ACL 2010

LAW IV

Fourth Linguistic Annotation Workshop

Proceedings of the Workshop

15-16 July 2010
Uppsala University
Uppsala, Sweden

Production and Manufacturing by
Taberg Media Group AB
Box 94, 562 02 Taberg
Sweden

©2010 The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-72-5 / 1-932432-72-8

Introduction

The Linguistic Annotation Workshop (The LAW) provides a forum to facilitate the exchange and propagation of research results concerned with the annotation, manipulation, and exploitation of corpora; work towards the harmonization and interoperability from the perspective of the increasingly large number of tools and frameworks for annotated language resources; and work towards a consensus on all issues crucial to the advancement of the field of corpus annotation. Although this year's LAW is officially the fourth edition, LAW itself is the convergence of several previous workshops—including NLPXML, FLAC, LINC, and Frontiers in Corpus Annotation—dating back to the first NLPXML in 2001. This series of workshops attests to the rapid developments in the creation and use of annotated data in both language technology and empirical approaches to linguistic studies over the past 10 years.

The response to this year's Call for Papers was enthusiastic: 60 submissions were received. After careful review, the program committee accepted 20 long papers and 24 posters. Selection of the papers was not an easy task, as the papers cover the full range of linguistic facts and their corresponding annotation frameworks, from predicate-argument to discourse structure, speech to social networks, and learner corpus to CVs. The papers also deal with a range of annotation levels, from the macro perspective on infrastructure for international collaboration and interoperability, to the micro perspective on tools to deal with inter-annotator inconsistencies. It is this richness of the topics that attest to the growing maturity of field. We would like to thank SIGANN for its continuing endorsement of the LAW workshops, as well as the support and comments from the ACL-IJCNLP 2009 workshop committee chairs: Pushpak Bhattacharyia and David Weir. We would also like to thank the ACL publication chairs Jing-Shin Chang and Philipp Koehn for their help in producing the LAW IV proceedings. Most of all, we would like to thank all our program committee members and reviewers for their dedication and helpful review comments. Without them, LAW IV could not be implemented successfully.

Nianwen Xue and Massimo Poesio, Program Committee Co-chairs

Nancy Ide and Adam Meyers, Organizers

Organizers:

Nancy Ide (Vassar College)
Adam Meyers (New York University)
Chu-Ren Huang (The Hong Kong Polytechnic University)
Antonio Pareja-Lora (SIC, UCM / OEG, UPM)
Sameer Pradhan (BBN Technologies)
Manfred Stede (Universität Potsdam)
Nianwen Xue (Brandeis University)

Program Committee:

Program Co-chairs:

Nianwen Xue (Brandeis University)
Massimo Poesio (University of Trento)

Members:

Nicoletta Calzolari (ILC/CNR)
Steve Cassidy (Macquarie University)
Tomaz Erjavec (Josef Stefan Institute)
Katrin Erk (University of Texas at Austin)
Alex Chengyu Fang (City University of Hong Kong)
Christiane Fellbaum (Princeton University)
Chu-Ren Huang (The Hong Kong Polytechnic University)
Nancy Ide (Vassar College)
Richard Johansson (University of Trento)
Aravind Joshi (University of Pennsylvania)
Sandra Kübler (Indiana University)
Seth Kulick (University of Pennsylvania)
Adam Meyers (New York University)
Olga Babko-Malaya (BAE Systems)
Eleni Miltsakaki (University of Pennsylvania)
Antonio Pareja-Lora (SIC, UCM / OEG, UPM)
Martha Palmer (University of Colorado)
Rebecca J. Passonneau (Columbia University)
Marta Recasens Potau (Universitat de Barcelona)
Sameer Pradhan (BBN Technologies)
Rashmi Prasad (University of Pennsylvania)
Arndt Riester (IMS Universität Stuttgart)
James Pustejovsky (Brandeis University)
Kepa Rodriguez (University of Trento)
Anna Rumshisky (Brandeis University)
Manfred Stede (Universität Potsdam)

Marc Verhagen (Brandeis University)
Theresa Wilson (University of Edinburgh)
Andreas Witt (Universität Tübingen)
Anja Nedoluzhko (Charles University)

Table of Contents

EmotiBlog: A Finer-Grained and More Precise Learning of Subjectivity Expression Models
Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco and Andrés Montoyo 1

Error-Tagged Learner Corpus of Czech
Jirka Hana, Alexandr Rosen, Svatava Škodová and Barbora Štindlová 11

Annotation Scheme for Social Network Extraction from Text
Apoorv Agarwal, Owen C. Rambow and Rebecca J. Passonneau 20

Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs
Bea Alex, Claire Grover, Rongzhou Shen and Mijail Kabadjov 29

Consistency Checking for Treebank Alignment
Markus Dickinson and Yvonne Samuelsson 38

Anveshan: A Framework for Analysis of Multiple Annotators' Labeling Behavior
Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi and Nancy Ide 47

Influence of Pre-Annotation on POS-Tagged Corpus Development
Karën Fort and Benoît Sagot 56

To Annotate More Accurately or to Annotate More
Dmitriy Dligach, Rodney Nielsen and Martha Palmer 64

Annotating Underquantification
Aurelie Herbelot and Ann Copestake 73

PropBank Annotation of Multilingual Light Verb Constructions
Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue and Martha Palmer 82

Retrieving Correct Semantic Boundaries in Dependency Structure
Jinho Choi and Martha Palmer 91

Complex Predicates Annotation in a Corpus of Portuguese
Iris Hendrickx, Amália Mendes, Sílvia Pereira, Anabela Gonçalves and Inês Duarte 100

Using an Online Tool for the Documentation of Edo Language
Ota Ogie 109

Cross-Lingual Validity of PropBank in the Manual Annotation of French
Lonneke van der Plas, Tanja Samardzic and Paola Merlo 113

Characteristics of High Agreement Affect Annotation in Text
Cecilia Ovesdotter Alm 118

The Deep Re-Annotation in a Chinese Scientific Treebank
Kun Yu, Xiangli Wang, Yusuke Miyao, Takuya Matsuzaki and Junichi Tsujii 123

The Unified Annotation of Syntax and Discourse in the Copenhagen Dependency Treebanks
Matthias Buch-Kromann and Iørn Korzen 127

Identifying Sources of Inter-Annotator Variation: Evaluating Two Models of Argument Analysis
Barbara White 132

Dependency-Based PropBanking of Clinical Finnish
Katri Haverinen, Filip Ginter, Timo Viljanen, Veronika Laippala and Tapio Salakoski 137

Building the Syntactic Reference Corpus of Medieval French Using NotaBene RDF Annotation Tool
Nicolas Mazziotta 142

Chunking German: An Unsolved Problem
Sandra Kübler, Kathrin Beck, Erhard Hinrichs and Heike Telljohann 147

Proposal for MWE Annotation in Running Text
Iris Hendrickx, Amália Mendes and Sandra Antunes 152

A Feature Type Classification for Therapeutic Purposes: A Preliminary Evaluation with Non-Expert Speakers
Gianluca E. Lebani and Emanuele Pianta 157

Annotating Korean Demonstratives
Sun-Hee Lee and Jae-young Song 162

Creating and Exploiting a Resource of Parallel Parses
Christian Chiarcos, Kerstin Eckart and Julia Ritz 166

From Descriptive Annotation to Grammar Specification
Lars Hellan 172

An Annotation Schema for Preposition Senses in German
Antje Müller, Olaf Hülscher, Claudia Roch, Katja Keßelmeier, Tobias Stadtfeld, Jan Strunk and Tibor Kiss 177

OTTO: A Transcription and Management Tool for Historical Texts
Stefanie Dipper, Lara Kresse, Martin Schnurrenberger and Seong-Eun Cho 182

Multimodal Annotation of Conversational Data
Philippe Blache, Roxane Bertrand, Emmanuel Bruno, Brigitte Bigi, Robert Espesser, Gaelle Ferre, Mathilde Guardiola, Daniel Hirst, Ning Tan, Edlira Cela, Jean-Claude Martin, Stéphane Rauzy, Mary-Annick Morel, Elisabeth Murisasco and Irina Nesterenko 186

Combining Parallel Treebanks and Geo-Tagging
Martin Volk, Anne Goehring and Torsten Marek 192

Challenges of Cheap Resource Creation
Jirka Hana and Anna Feldman 197

Discourse Relation Configurations in Turkish and an Annotation Environment
Berfin Aktaş, Cem Bozşahin and Deniz Zeyrek 202

An Overview of the CRAFT Concept Annotation Guidelines
Michael Bada, Miriam Eckert, Martha Palmer and Lawrence Hunter 207

Syntactic Tree Queries in Prolog
Gerlof Bouma 212

An Integrated Tool for Annotating Historical Corpora
Pablo Picasso Feliciano de Faria, Fabio Natanael Kepler and Maria Clara Paixão de Sousa 217

The Revised Arabic PropBank
Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan and Martha Palmer 222

PackPlay: Mining Semantic Data in Collaborative Games
Nathan Green, Paul Breimyer, Vinay Kumar and Nagiza Samatova 227

A Proposal for a Configurable Silver Standard
Udo Hahn, Katrin Tomanek, Elena Beisswanger and Erik Faessler 235

A Hybrid Model for Annotating Named Entity Training Corpora
Robert Voyer, Valerie Nygaard, Will Fitzgerald and Hannah Copperman 243

Anatomy of Annotation Schemes: Mapping to GrAF
Nancy Ide and Harry Bunt 247

Annotating Participant Reference in English Spoken Conversation
John Niekrasz and Johanna Moore 256

Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation
Anne Lacheret-Dujour, Nicolas Obin and Mathieu Avanzi 265

Depends on What the French Say - Spoken Corpus Annotation with and beyond Syntactic Functions
José Deulofeu, Lucie Duffort, Kim Gerdes, Sylvain Kahane and Paola Pietrandrea 274

The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations
Deniz Zeyrek, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, and Ümit Deniz Turan 282

Conference Program

Thursday, July 15, 2010

08:40–08:50 Opening remarks

08:50–10:30 Session I (Chair: Nianwen Xue)

08:50–09:15 *EmotiBlog: A Finer-Grained and More Precise Learning of Subjectivity Expression Models*
Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco and Andrés Montoyo

09:15–09:40 *Error-Tagged Learner Corpus of Czech*
Jirka Hana, Alexandr Rosen, Svatava Škodová and Barbora Štindlová

09:40–10:05 *Annotation Scheme for Social Network Extraction from Text*
Apoorv Agarwal, Owen C. Rambow and Rebecca J. Passonneau

10:05–10:30 *Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs*
Bea Alex, Claire Grover, Rongzhou Shen and Mijail Kabadjov

10:30–11:00 Break

11:00–12:40 Session II (Chair: Martha Palmer)

11:00–11:25 *Consistency Checking for Treebank Alignment*
Markus Dickinson and Yvonne Samuelsson

11:25–11:50 *Anveshan: A Framework for Analysis of Multiple Annotators' Labeling Behavior*
Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi and Nancy Ide

11:50–12:15 *Influence of Pre-Annotation on POS-Tagged Corpus Development*
Karën Fort and Benoît Sagot

12:15–12:40 *To Annotate More Accurately or to Annotate More*
Dmitriy Dligach, Rodney Nielsen and Martha Palmer

12:40–13:50 Lunch

Thursday, July 15, 2010 (continued)

13:50–15:30 Session III (Chair: Manfred Stede)

13:50–14:15 *Annotating Underquantification*
Aurelie Herbelot and Ann Copestake

14:15–14:40 *PropBank Annotation of Multilingual Light Verb Constructions*
Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue and Martha Palmer

14:40–15:05 *Retrieving Correct Semantic Boundaries in Dependency Structure*
Jinho Choi and Martha Palmer

15:05–15:30 *Complex Predicates Annotation in a Corpus of Portuguese*
Iris Hendrickx, Amália Mendes, Sílvia Pereira, Anabela Gonçalves and Inês Duarte

15:30–16:00 Break

16:00–17:30 Poster session (Chair: Nianwen Xue)

Using an Online Tool for the Documentation of Edo Language
Ota Ogie

Cross-Lingual Validity of PropBank in the Manual Annotation of French
Lonneke van der Plas, Tanja Samardzic and Paola Merlo

Characteristics of High Agreement Affect Annotation in Text
Cecilia Ovesdotter Alm

The Deep Re-Annotation in a Chinese Scientific Treebank
Kun Yu, Xiangli Wang, Yusuke Miyao, Takuya Matsuzaki and Junichi Tsujii

The Unified Annotation of Syntax and Discourse in the Copenhagen Dependency Treebanks
Matthias Buch-Kromann and Iørn Korzen

Identifying Sources of Inter-Annotator Variation: Evaluating Two Models of Argument Analysis
Barbara White

Thursday, July 15, 2010 (continued)

Dependency-Based PropBanking of Clinical Finnish
Katri Haverinen, Filip Ginter, Timo Viljanen, Veronika Laippala and Tapio Salakoski

Building the Syntactic Reference Corpus of Medieval French Using NotaBene RDF Annotation Tool
Nicolas Mazziotta

Chunking German: An Unsolved Problem
Sandra Kübler, Kathrin Beck, Erhard Hinrichs and Heike Telljohann

Proposal for MWE Annotation in Running Text
Iris Hendrickx, Amália Mendes and Sandra Antunes

A Feature Type Classification for Therapeutic Purposes: A Preliminary Evaluation with Non-Expert Speakers
Gianluca E. Lebani and Emanuele Pianta

Annotating Korean Demonstratives
Sun-Hee Lee and Jae-young Song

Creating and Exploiting a Resource of Parallel Parses
Christian Chiarcos, Kerstin Eckart and Julia Ritz

From Descriptive Annotation to Grammar Specification
Lars Hellan

An Annotation Schema for Preposition Senses in German
Antje Müller, Olaf Hülscher, Claudia Roch, Katja Keßelmeier, Tobias Stadtfeld, Jan Strunk and Tibor Kiss

OTTO: A Transcription and Management Tool for Historical Texts
Stefanie Dipper, Lara Kresse, Martin Schnurrenberger and Seong-Eun Cho

Multimodal Annotation of Conversational Data
Philippe Blache, Roxane Bertrand, Emmanuel Bruno, Brigitte Bigi, Robert Espesser, Gaelle Ferre, Mathilde Guardiola, Daniel Hirst, Ning Tan, Edlira Cela, Jean-Claude Martin, Stéphane Rauzy, Mary-Annick Morel, Elisabeth Murisasco and Irina Nesterenko

Combining Parallel Treebanks and Geo-Tagging
Martin Volk, Anne Goehring and Torsten Marek

Thursday, July 15, 2010 (continued)

Challenges of Cheap Resource Creation
Jirka Hana and Anna Feldman

Discourse Relation Configurations in Turkish and an Annotation Environment
Berfin Aktaş, Cem Bozşahin and Deniz Zeyrek

An Overview of the CRAFT Concept Annotation Guidelines
Michael Bada, Miriam Eckert, Martha Palmer and Lawrence Hunter

Syntactic Tree Queries in Prolog
Gerlof Bouma

An Integrated Tool for Annotating Historical Corpora
Pablo Picasso Feliciano de Faria, Fabio Natanael Kepler and Maria Clara Paixão de Sousa

The Revised Arabic PropBank
Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan and Martha Palmer

Friday, July 16, 2010

08:50–10:30 Session IV (Chair: Massimo Poesio)

08:50–09:15 *PackPlay: Mining Semantic Data in Collaborative Games*
Nathan Green, Paul Breimyer, Vinay Kumar and Nagiza Samatova

09:15–09:40 *A Proposal for a Configurable Silver Standard*
Udo Hahn, Katrin Tomanek, Elena Beisswanger and Erik Faessler

09:40–10:05 *A Hybrid Model for Annotating Named Entity Training Corpora*
Robert Voyer, Valerie Nygaard, Will Fitzgerald and Hannah Copperman

10:05–10:30 *Anatomy of Annotation Schemes: Mapping to GrAF*
Nancy Ide and Harry Bunt

10:30–11:00 Break

Friday, July 16, 2010 (continued)

11:00–12:40 Session V (Chair: Nancy Ide)

11:00–11:25 *Annotating Participant Reference in English Spoken Conversation*
John Niekrasz and Johanna Moore

11:25–11:50 *Design and Evaluation of Shared Prosodic Annotation for Spontaneous French Speech: From Expert Knowledge to Non-Expert Annotation*
Anne Lacheret-Dujour, Nicolas Obin and Mathieu Avanzi

11:50–12:15 *Depends on What the French Say - Spoken Corpus Annotation with and beyond Syntactic Functions*
José Deulofeu, Lucie Duffort, Kim Gerdes, Sylvain Kahane and Paola Pietrandrea

12:15–12:40 *The Annotation Scheme of the Turkish Discourse Bank and an Evaluation of Inconsistent Annotations*
Deniz Zeyrek, Işin Demirşahin, Ayişiği Sevdik-Çalli, Hale Ögel Balaban, and Ümit Deniz Turan

12:40–13:00 Closing remarks