.

Lorente, M., 2005. Ontología sobre economía y recuperación de información [on line]. *Hipertext.net*, (3). http://www.hipertext.net. [Consulted: january, 30, 2007].

Maynard, D. and S. Ananiadou, 1999. Identifying contextual information for term extraction. In *5th International Congress on Terminology and Knowledge Engineering (TKE'90)*.

Silva, J. F., G. Dias, S. Guilloré, and G. P. Lopes, 1999. Using LocalMaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. In *Progress in Artificial Intelligence*. LNAI, Springer-Verlag, pages 113–132.

AUTHOR INDEX

A
Abbasov Ali 468
Abderrahmen Mehdi Ben 300
Ahmed Mohamed Ben 265
Alabau Vicent 138
Alarcón Rodrigo 374
Amaral Rui 39
Anderson Michael L. 143
Arehart Mark 399
Argamon Shlomo 369
Athanaselis Theologos 94

B
Bach Carme 374
Bajkowski Leszek 167
Bakamidis Stelios 94
Baran Remigiusz 561
Bargelli Daniela 389
Barros Maria João 182
Bąk Jarosław 565
Bennani Younes 325
Biber Hanno 350
Bień Janusz S. 285, 428
Bilski Tomasz 554
Blackburn Patrick 384
Bloom Kenneth 369
Bosch Sonja E. 54
Breiteneder Evelyn 350
Brest Janez 448
Broda Bartosz 104
Brŭck Tim 270
Brun Caroline 330
Bucholc Krzysztof 554
Buczyński Aleksander 340

C
Calado António 182
Calzolari Nicoletta 295
Capet Philippe 524, 539
Casacuberta F. 453, 526
Caseiro Diamantino 79
Chauché Jacques 29
Chmiel Krzysztof 544
Clark Peter E. 226
Clergerie Éric 488
Cullen Caroline 177
Cybulka Jolanta 549
Czyżewski Andrzej 556

D
Daćko Dariusz 109
Dale Robert 315, 527
Darjaa Sachia 231
Das Pradip K. 129
Dąbrowski Adam 134, 558
Degórski Łukasz 473
Del Gratta Riccardo 295
Delavallade Thomas 524, 539
Delpech Estelle 493
Demenko Grażyna 560
Dernowicz Wiktor 310
Dębowski Łukasz 260
Dias Miguel Sales 182
Dizier Patrick Saint 493
Dobrowolski Grzegorz 568
Docío Susana Sotelo 359
Dolmatova Polina 530
Dologlou Ioannis 94
Dongilli Paolo 157
Draxler Christoph 69
Drgas Szymon 134, 558
Dura Elżbieta 305, 525
Dziech Andrzej 561

E
Ehrmann Maud 330
Esuli Andrea 221, 369

F
Fagni Tiziano 24
Falkowski Maciej 565
Farré Jacques 488
Fatullayev Abulfat 468
Fellbaum Christiane 226
Figiel Agnieszka 394
Fišer Darja 162
Forcada Mikel L. 280
Frankowski Gerard 563
Freitas João 182
Fukumoto Fumiyo 34, 275
Fults Scott 143

G
Gajewski Tadeusz 554
Gałka Jakub 74
Gargouri Bilel 300
Gawrońska Barbara 305, 525
Geetha T.V. 498
Ghassem-Sani Gholamreza 443
Gibbon Dafydd 537
Giménez A. 453, 526
Gledhill Christopher 255
Gnjatović Milan 152
Goggi Sara 364
González J. 526
González Jesús 453
González Jorge 453
González Rubio J. 526
Graliński Filip 463
Gregoriadou Maria 44
Grocholewska-Czuryło Anna 554
Grocholewski Stefan 560
Gubrynowicz Ryszard 64
Guillerm Erwan 503

H
Hanna Philip 177
Hartrumpf Sven 270
Hinderer Sebastien 384
Honarpisheh Mohamadali 443
Horák Aleš 433, 514

I
Idzikowska Ewa 554
Irimia Elena 408

J
Jacquelinet Julien 539, 524
Jacquet Guillaume 330
Jäensch Klaus 69
Jędrzejek Czesław 544, 565
Jmaiel Mohamed 300
Josyula Darsana P. 143
Joyce Terry 114
Józefowska Joanna 109
Junczys-Dowmunt Marcin 463

K
Kagirov Ildar 413
Kakkonen Tuomo 290
Kapuścinski Tomasz 202
Kardkovács Zsolt T. 211
Kaushik Lakshmish N 192
Kedrova Galina 197
Kiliçaslan Yilmaz 89
Kimmo Rossi 21
Kiślak-Malinowska Aleksandra 379
Koczan Paweł 519
Kolybasova Valentine 197
Konieczka Paweł 172, 514, 532
Konstantopoulos Stasinos 335
Koutny Ilona 246
Kováč Richard 231
Kovács Gábor 211
Kováø Vojtech 433
Kroustalli Stavroula 44
Krstev Cvetana 418
Kulików Sławomir 354
Kupść Anna 99
Kutzner Dariusz 528

L
Lagarda A. L. 453, 526
Lardilleux Adrien 458
Lejtovicz Katalin E. 211
Lembessi Penelope 44
Leontyeva Anastasia 413
Lepage Yves 458, 503
Ligozat Gérard 207
Lubaszewski Wiesław 567
Luján Miriam 138

Ł
Ławrynowicz Agnieszka 109

M
Magnini Bernardo 84
Majji Sreekanth 187
Malik Mohamed Mahdi 250
Mamouras Konstantinos 94
Manandhar Suresh 74, 124
Marasek Krzysztof 64
Marciniak Jacek 172, 514
Marciniak Małgorzata 64, 99
Martineau Claude 524, 539
Martinek Jacek 544, 565
Martínez Carlos D. 138
Martínez Marisol López 359
Mascarenhas Isabel 79
Mato Eduardo Miguel Moscoso 359
Maučec Mirjam Sepesy 448
Mazur Paweł 315, 527
Mazurek Cezary 563
Métais Elisabeth 265
Meyer Norbert 563
Migeot Julien 503
Miller Keith J. 399
Mirroshandel Seyed Abolghassem 443
Miyake Maki 114
Monachini Monica 216, 295
Mörth Karlheinz 350
Munoz Rafael 216
Muralishankar R. 241
Mŭŭrisep Kaili 483
Mykowiecka Agnieszka 64, 99

N
Nakamura Takuya 524, 539
Narayana Laxmi M. 192, 236
Navarro J. R. 453, 526
Nawarecki Edward 568
Nemec Petr 320
Nicolas Lionel 488
Nigol Helen 483
Nowak Jakub 207
Noya Eva Domínguez 359

O
O'Neill Ian 177
Obrębski Tomasz 172, 514, 532
Osherson Anne 226
Otero Pablo Gamallo 359
Ozimek Edward 528

P
Pala Karel 514
Pandian S. Lakshmana 498
Pankov Pavel 530
Pardelli Gabriella 364
Paul Fodor 148
Perlis Don 143
Petrovski Aleksandar 49
Piasecki Maciej 104, 519
Piskorski Jakub 99, 403
Pływaczewski Emil 536
Podlejski Bartosz 508
Popescu Octavian 84
Pretorius Laurette 54
Prince Violaine 29
Przepiórkowski Adam 340, 473

Q
Quochi Valeria 295

R
Rabiega-Wiśniewska Joanna 64
Ramakrishnan A. G. 187, 192, 236, 241
Rambousek Adam 514
Rau Zbigniew 536
Rebotier Aude 524, 539
Recanati Catherine 325
Rentoumi Vassiliki 335
Río María Paula Santalla 359
Rodríguez Fco. Mario Barcala 359
Rogovschi Nicoleta 325
Rojo Guillermo 359
Rösner Dietmar 152

Royauté Jean 250
Ruimy Nilda 119
Rusko Milan 231
Ruść Tomasz 561
Ryś Arkadiusz 565
Rzepecki Przemysław 172, 514

S
Sagot Benoît 423
Sándor Ágnes 524, 539
Sassi Manuela 364
Sassolini Eva 295
Sastre Javier M. 280
Schmitt Didier 207
Sebastiani Fabrizio 24, 221, 369
Serralheiro António 79
Sęk Aleksander 528
Shanker M. Ravi 241
Sielski Krzysztof 533
Sierra Gerardo 374
Simiński Krzysztof 345
Sobczak Mikołaj 569
Stefanescu Dan 255
Stępień Tomasz 508
Stokłosa Janusz 554
Stroiński Maciej 563
Suszczańska Nina 345
Suzuki Yoshimi 275
Suzuki Yoshimi 34
Swacha Jakub 438
Sydow Marcin 403
Szczuko Piotr 556
Szpakowicz Stanisław 104

Ś
Świdziński Marek 478, 531

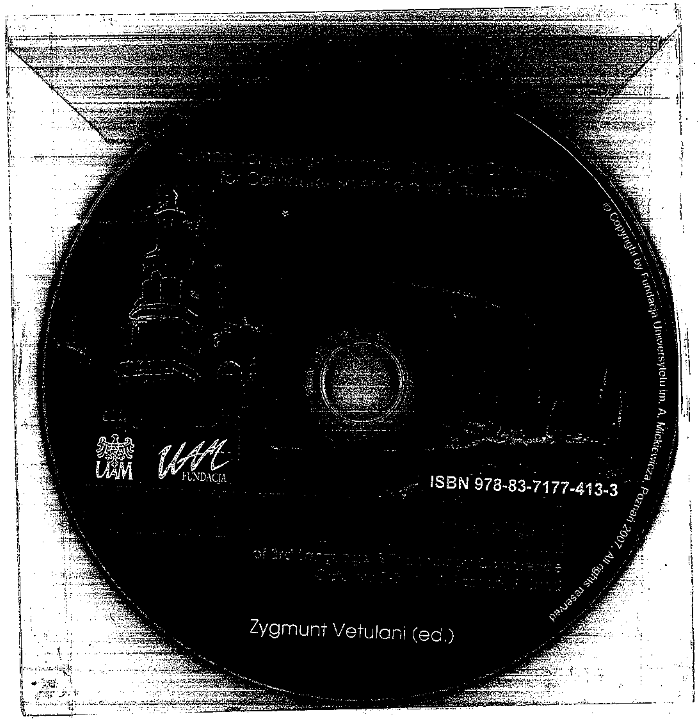
T
Tański Mariusz 533
Todirascu Amalia 255
Toral Antonio 216
Torjmen Aroua 265
Trancoso Isabel 39, 79
Tripathy B.K. 129
Tripathy Hrudaya Ku. 129
Trnka Marián 231

V
Vetulani Zygmunt 172, 514, 532, 533, 571
Viana Céu 79
Vitas Duško 418
Vossen Piek 20
Voyatzi Stavroula 524, 539

W
Walkowska Justyna 172, 514, 532
Wicher Andrzej 528
Wilson Richard C. 74, 124
Wilson Shomir 143
Woliński Marcin 59, 260
Wolski Marcin 563
Wójtowicz Beata 473
Wysocki Marian 202

Y
Yildirim Savaş 89
Yildiz Tugba 89

Z
Zdravkova Katerina 49
Zeja Andrzej 561
Ziółko Bartosz 74, 124
Ziółko Mariusz 74


UAM
FUNDACJA
ISBN 978-83-7177-413-3
Zygmunt Vetulani (ed.)
© Copyright by Fundacja Uniwersytetu im. A. Mickiewicza, Poznań 2007. All rights reserved