LREC-COLING 2024

Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)

Workshop Proceedings

Editors
Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre and Alexandre Rademacher

25 May, 2024
Torino, Italia

Proceedings of the LREC-COLING 2024 Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)

ISBN 978-2-493814-20-3
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

Multiword expressions (MWEs) are word combinations that exhibit lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies (Baldwin and Kim, 2010), such as *by and large, hot dog, pay a visit* and *pull someone's leg*. The notion encompasses closely related phenomena: idioms, compounds, light-verb constructions, phrasal verbs, rhetorical figures, collocations, institutionalized phrases, *etc*. Their behavior 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 modeling (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 (Constant et al., 2017). This joint workshop also marks the 20th anniversary of the MWE workshop series since 2003 (Bond et al., 2003). The organization of the workshops is sponsored by SIGLEX.[1]

Universal Dependencies (UD; De Marneffe et al., 2021) is a framework for cross-linguistically consistent treebank annotation that has so far been applied to over 100 languages. The framework aims to capture similarities as well as idiosyncrasies among typologically different languages (*e.g.*, morphologically rich languages, pro-drop languages, and languages featuring clitic doubling). The goal of developing UD was not only to support comparative evaluation and cross-lingual learning but also to facilitate multilingual natural language processing and enable comparative linguistic studies. Starting with the first UD workshop in 2017 (de Marneffe et al., 2017), this joint workshop is the 7th edition in the series.

For the current edition, the MWE and UD communities decided to organize a joint event, the MWE-UD workshop which is part of LREC-COLING 2024. This is a timely collaboration because the two communities clearly have overlapping interests. For instance, while UD has several dependency relations that are intended for annotation of MWEs, both annotation guidelines (i.e. is *syntactic irregularity and inflexibility* or *semantic non-compositionality* the leading criterion?) and annotation practice (both across treebanks for a single language and across languages) for MWEs can be improved (Schneider and Zeldes, 2021). For verbal MWEs, the PARSEME corpora for 26 languages provide annotation of MWEs consistent with UD annotation (Savary et al., 2023). Both communities share an interest in developing guidelines, data-sets, and tools that can be applied to a wide range of typologically diverse languages, raising fundamental questions about tokenization, lemmatization, and morphological decomposition of tokens. Proposals for harmonizing annotation practice between what has been achieved in PARSEME and UD and expanding PARSEME annotation to non-verbal MWEs are also central to the recently started UniDive[2] COST action (CA21167). UniDive also co-organizes and sponsors this joint workshop.

We are happy to have received 53 submissions, 29 long, 15 short, and 9 non-archival. 19 long, 7 short, and 9 non-archival contributions were selected for presentation at the workshop, bringing the overall acceptance rate for archival papers to 59%. The distribution over tracks is almost even: 8 of 12 archival submissions were accepted for the UD track, 9 of 16 for the MWE track, and 9 of 16 for the MWE+UD track. One long paper was withdrawn after acceptance.

An important theme in both the UD and MWE community is increasing the number of languages and language families that can be used as the object of study, for instance by making annotated data available in a standard format. The current workshop makes a substantial contribution towards that goal, as it includes contributions to Arabic, Hindi, Old Egyptian, Vedic, Northern Kurdish, Slovene, Dutch, Bavarian, South Asian languages, Turkic languages, Gujarati, Saraiki, Swedish, and more. Another important theme for research on MWEs has been the question

[1] https://siglex.org/

[2] https://unidive.lisn.upsaclay.fr

of to what extent Large Language Models deal adequately with the idiomatic meaning of multi-word expressions. This workshop also includes several contributions that explicitly deal with this question. Apart from these important and cross-disciplinary themes, there are also contributions on UD addressing such issues as assessing and enhancing the value of UD parsing for applications, improved automatic parsing procedures, and the interface between syntax and morphology. Contributions that are primarily concerned with MWEs address a.o. the role of lexical resources, automatic identification of MWEs, the proper annotation of idiomatic meanings in a corpus with fully structured meaning annotation, annotation in parallel corpora, and cross-lingually consistent annotation of MWEs with word senses. Some of these themes re-occur in the contributions that address both UD and MWEs, such as the interplay of lexicon and corpus annotations, the annotation of multiword functional categories, the annotation of light verb constructions, and the use of UD and MWEs in the task of stance detection.

Archna Bhatia, Gosse Bouma, A. Seza Doğruöz, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han, Joakim Nivre, Alexandre Rademacher.

Acknowledgements

MWE-UD Workshop 2024 has been co-organised and funded by The Special Interest Group for the Lexicon of the Association for Computational Linguistics (ACL-SIGLEX) and the CA21167 COST Action UniDive, supported by COST (European Cooperation in Science and Technology). ACL-SIGLEX funded two (2) participants, while UniDive provided funds for the travel and stay of 29 participants.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors. 2003. *Proceed- ings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.* Association for Computational Linguistics, Sapporo, Japan.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Marie-Catherine de Marneffe, Joakim Nivre, and Sebastian Schuster, editors. 2017. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. Parseme meets universal dependencies: getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9(1).

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria. Association for Computational Linguistics.

Organizing Committee

Workshop Organizers

Archna Bhatia, Institute for Human and Machine Cognition, USA
Gosse Bouma, Groningen University, Netherlands
A. Seza Doğruöz, Ghent University, Belgium
Kilian Evang, Heinrich Heine University Düsseldorf, Deutschland
Marcos Garcia, University of Santiago de Compostela, Galiza, Spain
Voula Giouli, Institute for Language & Speech Processing, ATHENA RC, Greece
Lifeng Han, University of Manchester, United Kingdom
Joakim Nivre, Uppsala University and Research Institutes of Sweden, Sweden
Alexandre Rademaker, IBM Research, Brazil

Program Committee

Verginica Barbu Mititelu, Cherifa Ben Kehlil, Philippe Blache, Francis Bond, Claire Bonial, Julia Bonn, Tiberiu Boroș, Marie Candito, Giuseppe G. A. Celano, Kenneth Church, Çağrı Çöltekin, Mathieu Constant, Monika Czerepowicka, Daniel Dakota, Miryam de Lhoneux, Marie-Catherine de Marneffe, Valeria de Paiva, Gaël Dias, Kaja Dobrovoljc, Rafael Ehren, Gülşen Eryiğit, Meghdad Farahmand, Christiane Fellbaum, Jennifer Foster, Aggeliki Fotopoulou, Stefan Th. Gries, Bruno Guillaume, Tunga Gungor, Eleonora Guzzi, Laura Kallmeyer, Cvetana Krstev, Timm Lichte, Irina Lobzhanidze, Teresa Lynn, Stella Markantonatou, John P. McCrae, Nurit Melnik, Johanna Monti, Dmitry Nikolaev, Jan Odijk, Petya Osenova, Yannick Parmentier, Agnieszka Patejuk, Pavel Pecina, Ted Pedersen, Prokopis Prokopidis, Manfred Sailer, Tanja Samardžić, Agata Savary, Nathan Schneider, Sabine Schulte im Walde, Sebastian Schuster, Matthew Shardlow, Joaquim Silva, Maria Simi, Ranka Stanković, Ivelina Stoyanova, Stan Szpakowicz, Shiva Taslimipoor, Beata Trawinski, Ashwini Vaidya, Marion Di Marco, Amir Zeldes, Daniel Zeman

Keynote Speakers

Natalia Levshina, Radboud University
Harish Tayyar Madabushi, University of Bath

Organized by

Funded by the European Union

Table of Contents

Workshop Program

09:00–09:05 **Welcome**

09:05–09:50 **Keynote Speaker 1**

Every Time We Hire an LLM, the Reasoning Performance of the Linguists Goes Up
Harish Tayyar Madabushi

09:50–10:30 **Session 1: Oral presentations**

09:50–10:10 Assessing BERT's sensitivity to idiomaticity
Li Liu and Francois Lareau

10:10–10:30 Sign of the Times: Evaluating the use of Large Language Models for Idiomaticity Detection
Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith and Aline Villavicencio

10:30–11:00 **Coffee Break**

11:00–12:00 **Session 2: Poster presentations**

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

12:00–12:20 Universal Feature-based Morphological Trees
Federica Gamba, Abishek Stephen and Zdeněk Žabokrtský

12:20–12:40 Light Verb Constructions in Universal Dependencies for South Asian Languages
Abishek Stephen and Daniel Zeman

12:40–13:00 Strategies for the Annotation of Pronominalised Locatives in Turkic Universal Dependency Treebanks
Jonathan Washington, Çağrı Çöltekin, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Gulnura Jumalieva, Aida Kasieva, Aslı Kuzgun, Büşra Marşan and Chihiro Taguchi

13:00–14:00 **Lunch**

14:00–14:45 **Keynote Speaker 2**

Using Universal Dependencies for testing hypotheses about communicative efficiency
Natalia Levshina

14:45–15:00 **Booster Session: Remote presentations**

15:00–16:00 **Session 4: Oral presentations**

15:00–15:20 To Leave No Stone Unturned: Annotating Verbal Idioms in the Parallel Meaning Bank
Rafael Ehren, Kilian Evang and Laura Kallmeyer

15:20–15:40 Annotation of Multiword Expressions in the SUK 1.0 Training Corpus of Slovene: Lessons Learned and Future Steps
Jaka Čibej, Polona Gantar and Mija Bon

15:40–16:00 Ad Hoc Compounds for Stance Detection
Qi Yu, Fabian Schlotterbeck, Henning Wang, Naomi Reichmann, Britta Stolterfoht, Regine Eckardt and Miriam Butt

16:00–16:30 **Coffee Break**

16:30–17:20 **Session 5: Poster presentations**

17:20–17:50 **Community Discussion**

17:50–18:00 **Closing and Awards**

Posters and Remote Presentations

Automatic Manipulation of Training Corpora to Make Parsers Accept Real-world Text
Hiroshi Kanayama, Ran Iwamoto, Masayasu Muraoka, Takuya Ohko and Kohtaroh Miyamoto

Identification and Annotation of Body Part Multiword Expressions in Old Egyptian
Roberto Díaz Hernández

Fitting Fixed Expressions into the UD Mould: Swedish as a Use Case
Lars Ahrenberg

Synthetic-Error Augmented Parsing of Swedish as a Second Language: Experiments with Word Order
Arianna Masciolini, Emilie Francis and Maria Irena Szawerna

The Vedic Compound Dataset
Sven Sellmer and Oliver Hellwig

A Universal Dependencies Treebank for Gujarati
Mayank Jobanputra, Maitrey Mehta and Çağrı Çöltekin

Overcoming Early Saturation on Low-Resource Languages in Multilingual Dependency Parsing
Jiannan Mao, Chenchen Ding, Hour Kaing, Hideki Tanaka, Masao Utiyama and Tadahiro Matsumoto

Part-of-Speech Tagging for Northern Kurdish
Peshmerge Morad, Sina Ahmadi and Lorenzo Gatti

Diachronic Analysis of Multi-word Expression Functional Categories in Scientific English
Diego Alves, Stefania Degaetano-Ortlieb, Elena Schmidt and Elke Teich

Lexicons Gain the Upper Hand in Arabic MWE Identification
Najet Hadj Mohamed, Agata Savary, Cherifa Ben Khelil, Jean-Yves Antoine, Iskandar keskes and Lamia Hadrich-Belguith

Revisiting VMWEs in Hindi: Annotating Layers of Predication
Kanishka Jain and Ashwini Vaidya

Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities
Cvetana Krstev, Ranka Stanković, Aleksandra M. Marković and Teodora Sofija Mihajlov

Combining Grammatical and Relational Approaches. A Hybrid Method for the Identification of Candidate Collocations from Corpora
Damiano Perri, Irene Fioravanti, Osvaldo Gervasi and Stefania Spina

Multiword Expressions between the Corpus and the Lexicon: Universality, Idiosyncrasy, and the Lexicon-Corpus Interface
Verginica Barbu Mititelu, Voula Giouli, Kilian Evang, Daniel Zeman, Petya Osenova, Carole Tiberius, Simon Krek, Stella Markantonatou, Ivelina Stoyanova, Ranka Stanković and Christian Chiarcos