MWE 2023

The 19th Workshop on Multiword Expressions (MWE 2023)

Proceedings of the Workshop

May 6, 2023

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 978-1-959429-59-3

Introduction

The 19th Workshop on Multiword Expressions (MWE 2023), colocated with EACL 2023 in Dubrovnik, Croatia, will take place as a hybrid event (on-site and virtual) on May 6, 2023. MWE 2023 is organized and sponsored by the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL).

Multiword expressions (MWEs) present an interesting research area due to the lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies they exhibit. Given their irregular nature, they pose complex problems in linguistic modeling (e.g. annotation), NLP tasks (e.g. parsing), and end-user applications (e.g. natural language understanding and MT). For the past two decades, modeling and processing MWEs for NLP has been the topic of the MWE workshop. Impressive progress has been made in the field, but our understanding of MWEs still requires much research considering their need and usefulness in NLP applications. This is also relevant to domain-specific NLP pipelines that need to tackle terminologies that often manifest as MWEs. For the 19th edition of the workshop, we identified the following topics on which contributions were particularly encouraged:

- MWE processing and identification in specialized languages and domains: Multiword terminology extraction from domain-specific corpora is of particular importance to various applications, such as MT, or for the identification and monitoring of neologisms and technical jargon. We expect approaches that deal with the processing of MWEs as well as the processing of terminology in specialized domains can benefit from each other.
- MWE processing to enhance end-user applications: MWEs have gained particular attention in end-user applications, including MT, simplification, language learning and assessment, social media mining, and abusive language detection. We believe that it is crucial to extend and deepen these first attempts to integrate and evaluate MWE technology in these and further end-user applications.
- MWE identification and interpretation in pre-trained language models: Most current MWE processing is limited to their identification and detection using pre-trained language models, but we still lack understanding about how MWEs are represented and dealt with therein, how to better model the compositionality of MWEs from semantics. Now that NLP has shifted towards end-to-end neural models like BERT, capable of solving complex tasks with little or no intermediary linguistic symbols, questions arise about the extent to which MWEs should be implicitly or explicitly modeled.
- MWE processing in low-resource languages: The PARSEME shared tasks, among others, have fostered significant progress in MWE identification, providing datasets that include low-resource languages, evaluation measures, and tools that now allow fully integrating MWE identification into end-user applications. A few efforts have recently explored methods for the automatic interpretation of MWEs, and their processing in low-resource languages. Resource creation and sharing should be pursued in parallel with the development of methods able to capitalize on small datasets.

Pursuing the tradition of MWE Section of SIGLEX to foster future synergies with other communities to address scientific challenges in the creation of resources, models and applications to deal with MWEs, and in accordance with one of our special topics in MWE 2023 on specialized languages and domains, we are organizing a special track on "MWEs in Clinical NLP" as part of the MWE 2023 Workshop, collaborating with the Clinical NLP Workshop (colocated with ACL 2023).

We received 21 submissions of original research papers (10 long and 11 short) and selected 14 of them (7 long and 7 short), with an overall acceptance rate of 66.67% for the archival submissions. 9 of the accepted papers will be presented orally and 5 will be presented as posters. Two of the 14 accepted papers will be presented in the Special Track on MWEs in Clinical NLP. In addition to the archival submissions,

we also invited and accepted two non-archival submissions (published at other venues) for presentation (1 oral and 1 poster). The papers range from focus on (i) tasks such as identification or detection of MWEs, detection of idiomaticity, probing for idiomaticity, or measuring idiomaticity in the clinical domain, processing and comprehension of MWEs (experiments to measure human and computational processing), or comprehension of verbal MWEs; (ii) their evaluation through a survey of papers, e.g., on MWE identification focusing on their experimental designs; (iii) annotation or corpus development efforts, for example, annotations for lexical bundles used as discourse connectives, release of an annotated multilingual corpus of verbal MWEs and related recent developments of technical infrastructure for various languages, automatic generation of difficulty-graded vocabulary lists with MWEs graded based on their semantic compositionally, automated generation of pronunciation information for multiword terms in Wiktionary; (iv) methods to evaluate corpora, e.g., evaluating MWE lexicon formalisms based on observational adequacy; or (v) their applications, for example, studying effects of identifying MWEs on topic modeling, or development of a tool to enable complex queries over instances of verbal MWEs. The papers cover a large number of languages and a number of domains demonstrating the pervasiveness of MWEs and usefulness of research and synergistic efforts involving this area.

In addition to the oral and poster presentations of the accepted papers, the workshop features two keynote talks and a panel discussion with distinguished guests from the MWEs community and the Clinical NLP community. In the main session, Dr. Leo Wanner (ICREA and University Pompeu Fabra) will deliver a keynote talk titled 'Lexical collocations: Explored a lot, still a lot more to explore'. In the special track on MWEs in Clinical NLP, Dr. Asma Ben Abacha (Microsoft) and Dr. Goran Nenadic (University of Manchester) will deliver a keynote talk titled 'MWEs in ClinicalNLP and Healthcare Text Analytics'.

We are grateful to the keynote speakers and panelists for agreeing to share their experiences and insights, the members of the Program Committee for their thorough and timely reviews to help us select an excellent technical program, and all members of the organizing committee for the fruitful collaboration. Our thanks also go to the EACL 2023 organizers for their support, to SIGLEX for their endorsement, and to the Clinical NLP workshop organizers for their efforts and interest in collaborating with MWE 2023 to create synergies between the two communities. Finally, we thank all the authors for their valuable contributions to the workshop and to all the workshop participants for their interest in the event.

Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Shiva Taslimipoor

Organizing Committee

Program Chairs

Marcos Garcia, Universidade de Santiago de Compostela, Galiza (Spain)
Voula Giouli, Institute for Language and Speech Processing, ATHENA RC, Greece
Shiva Taslimipoor, The University of Cambridge, England
Lifeng Han, The University of Manchester, United Kingdom

Coordination and communication Chair

Voula Giouli, Institute for Language and Speech Processing, ATHENA RC, Greece

Publication Chair

Archna Bhatia, Institute for Human and Machine Cognition, USA

Publicity Chair

Kilian Evang, Heinrich Heine University, Germany

Program Committee

Program Committee Members

Iñaki Alegria, University of the Basque Country
Margarita Alonso-Ramos, Universidade da Coruña
Tim Baldwin, University of Melbourne
Verginica Barbu Mititelu, Romanian Academy
Chris Biemann, Universität Hamburg
Alexandra Birch, University of Edinburgh
Francis Bond, Palacký University
Claire Bonial, U.S. Army Research Laboratory
Tiberiu Boroș, Adobe
Jill Burstein, Educational Testing Service
Miriam Butt, Universität Konstanz
Marie Candito, Université Paris Cité
Fabienne Cap, Uppsala University
Marine Carpuat, University of Maryland
Helena Caseli, Federal University of Sao Carlos
Anastasia Christofidou, Academy of Athens
Ken Church, Baidu
Simon Clematide, University of Zürich
Matthieu Constant, Université de Lorraine
Paul Cook, University of New Brunswick
Silvio Cordeiro, Bloomin
Monika Czerepowicka, University of Warmia and Mazury
Béatrice Daille, Nantes University
Myriam de Lhonneux, University of Copenhagen
Koenraad Desmedt, University of Bergen
Mona Diab, George Washington University
Gaël Dias, University of Caen Basse-Normandie
Rafael Ehren, Heinrich Heine University Düsseldorf
Ismail El Maarouf, Adarga Ltd
Gülşen Eryiğit, Istanbul Technical University
Meghdad Farahmand, University of Geneva
Christiane Fellbaum, Princeton University
Joaquim Ferreira da Silva, New University of Lisbon
Teresa Flera, Uni Warsaw
Karën Fort, Sorbonne Université
Aggeliki Fotopoulou, Institute for Language and Speech Processing, ATHENA RC
Daniela Gierschek, Uni Luxembourg
Stefan Th. Gries, UC Santa Barbara & JLU Giessen
Bruno Guillaume, Université de Lorraine
Dhouha Hadjmed, University of Sfax
Chikara Hashimoto, Yahoo!Japan
Christopher Hidey, Columbia University
Rebecca Hwa, University of Pittsburgh
Uxoa Iñurrieta, University of the Basque Country
Laura Kallmeyer, Heinrich Heine University Düsseldorf
Diptesh Kanojia, Surrey Institute for People-Centred AI, University of Surrey

Elma Kerz, RWTH Aachen
Ekaterina Kochmar, University of Cambridge
Dimitrios Kokkinakis, University of Gothenburg
Ioannis Korkontzelos, Edge Hill University
Iztok Kosem, Jožef Stefan Institute
Cvetana Krstev, University of Belgrade
Tita Kyriakopoulou, University Paris-Est Marne-la-Vallee
Eric Laporte, Gustave Eiffel University
Qinyuan Li, Trinity College Dublin
Timm Lichte, University of Tübingen
Irina Lobzhanidze, Ilia State University
Teresa Lynn, Mohamed bin Zayed University of Artificial Intelligence
Gunn Inger Lyse Samdal, University of Bergen
Alfredo Maldonado, Trinity College Dublin
Stella Markantonatou, Institute for Language and Speech Processing, ATHENA RC
Yuji Matsumoto, RIKEN Center for Advanced Intelligence Project
John P. McCrae, National University of Ireland, Galway
Nurit Melnik, The Open University of Israel
Laura A. Michaelis, University of Colorado Boulder
Jelena Mitrović, University of Passau
Johanna Monti, "L'Orientale" University of Naples
Preslav Nakov, Qatar Computing Research Institute, HBKU
Stella Neumann, RWTH Aachen
Sanni Nimb, Det Denske Sprog- og Litteraturselskab
Malvina Nissim, University of Groningen
Joakim Nivre, Uppsala University
Diarmuid Ó Séaghdha, University of Cambridge
Jan Odijk, University of Utrecht
Petya Osenova, Bulgarian Academy of Sciences
Yagmur Ozturk, Grenoble Alpes University
Martha Palmer, University of Colorado Boulder
Pan Pan, School of Foreign Studies, South China Normal University
Haris Papageorgiou, Institute for Language and Speech Processing
Yannick Parmentier, University of Lorraine
Carla Parra Escartín, Iconic Translation Machines
Caroline Pasquer, University of Tours
Agnieszka Patejuk, University of Oxford and Institute of Computer Science, Polish Academy of Sciences
Marie-Sophie Pausé, Independent researcher
Pavel Pecina, Charles University
Ted Pedersen, University of Minnesota
Miriam R.L. Petruck, International Computer Science Institute
Scott Piao, Lancaster University
Maciej Piasecki, Wroclaw University of Technology
Prisca Piccirilli, Uni. Stuttgart
Alain Polguère, Université de Lorraine
Vinodkumar Prabhakaran, Google
Behrang QuasemiZadeh, University of Duesseldorf
Alexandre Rademaker, IBM Research Brazil and EMAp/FGV
Carlos Ramisch, Aix Marseille University
Sonia Ramotowska, Uni Amsterdam

Livy Real, americanas s.a.
Martin Riedl, University of Hamburg
Matīss Rikters, University of Tokyo
Victoria Rosén, University of Bergen
Mike Rosner, University of Malta
Fatiha Sadat, Université du Québec à Montréal
Manfred Sailer, Goethe-Universität Frankfurt am Main
Bahar Salehi, The University of Melbourne
Magali Sanches Duran, University of São Paulo
Federico Sangati, Independent researcher
Agata Savary, Université Paris-Saclay
Nathan Schneider, Georgetown University
Sabine Schulte im Walde, University of Stuttgart
Matthew Shardlow, Manchester Metropolitan University
Vered Shwartz, Allen AI
Kiril Simov, Bulgarian Academy of Sciences
Noah Smith, University of Washington
Gyri Smørdal Losnegaard, University of Bergen
Jan Šnajder, University of Zagreb
Ranka Stanković, University of Belgrade
Ivelina Stoyanova, Bulgarian Academy of Sciences
Pavel Straňák, Charles University
Stan Szpakowicz, University of Ottawa
Harish Tayyar Madabushi, University of Bath
Carole Tiberius, Dutch Language Institute
Beata Trawinski, Leibniz Institute for the German Language
Yulia Tsvetkov, Carnegie Mellon University
Zdeňka Urešová, Charles University
Ruben Urizar, University of the Basque Country
Ashwini Vaidya, Indian Institute of Technology
Lonneke van der Plas, University of Malta
Bertram Vidgen, Alan Turing Institute
Aline Villavicencio, University of Sheffield
Veronika Vincze, Hungarian Academy of Sciences
Martin Volk, University of Zürich
Zeerak Talat, Simon Fraser University
Jakub Waszczuk, University of Duesseldorf
Eric Wehrli, University of Geneva
Marion Weller-Di Marco, Ludwig Maximilian University of Munich
Seid Muhie Yimam, Universität Hamburg

Keynote Speakers

Leo Wanner, ICREA and Universitat Pompeu Fabra
Asma Ben Abacha, Microsoft
Goran Nenadic, University of Manchester

Keynote Talk: Lexical collocations: Explored a lot, still a lot more to explore

Leo Wanner
ICREA and Universitat Pompeu Fabra
2023-05-06 –

Abstract: Lexical collocations: Explored a lot, still a lot more to explore
Lexical collocations, i.e., idiosyncratic binary lexical item combinations, have been an active research topic already for a number of years. State-of-the-art neural network models report to detect and classify specific types of lexical collocations with high accuracy, which might suggest that the problem has been solved. However, a cross-type and cross-language analysis of the results of one of these models raises several relevant research questions. In the first part of my talk, I will present our recent work on the identification and classification of lexical collocations with respect to the fine-grained taxonomy of lexical functions (LFs) in English, French, Spanish and Japanese. Drawing on the outcome of this work, I will focus, in the second part of my talk, on the comparative analysis of the "LF profiles" of English and Japanese material. In particular, I will discuss (i) how the considered LFs are distributed in the given corpora; (ii) how rich the repertoires of the LF instances are in each of them; (iii) whether the contexts of the LF instances overlap; and (iv) to what extent the "profile" of an LF correlates with the accuracy of the recognition of its instances. To conclude, I will formulate the research questions that arise from this analysis.

Bio: Dr. Leo Wanner, ICREA and Universitat Pompeu Fabra
Leo Wanner is ICREA Research Professor at the Pompeu Fabra University in Barcelona, with 230+ peer reviewed publications and 10 edited volumes. He is Associate Editor of the Computational Intelligence and Frontiers in AI, Language and Computation journals and serves as regular reviewer for a number of high-profile conferences and journals on Computational Linguistics. Throughout his career, Leo worked on a number of topics in the field, including natural language generation and summarization, concept extraction, conversational agents, hate speech recognition, and, in particular, also lexical collocation identification and classification.

Keynote Talk: MWEs in ClinicalNLP and Healthcare Text Analytics

Asma Ben Abacha and Goran Nenadic
Microsoft and University of Manchester (respectively)
2023-05-06 –

Abstract: MWEs in ClinicalNLP and Healthcare Text Analytics
MWEs are a common phenomenon in the clinical domain: for example, diagnoses and clinical findings are often expressed using complex, compositional multi-word expressions that contain references to a disease, its anatomy, laterality, severity, temporality etc. This applies both to the 'formal' clinical language (e.g. in clinical letters, clinical terminologies) and spoken or written healthcare discussions (e.g. patient-doctor conversations, healthcare social media). Despite advances in language modelling, the extraction and disentangling of clinical MWEs are still challenging tasks. In this talk, we will first look at the structure of multi-word disease descriptions in clinical letters, and discuss the challenges in mapping such free-text mentions to standard clinical vocabularies. We will then discuss how MWE extraction could be evaluated using various automatic evaluation metrics. We will compare several evaluation methods and metrics, and explore the correlation between automatic metrics and manual judgments, in particular in the context of the summarization of doctor-patient conversations and generation of clinical notes.

Bio: Dr. Asma Ben Abacha (Microsoft) and Dr. Goran Nenadic (University of Manchester)
Asma Abacha is a Senior Scientist at Microsoft, with over 80 peer reviewed publications. Her research interests include Natural Language Processing, Machine Learning, Artificial Intelligence and their applications in medicine and healthcare.
Goran Nenadic is a Professor in the Department of Computer Science at University of Manchester and a Turing Fellow at the Alan Turing Institute, with more than 250 peer reviewed publications. His research interests include Natural Language Processing, text mining, and health informatics.

Table of Contents

Program

Saturday, May 6, 2023

08:30 - 09:00 *Registration*

09:00 - 09:10 *Opening*

09:10 - 10:30 *Oral long paper presentations*

Are Frequent Phrases Directly Retrieved like Idioms? An Investigation with Self-Paced Reading and Language Models
Giulia Rambelli, Emmanuele Chersoni, Marco S. G. Senaldi, Philippe Blache and Alessandro Lenci

A Survey of MWE Identification Experiments: The Devil is in the Details
Carlos Ramisch, Abigail Walsh, Thomas Blanchard and Shiva Taslimipoor

The Better Your Syntax, the Better Your Semantics? Probing Pretrained Language Models for the English Comparative Correlative
Leonie Weissweiler, Valentin Hofmann, Abdullatif Koksal and Hinrich Schütze

Predicting Compositionality of Verbal Multiword Expressions in Persian
Mahtab Sarlak, Yalda Yarandi and Mehrnoush Shamsfard

10:30 - 11:15 *Morning coffee break*

11:15 - 12:15 *Keynote Talk, Leo Wanner: Lexical collocations: Explored a lot, still a lot more to explore*

12:15 - 12:50 *Oral paper presentations*

Romanian Multiword Expression Detection Using Multilingual Adversarial Training and Lateral Inhibition
Andrei Avram, Verginica Barbu Mititelu and Dumitru-Clementin Cercel

PARSEME corpus release 1.3
Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze and Abigail Walsh

12:50 - 14:15 *Lunch Break*

Saturday, May 6, 2023 (continued)

14:15 - 14:45 *Keynote Talk, Asma Abacha and Goran Nenadic: MWEs in ClinicalNLP and Healthcare Text Analytics*

14:45 - 15:15 *Oral short paper presentations*

Detecting Idiomatic Multiword Expressions in Clinical Terminology using Definition-Based Representation Learning
François Remy, Alfiya Khabibullina and Thomas Demeester

Investigating the Effects of MWE Identification in Structural Topic Modelling
Dimitrios Kokkinakis, Ricardo Sánchez, Sebastianus Bruinsma and Mia-Marie Hammarlin

15:15 - 15:45 *Panel discussion: Multiword Expressions in Knowledge-intensive Domains: Clinical Text as a Case Study*

15:45 - 16:30 *Afternoon coffee break*

16:30 - 17:15 *Poster session*

Idioms, Probing and Dangerous Things: Towards Structural Probing for Idiomaticity in Vector Space
Filip Klubička, Vasudevan Nedumpozhimana and John Kelleher

Simple and Effective Multi-Token Completion from Masked Language Models
Oren Kalinsky, Guy Kushilevitz, Alexander Libov and Yoav Goldberg

Annotation of lexical bundles with discourse functions in a Spanish academic corpus
Eleonora Guzzi, Margarita Alonso-Ramos, Marcos Garcia and Marcos García Salido

Enriching Multiword Terms in Wiktionary with Pronunciation Information
Lenka Bajcetic, Thierry Declerck and Gilles Sérasset

Automatic Generation of Vocabulary Lists with Multiword Expressions
John Lee and Adilet Uvaliyev

A MWE lexicon formalism optimised for observational adequacy
Adam Lion-Bouton, Agata Savary and Jean-Yves Antoine

Saturday, May 6, 2023 (continued)

17:15 - 17:45 *Oral short paper presentations*

Token-level Identification of Multiword Expressions using Pre-trained Multilingual Language Models
Raghuraman Swaminathan and Paul Cook

Graph-based multi-layer querying in Parseme Corpora
Bruno Guillaume

17:45 - 18:00 *Closing*