Preface

Welcome to The Linguistic Annotation Workshop (The LAW).

Linguistically annotated corpora play a major role in parsing, information extraction, question answering, machine translation and many other areas of computational linguistics, and provide an empirical testbed for theoretical linguistics research. This has led to a proliferation of annotation systems, frameworks, formats, and schemes. Recognition of the need to harmonize annotation practices and frameworks has become increasingly critical, as witnessed by numerous workshops dealing with different aspects of linguistic annotation over the past few years.

The LAW addresses all aspects of linguistic annotation in a single forum by merging two existing workshop series: NLPXML (Natural Language Processing and XML) and FLAC (Frontiers in Linguistically Annotated Corpora). The goals of the workshop include:

1. The exchange and propagation of research results with respect to the annotation, manipulation and exploitation of corpora, taking into account different applications and theoretical investigations in the field of language technology and research;

2. Working towards harmonization and interoperability from the perspective of the increasingly large number of tools and frameworks that support the creation, instantiation, manipulation, querying, and exploitation of annotated resources;

3. Working towards a consensus on all issues crucial to the advancement of the field of corpus annotation.

These proceedings include 11 long papers, 5 short papers, 4 demo descriptions and 8 posters selected by the program committee from 51 submissions for presentation at the workshop. In addition to these presentations, the workshop includes demonstrations of annotation tools, reports by working groups, and an open discussion session.

We would like to thank the members of the program committee for their timely reviews. We also thank the Workshops Chair and other organizers of ACL-2007 for their support. Finally, we congratulate Adriane Boyd, the winner of the Innovative Student Annotation Award for the paper Discontinuity Revisited: An Improved Conversion to Context-Free Representations.

Branimir Boguraev
Nancy Ide
Adam Meyers
Shigeko Nariyama
Manfred Stede
Janyce Wiebe
Graham Wilcock
Organizers

Workshop Chairs

Branimir Boguraev, IBM T. J. Watson Research Center
Nancy Ide, Vassar College
Adam Meyers, New York University
Shigeko Nariyama, University of Melbourne
Manfred Stede, University of Potsdam
Janyce Wiebe, University of Pittsburgh
Graham Wilcock, University of Helsinki

Program Committee

David Ahn, University of Amsterdam
Lars Ahrenberg, Linköping University
Timothy Baldwin, University of Melbourne
Francis Bond, NICT
Kalina Bontcheva, University of Sheffield
Paul Buitelaar, DFKI
Jean Carletta, University of Edinburgh
Key-Sun Choi, KAIST
Christopher Cieri, Linguistic Data Consortium/University of Pennsylvania
Hamish Cunningham, University of Sheffield
David Day, MITRE Corporation
Thierry Declerck, DFKI
Ludovic Denoyer, LIP6 - University of Paris 6
Tomaz Erjavec, Jozef Stefan Institute
David Farwell, New Mexico State University
Alex Chengyu Fang, City University of Hong Kong
Chuck Fillmore, International Computer Science Institute, Berkeley
Anette Frank, DFKI
John Fry, San Jose State University
Claire Grover, University of Edinburgh
Jan Hajic, Charles University
Ed Hovy, University of Southern California
Baden Hughes, University of Melbourne
Emi Izumi, NICT
Tsai Jia-Lin, Tung Nan Institute of Technology
Aravind Joshi, University of Pennsylvania
Ewan Klein, University of Edinburgh
Mounia Lalmas, University of London
Mike Maxwell, University of Maryland
Chieko Nakabasami, Toyo University
Stephan Oepen, University of Oslo
Kyonghee Paik, KLI
Martha Palmer, University of Colorado
Antonio Pareja-Lora, Universidad Complutense de Madrid / OEG - UPM
Manfred Pinkal, Saarland University
James Pustejovsky, Brandeis University
Owen Rambow, Columbia University
Laurent Romary, MPG-INRIA
Henry Thompson, University of Edinburgh
Erik Tjong Kim Sang, University of Amsterdam
Theresa Wilson, University of Edinburgh
Nainwen Xue, University of Colorado
# Table of Contents

**GrAF: A Graph-based Format for Linguistic Annotations**  
Nancy Ide and Keith Suderman ................................................................. 1

**Efficient Annotation with the Jena ANnotation Environment (JANE)**  
Katrin Tomanek, Joachim Wermter and Udo Hahn ........................................... 9

**Mining Syntactically Annotated Corpora with XQuery**  
Gosse Bouma and Geert Kloosterman ........................................................ 17

**Associating Facial Displays with Syntactic Constituents for Generation**  
Mary Ellen Foster ......................................................................................... 25

**An Annotation Type System for a Data-Driven NLP Pipeline**  
Udo Hahn, Ekaterina Buyko, Katrin Tomanek, Scott Piao, John McNaught, Yoshimasa Tsuruoka and Sophia Ananiadou ........................................... 33

**Discontinuity Revisited: An Improved Conversion to Context-Free Representations**  
Adriane Boyd .............................................................................................. 41

**Usage of XSL Stylesheets for the Annotation of the Sámi Language Corpora.**  
Saara Huhmarniemi, Sjur N. Moshagen and Trond Trosterud ......................... 45

**Criteria for the Manual Grouping of Verb Senses**  
Cecily Jill Duffield, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E. Vieweg, Jenny Davis and Martha Palmer ........................................... 49

**Semi-Automated Named Entity Annotation**  
Kuzman Ganchev, Fernando Pereira, Mark Mandel, Steven Carroll and Peter White ........ 53

**Querying Multimodal Annotation: A Concordancer for GeM**  
Martin Thomas .............................................................................................. 57

**Annotating Chinese Collocations with Multi Information**  
Ruifeng Xu, Qin Lu, Kam-Fai Wong and Wenjie Li ......................................... 61

**Computing Translation Units and Quantifying Parallelism in Parallel Dependency Treebanks**  
Matthias Buch-Kromann ................................................................................ 69

**Adding Semantic Role Annotation to a Corpus of Written Dutch**  
Paola Monachesi, Gerwert Stevens and Jantine Trapman ................................. 77

**A Search Tool for Parallel Treebanks**  
Martin Volk, Joakim Lundborg and Maël Mettler ......................................... 85

**Annotating Expressions of Appraisal in English**  
Jonathon Read, David Hope and John Carroll .............................................. 93
Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation
   Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll,
   Kevin Seppi and Deryle Lonsdale ................................................................. 101

Combining Independent Syntactic and Semantic Annotation Schemes
   Marc Verhagen, Amber Stubbs and James Pustejovsky .................................. 109

XARA: An XML- and Rule-based Semantic Role Labeler
   Gerwert Stevens ................................................................. 113

ITU Treebank Annotation Tool
   Gölşen Eryiğit ................................................................. 117

Two Tools for Creating and Visualizing Sub-sentential Alignments of Parallel Text
   Ulrich Germann ................................................................. 121

Building Chinese Sense Annotated Corpus with the Help of Software Tools
   Yunfang Wu, Peng Jin, Tao Guo and Shiwen Yu ........................................ 125

Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations
   Ryu Iida, Mamoru Komachi, Kentaro Inui and Yuji Matsumoto .......................... 132

Web-based Annotation of Anaphoric Relations and Lexical Chains
   Maik Stührenberg, Daniela Goecke, Nils Diewald, Alexander Mehler and Irene Cramer .... 140

Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus
   Kepa Joseba Rodríguez, Stefanie Dipper, Michael Götze, Massimo Poesio, Giuseppe Riccardi,
   Christian Raymond and Joanna Rabiega-Wiśniewska .................................... 148

PoCoS - Potsdam Coreference Scheme
   Olga Krasavina and Christian Chiarcos .................................................. 156

Multiple-step Treebank Conversion: From Dependency to Penn Format
   Cristina Bosco ................................................................. 164

Experiments with an Annotation Scheme for a Knowledge-rich Noun Phrase Interpretation System
   Roxana Girju ................................................................. 168

IGT-XML: An XML Format for Interlinearized Glossed Text
   Alexis Palmer and Katrin Erk .................................................. 176

Shared Corpora Working Group Report
   Adam Meyers, Nancy Ide, Ludovic Denoyer and Yusuke Shinyama ...................... 184

Discourse Annotation Working Group Report
   Manfred Stede, Janyc Wiebe, Eva Hajcova, Brian Reese, Simone Teufel, Bonnie Webber and
   Theresa Wilson ................................................................. 191
Workshop Program

Thursday, June 28, 2007

Session 1: Introduction and Long Papers

14:30–14:55  Introduction to the Workshop

14:55–15:20  GrAF: A Graph-based Format for Linguistic Annotations  
Nancy Ide and Keith Suderman

15:20–15:45  Efficient Annotation with the Jena ANnotation Environment (JANE)  
Katrin Tomaneek, Joachim Wermter and Udo Hahn

15:45–16:15  Break

Session 2: Long Papers

16:15–16:40  Mining Syntactically Annotated Corpora with XQuery  
Gosse Bouma and Geert Kloosterman

16:40–17:05  Assocating Facial Displays with Syntactic Constituents for Generation  
Mary Ellen Foster

17:05–17:30  An Annotation Type System for a Data-Driven NLP Pipeline  
Udo Hahn, Ekaterina Buyko, Katrin Tomaneek, Scott Piao, John McNaught,  
Yoshimasa Tsuruoka and Sophia Ananiadou

Demonstration and Poster Session

17:30–18:30  Demonstrations and Posters  
(Listed at end of program)
Friday, June 29, 2007

**Session 3: Working Groups**

09:30–10:00 Shared Corpora Working Group Report
Adam Meyers, Nancy Ide, Ludovic Denoyer and Yusuke Shinyama

10:00–10:45 Panel Session on Discourse Annotation
Manfred Stede, Eva Hajicova, Brian Reese, Simone Teufel, Bonnie Webber and Theresa Wilson

10:45–11:15 Break

**Session 4: Short Papers**

11:15–11:30 *Discontinuity Revisited: An Improved Conversion to Context-Free Representations*
Adriane Boyd

11:30–11:45 *Usage of XSL Stylesheets for the Annotation of the Sámi Language Corpora.*
Saara Huhmarniemi, Sjur N. Moshagen and Trond Trosterud

11:45–12:00 *Criteria for the Manual Grouping of Verb Senses*
Cecily Jill Duffield, Jena D. Hwang, Susan Windisch Brown, Dmitriy Dligach, Sarah E. Vieweg, Jenny Davis and Martha Palmer

12:00–12:15 *Semi-Automated Named Entity Annotation*
Kuzman Ganchev, Fernando Pereira, Mark Mandel, Steven Carroll and Peter White

12:15–12:30 *Querying Multimodal Annotation: A Concordancer for GeM*
Martin Thomas

12:30–14:30 Lunch
Friday, June 29, 2007 (continued)

Session 5: Long Papers

14:30–14:55 Annotating Chinese Collocations with Multi Information
Ruifeng Xu, Qin Lu, Kam-Fai Wong and Wenjie Li

14:55–15:20 Computing Translation Units and Quantifying Parallelism in Parallel Dependency Treebanks
Matthias Buch-Kromann

15:20–15:45 Adding Semantic Role Annotation to a Corpus of Written Dutch
Paola Monachesi, Gerwert Stevens and Jantine Trapman

15:45–16:15 Break

Session 6: Long Papers

16:15–16:40 A Search Tool for Parallel Treebanks
Martin Volk, Joakim Lundborg and Maël Mettler

16:40–17:05 Annotating Expressions of Appraisal in English
Jonathon Read, David Hope and John Carroll

17:05–17:30 Active Learning for Part-of-Speech Tagging: Accelerating Corpus Annotation
Eric Ringger, Peter McClanahan, Robbie Haertel, George Busby, Marc Carmen, James Carroll, Kevin Seppi and Deryle Lonsdale

Discussion Session

17:30–18:30 Open Discussion
Demonstrations and Posters

Demonstrations

*Combining Independent Syntactic and Semantic Annotation Schemes*
Marc Verhagen, Amber Stubbs and James Pustejovsky

*XARA: An XML- and Rule-based Semantic Role Labeler*
Gerwert Stevens

*ITU Treebank Annotation Tool*
Gülşen Eryiğit

*Two Tools for Creating and Visualizing Sub-sentential Alignments of Parallel Text*
Ulrich Germann

Posters

*Building Chinese Sense Annotated Corpus with the Help of Software Tools*
Yunfang Wu, Peng Jin, Tao Guo and Shiwen Yu

*Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations*
Ryu Iida, Mamoru Komachi, Kentaro Inui and Yuji Matsumoto

*Web-based Annotation of Anaphoric Relations and Lexical Chains*
Maik Stührenberg, Daniela Goecake, Nils Diewald, Alexander Mehler and Irene Cramer

*Standoff Coordination for Multi-Tool Annotation in a Dialogue Corpus*
Kepa Joseba Rodríguez, Stefanie Dipper, Michael Götze, Massimo Poesio, Giuseppe Riccardi, Christian Raymond and Joanna Rabiega-Wiśniewska

*PoCoS - Potsdam Coreference Scheme*
Olga Krasavina and Christian Chiarcos

*Multiple-step Treebank Conversion: From Dependency to Penn Format*
Cristina Bosco
Demonstrations and Posters (continued)

*Experiments with an Annotation Scheme for a Knowledge-rich Noun Phrase Interpretation System*
Roxana Girju

*IGT-XML: An XML Format for Interlinearized Glossed Text*
Alexis Palmer and Katrin Erk