MWE 2025

The 21st Workshop on Multiword Expressions (MWE 2025)

Proceedings of the Workshop

May 4, 2025

The MWE organizers gratefully acknowledge the support from the following organizations.

Gold

SIGLEX
ACL Special Interest Group

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 979-8-89176-243-5

Introduction

The 21st Workshop on Multiword Expressions (MWE 2025) took place on May 4, 2025, in Albuquerque, New Mexico, USA, and online, as a satellite event of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2025). The workshop was organized and sponsored by the Special Interest Group on the Lexicon (SIGLEX) (http://www.siglex.org) of the Association for Computational Linguistics (ACL)(https://www.aclweb.org/portal/).

The notion of multiword expressions (MWEs), i.e., word combinations that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies, encompasses closely related phenomena: idioms, compounds, light-verb constructions, phrasal verbs, rhetorical figures, collocations, institutionalized phrases, etc. Given their irregular nature, MWEs often pose complex problems in linguistic modeling (e.g. annotation), NLP tasks (e.g. parsing), and end-user applications (e.g. natural language understanding and Machine Translation), hence still representing an open issue for computational linguistics.

For this 21st edition of the workshop, our call for papers focused particularly on the following topics:

- MWE processing to enhance end-user applications;
- MWE processing and identification in the general language, as well as in specialized languages and domains;
- MWE processing in low-resourced languages;
- MWE identification and interpretation in LLMs
- new and enhanced representation of MWEs in language resources and computational models of compositionality as gold standards for formative intrinsic evaluation.

For this edition, all submitted papers were peer-reviewed by international experts and 75% of the submitted papers were accepted. Barbu Mititelu et al. paints the current state of the art of MWE lexica designed for NLP purposes. A diachronic perspective is adopted by Alves et al. when investigating the syntagmatic productivity of MWEs in English scientific writing.

The interest in endangered and low-resourced languages is still visible in the papers that report the development of new resources, dedicated to such languages. Thus, the paper authored by Adkins et al. focuses on Irish and on the recognition of named entities in this language, for which a tool is developed, while also producing a small gold-standard corpus annotated with named entities. Galician is the language for which a dataset of 240 ambiguous noun-adjective MWEs, contextualized in two sets of sentences, is manually rated for compositionality at token level, being also added information about frequency, ambiguity, and productivity. Markantonatou et al. propose the first Standard Modern Greek Universal Dependencies treebank annotated with Verbal MWEs, while also using it to evaluate the performance of models in MWEs identification tasks. A new resource for European Portuguese, namely a corpus annotated for verbal idioms, is reported by Antunes et al.

The development of multilingual resources is also an area of research represented in this workshop: Sentsova et al. introduce MultiCoPIE, a multilingual corpus of potentially idiomatic expressions in Catalan, Italian, and Russian, as well as the cross-lingual transfer of the potentially idiomatic expressions disambiguation task from English to the three languages in this new resource.

LLMs are found in several tasks: Kissane et al. examine how LLMs capture lexical and syntactic properties of phrasal verbs and prepositional verbs at different neural network layers. Adkins et al. compare both monolingual and multilingual BERT models fine-tuned on named entity recognition task. LLMs are also used to generate synthetic idiom datasets and to evaluate their effectiveness in training task-specific models for idiomaticity detection.

Verginica Barbu Mititelu, Mathieu Constant, A. Seza Doğruöz, Voula Giouli, Gražina Korvel, Atul Kr. Ojha, Alexandre Rademaker (MWE-2025 Organizers and Co-Chairs)

Organizing Committee

Workshop Chairs

Mathieu Constant, Université de Lorraine, CNRS, ATILF
A. Seza Doğruöz, Ghent University
Voula Giouli, Aristotle University of Thessaloniki and ILSP - "Athena" Research Center
Gražina Korvel, Vilnius University
Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence
Atul Kr. Ojha, Insight Research Ireland Centre for Data Analytics, DSI, University of Galway, Ireland and Panlingua Languague Processing LLP, India
Alexandre Rademaker, School of Applied Mathematics of Getulio Vargas Foundation

Program Committee

Agata Savary, Université Paris-Saclay
Beata Trawinski, Leibniz Institute for the German Language
Carlos Ramisch, LIS - Laboratoire d'Informatique et Systèmes
Chikara Hashimoto, Rakuten Institute of Technology
Cvetana Krstev, University of Belgrade, Faculty of Philology
Eric G C Laporte, Université Gustave Eiffel
Francis Bond, Palacký University Olomouc
Gaël Dias, University of Caen Normandy
Gražina Korvel, Vilnius University
Irina Lobzhanidze, Ilia Chavchavadze State University
Ismail El Maarouf, Imprevicible
Ivelina Stoyanova, Deaf Studies Institute
Jan Odijk, Utrecht University
John Philip McCrae, University of Galway
Kenneth Church, Northeastern University
Manfred Sailer, Johann Wolfgang Goethe Universität Frankfurt am Main
Mathieu Constant, Université de Lorraine, CNRS, ATILF
Matthew Shardlow, The Manchester Metropolitan University
Meghdad Farahmand, University of Genoa
Miriam Butt, Universität Konstanz
Paul Cook, University of New Brunswick
Pavel Pecina, Charles University
Petya Osenova, Sofia University St. Kliment Ohridski
Ranka Stanković Stanković, University of Belgrade
Sabine Schulte im Walde, University of Stuttgart
Shiva Taslimipoor, University of Cambridge
Stan Szpakowicz, University of Ottawa
Stella Markantonatou, ATHENA RIC
Tiberiu Boros, Adobe Systems
Tunga Gungor, Bogazici University

Keynote Talk: Meaning Construction at the Syntax-Lexis Nexus

Nathan Schneider
Associate Professor of Linguistics and Computer Science at Georgetown University (USA)

Abstract: When words and grammar come into contact, things sometimes get messy: idiosyncratic expressions and patterns disobey ordinary principles of regularity and compositionality. A useful point of reference is the theoretical perspective of Construction Grammar, which exhorts us to view linguistic knowledge in terms of form-function mappings—at all levels of granularity. How can this perspective inform a broad-coverage, multilingual approach to lexicosyntactic conundrums? First, I will discuss implications for corpus annotation: while some multiword expressions and names (e.g. "at least", "in order to", "Chapter 1") test the limits of categorical annotation standards like Universal Dependencies, UD treebanks nevertheless enable empirical investigation of some functionally-defined constructions across languages. Second, I will discuss efforts to interpret the latent representations of constructional form and meaning in transformer language models, with the NPN construction (noun-preposition-noun, as in "face to face") as a case study.

Bio: Nathan Schneider is a computational linguist. As Associate Professor of Linguistics and Computer Science at Georgetown University, he leads the NERT lab, looking for synergies between practical language technologies and the scientific study of language, with an emphasis on how words, grammar, and context conspire to convey meaning. He is the recipient of an NSF CAREER award to study NLP vis-à-vis metalinguistic enterprises like language learning, linguistics, and legal interpretation. Recently, he has weighed in on specific interpretive debates in U.S. law; one of these analyses was cited by U.S. Supreme Court justices in a major firearms case. He is active in the NLP community—especially ACL's SIGANN and SIGLEX—and the Universal Dependencies project; and cofounded the SOLID forum for empirical research on legal interpretation. Prior to Georgetown, he inhabited UC Berkeley, Carnegie Mellon University, and the University of Edinburgh. Apart from annotation scheming and computational modeling, he enjoys classical music and chocolate chip cookies.

Table of Contents

Program

Sunday, May 4, 2025

09:15 - 09:30	*Welcome and Introduction to 21st MWE Workshop*
09:30 - 10:30	*Invited Talk*
10:30 - 11:00	*Coffee Break*
11:00 - 12:30	*Oral Session-I*
	Survey on Lexical Resources Focused on Multiword Expressions for the Purposes of NLP Verginica Mititelu, Voula Giouli, Gražina Korvel, Chaya Liebeskind, Irina Lobzhanidze, Rusudan Makhachashvili, Stella Markantonatou, Aleksandra Markovic and Ivelina Stoyanova
	Named Entity Recognition for the Irish Language Jane Adkins, Hugo Collins, Joachim Wagner, Abigail Walsh and Brian Davis
	Gathering Compositionality Ratings of Ambiguous Noun-Adjective Multiword Expressions in Galician Laura Castro and Marcos Garcia
	VMWE identification with models trained on GUD (a UDv.2 treebank of Standard Modern Greek) Stella Markantonatou, Vivian Stamou, Stavros Bompolas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos and Antonios Anastasopoulos
12:30 - 14:00	*Lunch*
14:00 - 15:00	*Oral Session-II*
	A European Portuguese corpus annotated for verbal idioms David Antunes, Jorge Baptista and Nuno J. Mamede
	MultiCoPIE: A Multilingual Corpus of Potentially Idiomatic Expressions for Cross-lingual PIE Disambiguation Uliana Sentsova, Debora Ciminari, Josef Van Genabith and Cristina España-Bonet
	Probing Internal Representations of Multi-Word Verbs in Large Language Models Hassane Kissane, Achim Schilling and Patrick Krauss

Sunday, May 4, 2025 (continued)

	Syntagmatic Productivity of MWEs in Scientific English Diego Alves, Stefan Fischer and Elke Teich
15:30 - 16:00	*Coffee Break*
16:00 - 18:00	*Oral Session III, panel and community discussion*
	Using LLMs to Advance Idiom Corpus Construction Doğukan Arslan, Hüseyin Anıl Çakmak, Gulsen Eryigit and Joakim Nivre
16:21 - 17:20	*Panel: Tokenization in the era of LLMs*
17:21 - 17:50	*Community discussion*
17:51 - 18:00	*Best Paper Awards and Concluding Remarks*