LAW-MWE-CxG 2018

Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions

Proceedings of the Workshop

August 25-26, 2018
Santa Fe, New Mexico, USA

Copyright of each paper stays with the respective authors (or their employers).

ISBN 978-1-948087-51-3

Introduction

The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)[1] took place on August 25-26, 2018 in Santa Fe (USA), in conjunction with the 27th International Conference on Computational Linguistics (COLING 2018). This was simultaneously the 12th edition of the Linguistic Annotation Workshop (LAW XII) and the 14th edition of the Workshop on Multiword Expressions (MWE 2018). The event was organized and sponsored by the Special Interest Group for Annotation (SIGANN)[2] and the Special Interest Group on the Lexicon (SIGLEX)[3] of the Association for Computational Linguistics (ACL). It was also endorsed by the Special Interest Group on Computational Semantics (SIGSEM)[4].

The workshop brought together three divergent, but overlapping, research communities studying linguistic annotation, multiword expressions, and grammatical constructions. *Linguistic annotation* of natural language corpora is the backbone of supervised methods for statistical natural language processing. It is also essential for evaluation of both rule-based and supervised systems and can help formalize and study linguistic phenomena. *Multiword expressions* (MWEs) are word combinations, such as *all of a sudden*, *hot dog*, *to pay a visit* or *to pull one's leg*, which exhibit lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies. Computational research on MWEs encompasses NLP modeling and processing, as well as annotation. *Grammatical constructions* (which include MWEs) are conventional associations of lexical, syntactic, and pragmatic information, such as *the-ComparAdjP-the-ComparAdjP* (*the more the merrier*, *the higher the better*, etc.). In the framework of Construction Grammar (CxG), linguistic knowledge about constructions is captured in an inventory of form-meaning pairings of varying degrees of internal complexity and lexical fixedness.

In order to promote synergies between these three largely complementary scientific topics, we called for papers in two tracks: the regular *research track* and the *shared task track*. The topics promoted in the research track included, but were not limited, to:

- Joint topics on constructions, annotation, and MWEs:
  - MWE and construction annotation in corpora and treebanks
  - MWE and construction representation in manually and automatically constructed lexical resources
  - Extending MWE discovery and identification methods to constructions
  - MWEs and constructions (and their annotations) in language acquisition and in non-standard language (e.g. tweets, forums, spontaneous speech)
  - Evaluation of MWE and construction annotation and processing techniques
  - Computationally-applicable theoretical studies on MWEs and constructions in psycholinguistics, corpus linguistics and grammar formalisms, and/or how such studies can impact annotation of constructions
- Annotation-specific topics:
  - Annotation procedures, whether manual or automatic, including machine learning and knowledge-based methods
  - Maintenance and interactive exploration of annotation structures and annotated data
  - Qualitative and quantitative annotation evaluation
  - Linguistic considerations, representation formats and exploration tools for merged annotations of different phenomena

[1] http://multiword.sourceforge.net/lawmwecxg2018/
[2] https://www.cs.vassar.edu/sigann/
[3] http://siglex.org/
[4] http://www.sigsem.org

  - Standards, best practices, documentation, interoperability, and comparison of annotation schemes
  - Development, evaluation and innovative use of annotation software frameworks
- MWE-specific topics:
  - Original MWE discovery and identification methods
  - MWE processing in syntactic and semantic frameworks (e.g. HPSG, LFG, TAG, Universal Dependencies, WSD, semantic parsing), and in end-user applications (e.g. summarization, machine translation)

We received 34 submissions (22 long and 12 short papers) in the research track, one of them was further withdrawn. We selected 16 long papers and 6 short ones. From those, 9 papers were presented orally and the remaining 13 as posters. The overall selectivity rate was 65%. Of the 22 presented papers, 16 concerned linguistic annotation, 14 multiword expressions, and 5 constructions. As many as 11 papers addressed at least 2 of the 3 workshop topics, which makes us believe that the intended synergy effect has been achieved.

The shared task track was the culmination of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions[5], preceded by a corpus annotation campaign in 20 languages, coordinated by the PARSEME[6] scientific network. The shared task attracted 12 teams, which presented 17 systems, most of them highly multilingual. Eight of those teams submitted system description papers, all were selected and presented as posters. The reviewing modalities were different in this track (notably the requirement of originality did not apply), therefore we do not count these papers in the workshop selectivity rate.

In addition to the oral and poster sessions, the workshop featured three invited talks:

- Lori Levin, Carnegie Mellon University (Pittsburgh, USA), "Annotation Schemes for Surface Construction Labeling"
- Adam Przepiórkowski, University of Warsaw and Polish Academy of Sciences (Warsaw, Poland), "From Lexical Functional Grammar to Enhanced Universal Dependencies"
- Nathan Schneider, Georgetown University (USA), "Leaving no token behind: comprehensive (and delicious) annotation of MWEs and supersenses"

We are grateful to the paper authors for their valuable contributions, the members of the Program Committee for their thorough and timely reviews, all members of the organizing committee for the fruitful collaboration, the shared task organizers, annotators, and system developers for their hard work, and all the workshop participants for their interest in this event. Our thanks also go to the COLING 2018 organizers for their support, as well as to SIGLEX, SIGANN and SIGSEM, for their endorsement.

Agata Savary, Carlos Ramisch, Nancy Ide and Adam Meyers

[5] http://multiword.sourceforge.net/sharedtask2018
[6] http://www.parseme.eu

Organizers

Workshop Organizers

Nancy Ide, Vassar College, USA
Adam Meyers, New York University, USA
Carlos Ramisch, Aix Marseille University, France
Agata Savary, University of Tours, France

Program Committee Chairs

Jena D. Hwang, Institute for Human and Machine Cognition, USA
Miriam R L Petruck, ICSI, USA
Sameer Pradhan, cemantix.org and Vassar College, New York, USA
Carlos Ramisch, Aix Marseille University, France
Agata Savary, University of Tours, France
Nathan Schneider, Georgetown University, USA

Publication Chairs

Melanie Andresen, Universität Hamburg, Germany
Agata Savary, University of Tours, France

Publicity Chairs

Adam Meyers, New York University, USA
Agata Savary, University of Tours, France

Programme Committee

Omri Abend, The Hebrew University of Jerusalem, Israel
Ron Artstein, Institute for Creative Technologies, USC, USA
Timothy Baldwin, University of Melbourne, Australia
Libby Barak, University of Toronto, Canada
Riyaz A. Bhat, University of Colorado Boulder, USA
Archna Bhatia, Carnegie Mellon University, USA
Ann Bies, Linguistic Data Consortium, University of Pennsylvania, USA
Francis Bond, Nanyang Technological University, Singapore
Claire Bonial, Air-Force Research Lab, USA
Antonio Branco, University of Lisbon, Portugal
Miriam Butt, Universität Konstanz, Germany
Aoife Cahill, Educational Testing Service, USA
Nicoletta Calzolari, National Research Council of Italy, Italy
Marie Candito, Paris Diderot University, France
Özlem Çetinoğlu, University of Stuttgart, Germany
Christian Chiarcos, University of Frankfurt, Germany
Anastasia Christofidou, Academy of Athens, Greece
Kathryn Conger, University of Colorado Boulder, USA
Matthieu Constant, Université de Lorraine, France
Paul Cook, University of New Brunswick, Canada
Silvio Ricardo Cordeiro, Federal University of Rio Grande do Sul, Brazil
Béatrice Daille, Nantes University, France

Gaël Dias, University of Caen Basse-Normandie, France
Stefanie Dipper, RUHR University at Bochum, Germany
Ellen Dodge, ICSI, University of California at Berkeley, USA
Jonathan Dunn, Illinois Institute of Technology, USA
Kilian Evang, Institute for Language and Information Science, Germany
Federico Fancellu, University of Edinburgh, UK
Pablo Faria, University of Campinas, Brazil
Joaquim Ferreira da Silva, New University of Lisbon, Portugal
Aggeliki Fotopoulou, ILSP/RC "Athena", Greece
Mirjam Fried, Charles University, Czech Republic
Andrew Gargett, University of Birmingham, UK
Kim Gerdes, University of Paris, France
Voula Giouli, Institute for Language and Speech Processing, Greece
Roxana Girju, University of Illinois at Urbana-Champaign, USA
Udo Hahn, Jena University, Germany
Chikara Hashimoto, Yahoo!Japan, Japan
Kyoko Hirose Ohara, Keio University, Japan
Nancy Ide, Vassar College, USA
Kyo Kageura, University of Tokyo, Japan
Dimitrios Kokkinakis, University of Gothenburg, Sweden
Valia Kordoni, Humboldt University Berlin, Germany
Ioannis Korkontzelos, Edge Hill University, UK
Brigitte Krenn, Austrian Research Institute for Artificial Intelligence, Austria
Cvetana Krstev, University of Belgrade, Serbia
Tita Kyriakopoulou, University Paris-Est Marne-la-Vallee, France
Vicky Lai, University of Arizona, USA
Eric Laporte, University Paris-Est Marne-la-Vallee, France
John S. Y. Lee, City University of Hong Kong, Hong Kong
Els Lefever, Ghent University, Belgium
Lori Levin, Carnegie Mellon University, USA
Ben Lyngfelt, University of Gothenburg, Sweden
Stella Markantonatou, Institute for Language and Speech Processing, Greece
Hector Martínez Alonso, INRIA, France
Amália Mendes, University of Lisbon, Portugal
Adam Meyers, New York University, USA
Laura A. Michaelis, University of Colorado Boulder, USA
Johanna Monti, "L'Orientale" University of Naples, Italy
Éva Mújdricza-Maydt, University of Heidelberg, Germany
Stefan Müller, Humboldt University, Germany
Preslav Nakov, Qatar Computing Research Institute, HBKU, Qatar
Anna Nedoluzhko, Charles University, Czech Republic
Kiki Nikiforidou, School of Philosophy, Greece
Joakim Nivre, Uppsala University, Sweden
Michael Oakes, University of Wolverhampton, UK
Jan Odijk, University of Utrecht, The Netherlands
Kemal Oflazer, Carnegie Mellon University, USA
Gertjan van Noord, University of Groningen, The Netherlands
Petya Osenova, Bulgarian Academy of Sciences, Bulgaria
Simon Ostermann, Saarland University, Germany
Lilja Øvrelid, University of Oslo, Norway
Alexis Palmer, University of North Texas, USA

Haris Papageorgiou, Institute for Language and Speech Processing, Greece
Antonio Pareja-Lora, Universidad Complutense de Madrid, DMEG and ATLAS, Spain
Yannick Parmentier, Université d'Orléans, France
Carla Parra Escartín, Dublin City University, Ireland
Agnieszka Patejuk, Institute of Computer Science, Polish Academy of Sciences, Poland
Pavel Pecina, Charles University, Czech Republic
Scott Piao, Lancaster University, UK
Thierry Poibeau, CNRS and École Normale Supérieure, France
James Pustejovsky, Brandeis University, USA
Ines Rehbein, University of Heidelberg, Germany
Arndt Riester, University of Stuttgart, Germany
Josef Ruppenhofer, Heidelberg University, Germany
Manfred Sailer, Goethe-Universität Frankfurt am Main, Germany
Sabine Schulte im Walde, University of Stuttgart, Germany
Djamé Seddah, Paris-Sorbonne University, France
Michael Spranger, Sony Labs, Japan
Manfred Stede, University of Potsdam, Germany
Sara Stymne, Uppsala University, Sweden
Stan Szpakowicz, University of Ottawa, Canada
Tiago Torrent, Federal University of Juiz de Fora, Brazil
Beata Trawinski, Institut für Deutsche Sprache Mannheim, Germany
Yuancheng Tu, Microsoft, USA
Ruben Urizar, University of the Basque Country, Spain
Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil
Veronika Vincze, Hungarian Academy of Sciences, Hungary
Michael Wiegand, Saarland University, Germany
Susan Windisch Brown, University of Colorado Boulder, USA
Shuly Wintner, University of Haifa, Israel
Andreas Witt, Institut fur Deutsche Sprache, Germany
Amir Zeldes, Georgetown University, USA
Heike Zinsmeister, Universität Hamburg, Germany

Invited Speakers

Lori Levin, Carnegie Mellon University, Pittsburgh, USA
Adam Przepiórkowski, University of Warsaw and Polish Academy of Sciences, Warsaw, Poland
Nathan Schneider, Georgetown University, USA

Table of Contents

Short Papers

Shared Task Papers

Conference Program

Saturday, August 25, 2018

8:55–9:00 *Opening*

Session 1: Multiword Expressions

9:00–10:00 **Invited talk:** *Leaving no token behind: comprehensive (and delicious) annotation of MWEs and supersenses*
Nathan Schneider

10:00–10:30 *Poster boosters of research papers*

10:30–11:00 *Coffee break*

Session 2: Multiword Expressions

11:00–11:30 *Fixed Similes: Measuring aspects of the relation between MWE idiomatic semantics and syntactic flexibility*
Stella Markantonatou, Panagiotis Kouris and Yanis Maistros

11:30–12:00 *Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions*
Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya and Abigail Walsh

12:00–12:10 *TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model*
Jakub Waszczuk

12:10–12:20 *TRAPACC and TRAPACCS at PARSEME Shared Task 2018: Neural Transition Tagging of Verbal Multiword Expressions*
Regina Stodden, Behrang QasemiZadeh and Laura Kallmeyer

12:20–12:30 *Poster boosters of 6 other shared task papers*

12:30–13:50 *Lunch break*

13:50–15:50 **Session 3: Posters**

The RST Spanish-Chinese Treebank
Shuyuan Cao, Iria da Cunha and Mikel Iruskieta

The Other Side of the Coin: Unsupervised Disambiguation of Potentially Idiomatic Expressions by Contrasting Senses
Hessel Haagsma, Malvina Nissim and Johan Bos

Fine-Grained Termhood Prediction for German Compound Terms Using Neural Networks
Anna Hätty and Sabine Schulte im Walde

Verbal Multiword Expressions in Basque Corpora
Uxoa Iñurrieta, Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez-Dios, Antton Gurrutxaga, Ruben Urizar and Iñaki Alegria

Towards a Computational Lexicon for Moroccan Darija: Words, Idioms, and Constructions
Jamal Laoudi, Claire Bonial, Lucia Donatelli, Stephen Tratz and Clare Voss

Cooperating Tools for MWE Lexicon Management and Corpus Annotation
Yuji Matsumoto, Akihiko Kato, Hiroyuki Shindo and Toshio Morita
Developing and Evaluating Annotation Procedures for Twitter Data during Hazard Events
Kevin Stowe, Martha Palmer, Jennings Anderson, Marina Kogan, Leysia Palen, Kenneth M. Anderson, Rebecca Morss, Julie Demuth and Heather Lazrus
CRF-Seq and CRF-DepTree at PARSEME Shared Task 2018: Detecting Verbal MWEs using Sequential and Dependency-Based Approaches
Erwan Moreau, Ashjan Alsulaimani, Alfredo Maldonado and Carl Vogel
Mumpitz at PARSEME Shared Task 2018: A Bidirectional LSTM for the Identification of Verbal Multiword Expressions
Rafael Ehren, Timm Lichte and Younes Samih
TRAVERSAL at PARSEME Shared Task 2018: Identification of Verbal Multiword Expressions Using a Discriminative Tree-Structured Model
Jakub Waszczuk
Veyn at PARSEME Shared Task 2018: Recurrent neural networks for VMWE identification
Nicolas Zampieri, Manon Scholivet, Carlos Ramisch and Benoit Favre

15:50–16:20 *Coffee break*

16:20–18:00 *Session 4: Posters*
"Fingers in the Nose": Evaluating Speakers' Identification of Multi-Word Expressions Using a Slightly Gamified Crowdsourcing Platform
Karën Fort, Bruno Guillaume, Matthieu Constant, Nicolas Lefèbvre and Yann-Alan Pilatte
Do Character-Level Neural Network Language Models Capture Knowledge of Multiword Expression Compositionality?
Ali Hakimi Parizi and Paul Cook
A Treebank for the Healthcare Domain
Nganthoibi Oinam, Diwakar Mishra, Pinal Patel, Narayan Choudhary and Hitesh Desai
All Roads Lead to UD: Converting Stanford and Penn Parses to English Universal Dependencies with Multilayer Annotations
Siyao Peng and Amir Zeldes
Constructing an Annotated Corpus of Verbal MWEs for English
Abigail Walsh, Claire Bonial, Kristina Geeraert, John P. McCrae, Nathan Schneider and Clarissa Somers
A Syntax-Based Scheme for the Annotation and Segmentation of German Spoken Language Interactions
Swantje Westpfahl and Jan Gorisch
Deep-BGT at PARSEME Shared Task 2018: Bidirectional LSTM-CRF Model for Verbal Multiword Expression Identification
Gözde Berk, Berna Erden and Tunga Güngör
GBD-NER at PARSEME Shared Task 2018: Multi-Word Expression Detection Using Bidirectional Long-Short-Term Memory Networks and Graph-Based Decoding
Tiberiu Boroș and Ruxandra Burtica
TRAPACC and TRAPACCS at PARSEME Shared Task 2018: Neural Transition Tagging of Verbal Multiword Expressions
Regina Stodden, Behrang QasemiZadeh and Laura Kallmeyer
VarIDE at PARSEME Shared Task 2018: Are Variants Really as Alike as Two Peas in a Pod?
Caroline Pasquer, Carlos Ramisch, Agata Savary and Jean-Yves Antoine

Sunday, August 26, 2018

Session 5: Constructions

9:00–10:00 **Invited talk**: *Annotation Schemes for Surface Construction Labeling*
Lori Levin

10:00–10:30 *The Interplay of Form and Meaning in Complex Medical Terms: Evidence from a Clinical Corpus*
Leonie Grön, Ann Bertels and Heylen Kris

10:30–11:00 *Coffee break*

Session 6: Constructions

Processing MWEs: Neurocognitive Bases of Verbal MWEs and Lexical Cohesiveness within MWEs
Shohini Bhattasali, Murielle Fabre and John Hale

Discourse and lexicons: lexemes, MWEs, grammatical constructions and compositional word combinations to signal discourse relations
Laurence Danlos

From Chinese word segmentation to extraction of constructions: two sides of the same algorithmic coin
Jean-Pierre Colson

12:30–13:50 *Lunch break*

Session 7: Linguistic annotation

13:50–14:50 **Invited talk:** *From Lexical Functional Grammar to Enhanced Universal Dependencies*
Adam Przepiórkowski (joint work with Agnieszka Patejuk)

14:50–15:20 *Annotation of Tense and Aspect Semantics for Sentential AMR*
Lucia Donatelli, Michael Regan, William Croft and Nathan Schneider

15:20–15:50 *An Annotated Corpus of Picture Stories Retold by Language Learners*
Christine Köhn and Arne Köhn

15:50–16:20 *Coffee break*

Session 8: Linguistic annotation

16:20–16:40 *Improving Domain Independent Question Parsing with Synthetic Treebanks*
Halim-Antoine Boukaram, Nizar Habash, Micheline Ziadee and Majd Sakr

16:40–17:40 *Business meeting*