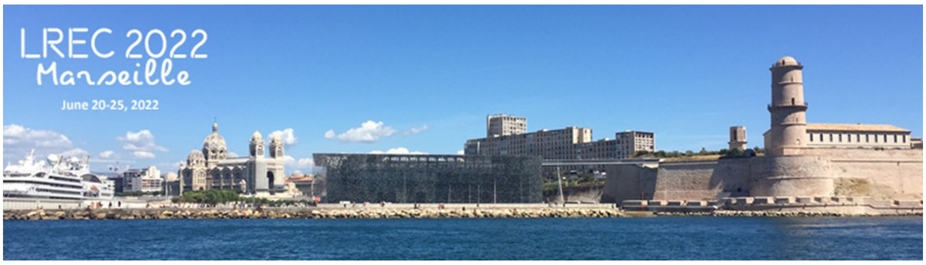

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

# 18th Workshop on Multiword Expressions
# MWE 2022

# PROCEEDINGS

Editors:
Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia and Carlos Ramisch

# Proceedings of the LREC 2022 workshop on 18th Workshop on Multiword Expressions (MWE 2022)

Edited by: Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia and Carlos Ramisch

**ISBN: 979-10-95546-90-0**
**EAN: 9791095546900**

**For more information:**
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

# Introduction

The 18th Workshop on Multiword Expressions (MWE 2022)[1] took place on a hybrid (on-site/remote) format on June 25, 2022 in Marseille (France), in conjunction with the 13th Edition of the Language Resources and Evaluation Conference (LREC 2022). MWE 2022 was organized and sponsored by the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL).

Multiword expressions (MWEs) are word combinations which exhibit lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies, such as by and large, hot dog, pay a visit and pull one's leg. The notion encompasses closely related phenomena: idioms, compounds, light-verb constructions, phrasal verbs, rhetorical figures, collocations, institutionalised phrases, etc. Their behaviour is often unpredictable; for example, their meaning often does not result from the direct combination of the meanings of their parts. Given their irregular nature, MWEs often pose complex problems in linguistic modelling (e.g. annotation), NLP tasks (e.g. parsing), and end-user applications (e.g. natural language understanding and MT), hence still representing an open issue for computational linguistics.

For almost two decades, modelling and processing MWEs for NLP has been the topic of the MWE workshop organised by the MWE section of SIGLEX in conjunction with major NLP conferences since 2003. Impressive progress has been made in the field, but our understanding of MWEs still requires much research considering its need and usefulness in NLP applications. For this 18th edition of the workshop, we identified three topics on which contributions are particularly encouraged:

- 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 automatic interpretation of MWEs. Pursuing similar efforts on understanding MWEs in low-resource languages is beneficial. there are some recent efforts on processing of MWEs in low-resource languages. Resource creation and sharing should be pursued in parallel to the development of methods able to capitalize on small datasets.
- 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 lack understanding about how MWEs are represented and dealt with therein. Now that NLP has shifted towards end-to-end neural models like BERT, capable of solving complex end-user tasks with little or no intermediary linguistic symbols, questions arise about the extent to which MWEs should be implicitly or explicitly modelled in such models.
- MWE processing to enhance end-user applications: As underlined by the MWE 2021 call for papers, MWEs 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.

We received 23 submissions of original research papers (12 long and 11 short). We selected 15 papers (9 long and 6 short), 10 presented orally and 5 as posters. The overall acceptance rate was 65%. As a novelty in this edition, we also called for non-archival submissions of abstracts (describing preliminary results, work in progress, or abstract of papers recently submitted or published at other venues), considered for

[1] https://multiword.org/mwe2022/

presentation but not included in the proceedings. We received 7 non-archival submission, from which we selected 5 for presentation.

Moreover, we organised a joint session with the workshop of the Special Interest Group on Under-resourced Languages, SIGUL 2022, to foster future synergies that could address scientific challenges in the creation of resources, models and applications to deal with multiword expressions and related phenomena in low-resource scenarios, in accordance with one of our special topics in MWE 2022.

In addition to the oral and poster sessions, the workshop featured two invited talks, given by Sabine Schulte im Walde (University of Stuttgart, Germany) and by Steven Bird (Charles Darwin University, Australia).

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, and to all the workshop participants for their interest in this event. Our thanks also go to the LREC 2022 organizers for their support, to SIGLEX for their endorsement, and to SIGUL for their efforts and interest in organising the MWE-SIGUL joint session.

*Archna Bhatia, Paul Cook, Shiva Taslimipoor, Marcos Garcia, Carlos Ramisch*

## Organizers

### Program Chairs

Archna Bhatia – Florida Institute for Human & Machine Cognition
Paul Cook – University of New Brunswick – Faculty of Computer Science
Shiva Taslimipoor – University of Cambridge – NLIP Group

### Publication Chair

Marcos Garcia – Universidade de Santiago de Compostela – CiTIUS Research Centre

### Communication Chair

Carlos Ramisch – Aix Marseille University – TALEP Research Group

## Program Committee

Tim Baldwin, University of Melbourne (Australia)
Verginica Barbu Mititelu, Romanian Academy (Romania)
Francis Bond, Palacký University (Czech Republic)
Claire Bonial, U.S. Army Research Laboratory (USA)
Tiberiu Boroș, Adobe (Romania)
Marie Candito, Université Paris Cité (France)
Anastasia Christofidou, Academy of Athens (Greece)
Ken Church, Baidu (USA)
Matthieu Constant, Université de Lorraine (France)
Monika Czerepowicka, University of Warmia and Mazury (Poland)
Myriam de Lhonneux, University of Copenhagen (Denmark)
Gaël Dias, University of Caen Basse-Normandie (France)
Gülşen Eryiğit, Istanbul Technical University (Turkey)
Meghdad Farahmand, University of Geneva (Switzerland)
Christiane Fellbaum, Princeton University (USA)
Joaquim Ferreira da Silva, New University of Lisbon (Portugal)
Aggeliki Fotopoulou, Institute for Language and Speech Processing/RC "Athena" (Greece)
Voula Giouli, Institute for Language and Speech Processing (Greece)
Stefan Th. Gries, UC Santa Barbara (USA) & JLU Giessen (Germany)
Uxoa Iñurrieta, University of the Basque Country (Spain)
Diptesh Kanojia, Surrey Institute for People-Centred AI, University of Surrey (UK)
Ioannis Korkontzelos, Edge Hill University (UK)
Cvetana Krstev, University of Belgrade (Serbia)
Eric Laporte, Gustave Eiffel University (France)
Timm Lichte, University of Tübingen (Germany)
Irina Lobzhanidze, Ilia State University (Georgia)
Teresa Lynn, ADAPT Centre (Ireland)
Gunn Inger Lyse Samdal, University of Bergen (Norway)
Stella Markantonatou, Institute for Language and Speech Processing (Greece)
Yuji Matsumoto, RIKEN Center for Advanced Intelligence Project (Japan)

Jan Odijk, University of Utrecht (Netherlands)
Haris Papageorgiou, Institute for Language and Speech Processing (Greece)
Yannick Parmentier, Université d'Orléans (France)
Pavel Pecina, Charles University (Czech Republic)
Ted Pedersen, University of Minnesota (USA)
Scott Piao, Lancaster University (UK)
Alain Polguère, Université de Lorraine (France)
Livy Real, americanas s.a. (Brazil)
Fatiha Sadat, Université du Québec à Montréal (Canada)
Magali Sanches Duran, University of São Paulo (Brazil)
Sabine Schulte im Walde, University of Stuttgart (Germany)
Matthew Shardlow, Manchester Metropolitan University (UK)
Ivelina Stoyanova, Bulgarian Academy of Sciences (Bulgaria)
Pavel Straňák, Charles University (Czech Republic)
Stan Szpakowicz, University of Ottawa (Canada)
Carole Tiberius, Dutch Language Institute (Netherlands)
Beata Trawinski, Leibniz Institute for the German Language (Germany)
Zdeňka Urešová, Charles University (Czech Republic)
Ruben Urizar, University of the Basque Country (Spain)
Lonneke van der Plas, University of Malta (Malta)
Veronika Vincze, Hungarian Academy of Sciences (Hungary)
Martin Volk, University of Zürich (Switzerland)
Zeerak Talat, Digital Democracies Institute, Simon Fraser University (Canada)
Marion Weller-Di Marco, Ludwig Maximilian University of Munich (Germany)
Jelena Mitrović, University of Passau (Germany)
Petya Osenova , Bulgarian Academy of Sciences (Bulgaria)
Ashwini Vaidya, Indian Institute of Technology, Delhi (India)

# Table of Contents

# Conference Program

***9:00–9:10*** ***Opening***

**9:10–10:30** **Session 1: Oral presentations**

9:10–9:25 *A General Framework for Detecting Metaphorical Collocations*
Marija Brkić Bakarić, Lucia Načinović Prskalo and Maja Popović

9:25–9:40 *Improving Grammatical Error Correction for Multiword Expressions*
Shiva Taslimipoor, Christopher Bryant and Zheng Yuan

9:40–9:50 *Native and Non-native Speakers' Idiom Production: What Can Read Speech Tell Us?* (non-archival paper)
Jing Liu and Helmer Strik

9:50–10:10 *An Analysis of Attention in German Verbal Idiom Disambiguation*
Rafael Ehren, Laura Kallmeyer and Timm Lichte

10:10–10:30 *Support Verb Constructions across the Ocean Sea*
Jorge Baptista, Nuno Mamede and Sónia Reis

*10:30–11:00* *Coffee break*

**11:00–12:00** **Session 2: Invited Talk #1**

*Figurative Language in Noun Compound Models across Target Properties, Domains and Time*
Sabine Schulte im Walde

**12:00–13:00** **Session 3: Oral presentations**

12:00–12:20 *A Matrix-Based Heuristic Algorithm for Extracting Multiword Expressions from a Corpus*
Orhan Bilgin

12:20–12:40 *Multi-word Lexical Units Recognition in WordNet*
Marek Maziarz, Ewa Rudnicka and Łukasz Grabowski

12:40–13:00 *Automatic Detection of Difficulty of French Medical Sequences in Context*
Anaïs Koptient and Natalia Grabar

*13:00–14:00* *Lunch break*

**14:00–15:00** **Session 4: Poster Session (joint with SIGUL)**

*Annotating "Particles" in Multiword Expressions in te reo Māori for a Part-of-Speech Tagger*
Aoife Finn, Suzanne Duncan, Peter-Lucas Jones, Gianna Leoni and Keoni Mahelona

*Metaphor Detection for Low Resource Languages: From Zero-Shot to Few-Shot Learning in Middle High German*
Felix Schneider, Sven Sickert, Phillip Brandes, Sophie Marshall and Joachim Denzler

*Automatic Bilingual Phrase Dictionary Construction from GIZA++ Output*
Albina Khusainova, Vitaly Romanov and Adil Khan

*A BERT's Eye View: Identification of Irish Multiword Expressions Using Pre-trained Language Models*
Abigail Walsh, Teresa Lynn and Jennifer Foster

*Enhancing the PARSEME Turkish Corpus of Verbal Multiword Expressions*
Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton and Agata Savary

*German Light Verb Constructions in Business Process Models* (non-archival paper)
Kristin Kutzner and Ralf Laue

*BPE beyond Word Boundary: How NOT to Use Multi Word Expressions in Neural Machine Translation* (non-archival paper)
Avijit Thawani and Dipesh Kumar

**15:00–16:00 Session 5: Invited Talk #2 (joint with SIGUL)**

*Multiword Expressions and the Low-Resource Scenario from the Perspective of a Local Oral Culture*
Steven Bird

*16:00–16:30 Coffee break*

**16:30–17:40 Session 6: Oral Presentations**

16:30–16:40 *Compound-internal Anaphora: Evidence from Acceptability Judgements on Italian Argumental Compounds* (non-archival paper)
Irene Lami and Joost van de Weijer

16:40–16:50 *Light Verb Constructions in Corpora of Historical English* (non-archival paper)
Eva Zehentner

16:50–17:05 *Sample Efficient Approaches for Idiomaticity Detection*
Dylan Phelps, Xuan-Rui Fan, Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton and Aline Villavicencio

17:05–17:20 *mwetoolkit-lib: Adaptation of the mwetoolkit as a Python Library and an Application to MWE-based Document Clustering*
Fernando Rezende Zagatti, Paulo Augusto de Lima Medeiros, Esther da Cunha Soares, Lucas Nildaimon dos Santos Silva, Carlos Ramisch and Livy Real

17:20–17:40 *Handling Idioms in Symbolic Multilingual Natural Language Generation*
Michaelle Dubé and François Lareau

**17:40–18:00 MWE Community Discussion**