Preface

The Linguistic Annotation Workshop (The LAW) is organized annually by the Association for Computational Linguistics Special Interest Group for Annotation (ACL SIGANN). It provides a forum to facilitate the exchange and propagation of research results concerned with the annotation, manipulation, and exploitation of corpora; work towards 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.

Last fall, when workshop proposals were solicited, we were asked for a tagline that would telegraph the essence of LAW. Naturally, this prompted a healthy dose of wordsmithing on the part of the organizing committee. The initial suggestions (including “Annotation schemers unite!” and “Don’t just annoTATE—annoGREAT!”) were deemed too corny. Then, playing on the abbreviation “LAW”, various legal puns emerged: “the fine print of linguistic annotation”; “the letter and spirit of linguistic annotation”; “LAW, the authority on linguistic annotation”; “LAW, where language is made to behave”; and so forth. “Annotation schemers on parole” took the punning to the extreme (as students of Saussure will recognize).

In the end, we settled on “LAW: Due process for linguistic annotation”. The concept of “due process” underscores the care required not just to annotate, but to annotate well. To produce a high-quality linguistic resource, diligence is required in all phases: assembling the source data; designing and refining the annotation scheme and guidelines; choosing or developing appropriate annotation software and data formats; applying automatic tools for preprocessing and provisional annotation; selecting and training human annotators; implementing quality control procedures; and documenting and distributing the completed resource.

The 14 papers in this year’s workshop study methods for annotation in the domains of emotion and attitude; conversations and discourse structure; events and causality; semantic roles; and translation (among others). Compared to previous years, syntax plays a much smaller role: indeed, this may be the first ever LAW where no paper has the word “treebank” in the title. (We leave it to the reader to speculate whether this reflects broader trends in the field or has an innocuous explanation.) Also groundbreaking in this year’s LAW will be a best paper award, to be announced at the workshop.

LAW XI would not have been possible without the fine contributions of the authors; the remarkably thorough and thoughtful reviews from the program committee; and the sage guidance of the organizing committee. Two invited talks will add multilingual perspective to the program, Deniz Zeyrek and Johan Bos having generously agreed to share their wisdom. We thank our publicity chair, Marc Verhagen, as well as those who have worked to coordinate the various aspects of EACL workshops, including logistics and publications.

We hope that after reading the collected wisdom in this volume, you will be empowered to give the linguistic annotation process its due.

Nathan Schneider and Nianwen Xue, program co-chairs
Program Co-chairs:

Nathan Schneider  Georgetown University  
Nianwen Xue  Brandeis University

Publicity Chair:

Marc Verhagen  Brandeis University

Program Committee:

Adam Meyers  New York University  
Alexis Palmer  University of North Texas  
Amália Mendes  University of Lisbon  
Amir Zeldes  Georgetown University  
Andrew Gargett  University of Birmingham  
Annemarie Friedrich  Saarland University  
Antonio Pareja-Lora  Universidad Complutense de Madrid / ATLAS, UNED  
Archna Bhatia  IHMC  
Benoît Sagot  Université Paris Diderot  
Bonnie Webber  University of Edinburgh  
Collin Baker  ICSI Berkeley  
Dirk Hovy  University of Copenhagen  
Djamé Seddah  Paris-Sorbonne University  
Els Lefever  Ghent University  
Heike Zinsmeister  University of Hamburg  
Ines Rehbein  Leibniz Science Campus, Institute for German Language and Heidelberg University  
Joel Tetreault  Grammarly  
John S. Y. Lee  City University of Hong Kong  
Josef Ruppenhofer  Leibniz Science Campus, Institute for German Language and Heidelberg University  
Katrin Tomanek  OpenTable  
Kemal Oflazer  Carnegie Mellon University—Qatar  
Kilian Evang  University of Groningen  
Kim Gerdes  Sorbonne Nouvelle  
Kiril Simov  Bulgarian Academy of Sciences  
Lori Levin  Carnegie Mellon University  
Manfred Stede  University of Potsdam  
Marie Candito  Université Paris Diderot  
Markus Dickinson  Indiana University  
Martha Palmer  University of Colorado at Boulder  
Massimo Poesio  University of Essex  
Nancy Ide  Vassar College  
Nicoletta Calzolari  CNR-ILC  
Nizar Habash  New York University Abu Dhabi  
Özlem Çetinoğlu  University of Stuttgart  
Pablo Faria  State University of Campinas  
Ron Artstein  Institute for Creative Technologies, USC  
Sandra Kübler  Indiana University  
Stefanie Dipper  Ruhr University Bochum  
Tomaž Erjavec  Jožef Stefan Institute
Udo Hahn  Jena University  
Valia Kordoni  Humboldt University of Berlin  

**SIGANN Organizing Committee:**

- Stefanie Dipper  Ruhr University Bochum  
- Annemarie Friedrich  Saarland University  
- Chu-Ren Huang  The Hong Kong Polytechnic University  
- Nancy Ide  Vassar College  
- Lori Levin  Carnegie Mellon University  
- Adam Meyers  New York University  
- Antonio Pareja-Lora  Universidad Complutense de Madrid / ATLAS, UNED  
- Massimo Poesio  University of Essex  
- Sameer Pradhan  Boulder Learning, Inc.  
- Ines Rehbein  Leibniz Science Campus, Institute for German Language and Heidelberg University  
- Manfred Stede  University of Potsdam  
- Katrin Tomanek  OpenTable  
- Fei Xia  University of Washington  
- Heike Zinsmeister  University of Hamburg
# Table of Contents

**Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation**  
Sven Buechel and Udo Hahn ................................................................. 1

**Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus**  
Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato and Brian Provenzale .......... 13

**Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task**  
Merel Scholman and Vera Demberg .................................................. 24

**A Code-Switching Corpus of Turkish-German Conversations**  
Özlem Çetinoğlu ................................................................. 34

**Annotating omission in statement pairs**  
Héctor Martínez Alonso, Amaury Delamaire and Benoît Sagot ................................. 41

**Annotating Speech, Attitude and Perception Reports**  
Corien Bary, Leopold Hess, Kees Thijs, Peter Berck and Iris Hendrickx ......................... 46

**Consistent Classification of Translation Revisions: A Case Study of English-Japanese Student Translations**  
Atsushi Fujita, Kikuko Tanabe, Chiho Toyoshima, Mayuka Yamamoto, Kyo Kageura and Anthony Hartley .................................................. 57

**Representation and Interchange of Linguistic Annotation: An In-Depth, Side-by-Side Comparison of Three Designs**  
Richard Eckart de Castilho, Nancy Ide, Emanuele Lapponi, Stephan Oepen, Keith Suderman, Erik Velldal and Marc Verhagen .................................................. 67

**TDB 1.1: Extensions on Turkish Discourse Bank**  
Deniz Zeyrek and Murathan Kurfalı .................................................. 76

**Two Layers of Annotation for Representing Event Mentions in News Stories**  
Maria Pia di Buono, Martin Tutek, Jan Šnajder, Goran Glavaš, Bojana Dalbelo Bašić and Natasa Milic-Frayling .................................................. 82

**Word Similarity Datasets for Indian Languages: Annotation and Baseline Systems**  
Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava and Manish Shrivastava. 91

**The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations**  
Jesse Dunietz, Lori Levin and Jaime Carbonell .................................. 95

**Catching the Common Cause: Extraction and Annotation of Causal Relations and their Participants**  
Ines Rehbein and Josef Ruppenhofer .................................................. 105

**Assessing SRL Frameworks with Automatic Training Data Expansion**  
Silvana Hartmann, Ėva Mújdricza-Maydt, Iliya Kuznetsov, Iryna Gurevych and Anette Frank . . 115
Workshop Program

Monday, April 3, 2017

9:30–9:40  Welcome

9:40–10:40  Invited Talk I: The TED-Multilingual Discourse Bank
Deniz Zeyrek

10:40–11:00  Emotion

10:40–11:00  Readers vs. Writers vs. Texts: Coping with Different Perspectives of Text Understanding in Emotion Annotation
Sven Buechel and Udo Hahn

11:00–11:30  Coffee Break

11:30–12:10  Conversations & Discourse

11:30–11:50  Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus
Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato and Brian Provenzale

11:50–12:10  Crowdsourcing discourse interpretations: On the influence of context and the reliability of a connective insertion task
Merel Scholman and Vera Demberg
Monday, April 3, 2017 (continued)

12:10–13:00 Posters

A Code-Switching Corpus of Turkish-German Conversations
Özlem Çetinoğlu

Annotating omission in statement pairs
Héctor Martínez Alonso, Amaury Delamaire and Benoît Sagot

Annotating Speech, Attitude and Perception Reports
Corien Bary, Leopold Hess, Kees Thijs, Peter Berck and Iris Hendrickx

Consistent Classification of Translation Revisions: A Case Study of English-Japanese Student Translations
Atsushi Fujita, Kikuko Tanabe, Chiho Toyoshima, Mayuka Yamamoto, Kyo Kageura and Anthony Hartley

Representation and Interchange of Linguistic Annotation: An In-Depth, Side-by-Side Comparison of Three Designs
Richard Eckart de Castilho, Nancy Ide, Emanuele Lapponi, Stephan Oepen, Keith Suderman, Erik Velldal and Marc Verhagen

TDB 1.1: Extensions on Turkish Discourse Bank
Deniz Zeyrek and Murathan Kurfalı

Two Layers of Annotation for Representing Event Mentions in News Stories
Maria Pia di Buono, Martin Tutek, Jan Šnajder, Goran Glavaš, Bojana Dalbelo Bašić and Natasa Milic-Frayling

Word Similarity Datasets for Indian Languages: Annotation and Baseline Systems
Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava and Manish Shrivastava

13:00–14:30 Lunch
Monday, April 3, 2017 (continued)

14:30–15:00 Posts (contd.)

15:00–16:00 Causality & Semantic Roles

15:00–15:20 *The BECauSE Corpus 2.0: Annotating Causality and Overlapping Relations*
Jesse Dunietz, Lori Levin and Jaime Carbonell

15:20–15:40 *Catching the Common Cause: Extraction and Annotation of Causal Relations and their Participants*
Ines Rehbein and Josef Ruppenhofer

15:40–16:00 *Assessing SRL Frameworks with Automatic Training Data Expansion*
Silvana Hartmann, Éva Mújdricza-Maydt, Ilia Kuznetsov, Iryna Gurevych and Anette Frank

16:00–16:30 Coffee Break

16:30–17:30 Invited Talk II: Cross-lingual Semantic Annotation
Johan Bos

17:30–18:00 Discussion and Wrap-up