MWE-LEX 2020

Joint Workshop on Multiword Expressions and Electronic Lexicons

Proceedings of the Workshop

December 13, 2020
Barcelona, Spain (Online)

ISBN 978-1-952148-50-7

Introduction

The Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020)[1] took place in an online format on December 13, 2020 in conjunction with COLING 2020.

This was the 16th edition of the Workshop on Multiword Expressions (MWE 2020). The event was organized and sponsored by the Special Interest Group on the Lexicon (SIGLEX)[2] of the Association for Computational Linguistics (ACL) and by ELEXIS[3] - European Lexicographic Infrastructure.

The joint MWE-LEX workshop addressed two domains – multiword expressions and (electronic) lexicons – with partly overlapping communities and research interests, but divergent practices and terminologies.

Multiword expressions (MWEs) are word combinations, such as in *the middle of nowhere*, *hot dog*, *to make a decision* or *to kick the bucket*, displaying lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies. Because of their unpredictable behavior, notably their non-compositional semantics, MWEs pose problems in linguistic modelling (e.g. treebank annotation, grammar engineering), NLP pipelines (notably when orchestrated with parsing), and end-user applications (e.g. information extraction).

Because MWE-hood is a largely lexical phenomenon, appropriately built electronic MWE lexicons turn out to be quite important for NLP. Large standardised multilingual, possibly interconnected, NLP-oriented MWE lexicons prove indispensable for NLP tasks such as MWE identification, due to its critical sensitivity to unseen data. But the development of such lexicons is challenging and calls for tools which would leverage, on the one hand, MWEs encoded in pre-existing NLP-unaware lexicons and, on the other hand, automatic MWE discovery in large non-annotated corpora.

In order to allow better convergence and scientific innovation within these two largely complementary scientific communities, we called for papers on joint topics on MWEs and e-lexicons, on the one hand, and on MWE-specific topics, on the other hand.

Joint topics on MWEs and e-lexicons

- Extracting and enriching MWE lists from traditional human-readable lexicons for NLP use
- Formats for NLP-applicable MWE lexicons
- Interlinking MWE lexicons with other language resources
- Using MWE lexicons in NLP tasks (identification, parsing, translation)
- MWE discovery in the service of lexicography
- Multiword terms in specialized lexicons
- Representing semantic properties of MWEs in lexicons
- Paving the way towards encoding lexical idiosyncrasies in constructions

[1] http://multiword.sourceforge.net/mwelex2020
[2] http://alt.qcri.org/siglex/
[3] https://elex.is/

MWE-specific topics

- Computationally-applicable theoretical work on MWEs and constructions in psycholinguistics, corpus linguistics and formal grammars
- MWE and construction annotation in corpora and treebanks
- Processing of MWEs and constructions in syntactic and semantic frameworks (e.g. CCG, CxG, HPSG, LFG, TAG, UD, etc.), and in end-user applications (e.g. information extraction, machine translation and summarization)
- Original discovery and identification methods for MWEs and constructions
- MWEs and constructions in language acquisition and in non-standard language (e.g. tweets, forums, spontaneous speech)
- Evaluation of annotation and processing techniques for MWEs and constructions
- Retrospective comparative analyses from the PARSEME shared tasks on automatic identification of MWEs

We received 25 submissions (14 long and 11 short papers). We selected 6 long papers and 7 short ones. All 13 accepted papers were presented orally. The overall acceptance rate was 52 %.

In addition to the oral sessions, the workshop featured an invited talk that given by Roberto Navigli from Sapienza University of Rome.

The workshop also organized the PARSEME Shared Task on Semi-Supervised Identification of Verbal MWEs (edition 1.2). This was a follow-up of editions 1.0 (2017), and 1.1 (2018). Edition 1.2 featured (a) improved and extended corpora annotated with MWEs, (b) complementary unannotated corpora for unsupervised MWE discovery, and (c) a new evaluation methodology focusing on unseen MWEs. Following the synergy with Elexis, our aim was to foster the development of unsupervised methods for MWE lexicon induction, which in turn can be used for identification.

Seven teams submitted 9 system results to the shared task (some teams made 2 submissions): 2 to the closed track (where only the data provided by the organizers could be used on input) and 7 to the open track (where external data were also allowed). Out of these 9 results, 4 covered all 14 languages for which data were made available by the organizers. Six teams also submitted systems description papers, all of which were accepted and presented orally.

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 all the workshop participants for their interest in this event. Our thanks also go to the COLING 2020 organizers for their support, as well as to SIGLEX and ELEXIS for their endorsement.

Stella Markantonatou, John McCrae, Jelena Mitrović, Carole Tiberiu, Carlos Ramisch, Ashwini Vaidya, Petya Osenova, Agata Savary

Organizers:

Stella Markantonatou, Institute for Language and Speech Processing, R.C. "Athena" (Greece)
John McCrae, National University of Ireland Galway (Ireland)
Jelena Mitrović, University of Passau (Germany)
Carole Tiberius, Dutch Language Institute in Leiden (Netherlands)
Carlos Ramisch, Aix Marseille University (France)
Ashwini Vaidya, Indian Institute of Technology in Delhi (India)
Petya Osenova, Institute of Information and Communication Technologies (Bulgaria)
Agata Savary, University of Tours (France)

Program Committee:

Tim Baldwin, University of Melbourne (Australia)
Verginica Barbu Mititelu, Romanian Academy (Romania)
Archna Bhatia, Florida Institute for Human and Machine Cognition (USA)
Francis Bond, Nanyang Technological University (Singapore)
Tiberiu Boroș, Adobe (Romania)
Marie Candito, Paris Diderot University (France)
Helena Caseli, Federal University of Sao Carlos (Brazil)
Anastasia Christofidou, Academy of Athens (Greece)
Ken Church, IBM Research (USA)
Matthieu Constant, Université de Lorraine (France)
Paul Cook, University of New Brunswick (Canada)
Monika Czerepowicka, University of Warmia and Mazury (Poland)
Béatrice Daille, Nantes University (France)
Gerard de Melo, Rutgers University (USA)
Thierry Declerck, DFKI (Germany)
Gaël Dias, University of Caen Basse-Normandie (France)
Meghdad Farahmand, University of Geneva (Switzerland)
Christiane Fellbaum, Princeton University (USA)
Joaquim Ferreira da Silva, New University of Lisbon (Portugal)
Aggeliki Fotopoulou, ILSP/RC "Athena" (Greece)
Francesca Frontini, Université Paul-Valéry Montpellier (France)
Marcos Garcia, CITIC (Spain)
Voula Giouli, Institute for Language and Speech Processing (Greece)
Chikara Hashimoto, Yahoo!Japan (Japan)
Kyo Kageura, University of Tokyo (Japan)
Diptesh Kanojia, IITB-Monash Research Academy (India)
Dimitris Kokkinakis, University of Gothenburg (Sweden)
Ioannis Korkontzelos, Edge Hill University (UK)
Iztok Kosem, Jožef Stefan Institute (Slovenia)
Cvetana Krstev, University of Belgrade (Serbia)
Malhar Kulkarni, Indian Institute of Technology, Bombay (India)
Eric Laporte, University Paris-Est Marne-la-Vallee (France)
Timm Lichte, University of Duesseldorf (Germany)
Irina Lobzhanidze, Ilia State University (Georgia)

Ismail el Maarouf, Adarga Ltd (UK)
Yuji Matsumoto, Nara Institute of Science and Technology (Japan)
Nurit Melnik, The Open University of Israel (Israel)
Elena Montiel-Ponsoda, Universidad Politecnica de Madrid (Spain)
Sanni Nimb, Det Denske Sprog- og Litteraturselskab (Denmark)
Haris Papageorgiou, Institute for Language and Speech Processing (Greece)
Carla Parra Escartín, Unbabel (Portugal)
Marie-Sophie Pausé, independent researcher (France)
Pavel Pecina, Charles University (Czech Republic)
Scott Piao, Lancaster University (UK)
Alain Polguère, Université de Lorraine (France)
Alexandre Rademaker, IBM Research Brazil and EMAp/FGV (Brazil)
Laurent Romary, INRIA & HUB-ISDL (France)
Mike Rosner, University of Malta (Malta)
Manfred Sailer, Goethe-Universität Frankfurt am Main (Germany)
Magali Sanches Duran, University of São Paulo (Brazil)
Nathan Schneider, Georgetown University (USA)
Sabine Schulte im Walde, University of Stuttgart (Germany)
Kiril Simov, Bulgarian Academy of Sciences (Bulgaria)
Ranka Stanković, University of Belgrade (Serbia)
Ivelina Stoyanova, Bulgarian Academy of Sciences (Bulgaria)
Stan Szpakowicz, University of Ottawa (Canada)
Shiva Taslimipoor, University of Wolverhampton (UK)
Arvi Tavast, Qlaara, Tallinn (Estonia)
Beata Trawinski, Institut für Deutsche Sprache Mannheim (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)
Jakub Waszczuk, University of Duesseldorf (Germany)
Eric Wehrli, University of Geneva (Switzerland)
Seid Muhie Yimam, Universität Hamburg (Germany)

Invited Speaker:

Roberto Navigli, Sapienza University of Rome

Table of Contents

CollFrEn: Rich Bilingual English–French Collocation Resource
Beatriz Fisas, Joan Codina-Filbá, Luis Espinosa Anke and Leo Wanner 1

Filling the ___-s in Finnish MWE lexicons
Frankie Robertson 13

Hierarchy-aware Learning of Sequential Tool Usage via Semi-automatically Constructed Taxonomies
Nima Nabizadeh, Martin Heckmann and Dorothea Kolossa 22

Scalar vs. mereological conceptualizations of the N-BY-N and NUM-BY-NUM adverbials
Lucia Vlášková and Mojmír Dočekal 27

Polish corpus of verbal multiword expressions
Agata Savary and Jakub Waszczuk 32

AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations
Lifeng Han, Gareth Jones and Alan Smeaton 44

Annotating Verbal MWEs in Irish for the PARSEME Shared Task 1.2
Abigail Walsh, Teresa Lynn and Jennifer Foster 58

VMWE discovery: a comparative analysis between Literature and Twitter Corpora
Vivian Stamou, Artemis Xylogianni, Marilena Malli, Penny Takorou and Stella Markantonatou . 66

Generationary or: "How We Went beyond Sense Inventories and Learnedto Gloss"
Roberto Navigli 73

Multi-word Expressions for Abusive Speech Detection in Serbian
Ranka Stankovic, Jelena Mitrović, Danka Jokic and Cvetana Krstev 74

Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings
Murathan Kurfalı and Robert Östling 85

Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality
Thomas Pickard 95

Automatic detection of unexpected/erroneous collocations in learner corpus
Jen-Yu Li and Thomas Gaillat 101

Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions
Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Gungor, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh and Hongzhi Xu 107

HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and unseen-in-training MWEs
Jean-Pierre Colson 119

Seen2Unseen at PARSEME Shared Task 2020: All Roads do not Lead to Unseen Verb-Noun VMWEs
Caroline Pasquer, Agata Savary, Carlos Ramisch and Jean-Yves Antoine 124

ERMI at PARSEME Shared Task 2020: Embedding-Rich Multiword Expression Identification
Zeynep Yirmibeşoğlu and Tunga Gungor 130

TRAVIS at PARSEME Shared Task 2020: How good is (m)BERT at seeing the unseen?
Murathan Kurfalı 136

MTLB-STRUCT @Parseme 2020: Capturing Unseen Multiword Expressions Using Multi-task Learning and Pre-trained Masked Language Models
Shiva Taslimipoor, Sara Bahaadini and Ekaterina Kochmar 142

MultiVitaminBooster and MultiVitaminRegressor at PARSEME Shared Task 2020: Combining Window- and Dependency-Based Features with Multilingual Contextualized Word Embeddings for Detecting Verbal Multiword Expressions
Sebastian Gombert and Sabine Bartsch 149

Workshop Program

Sunday, December 13, 2020

14:00–14:10 ***Welcoming and preparation***

14:10–14:40 ***Session 1: MWE Resources and Linguistics***

CollFrEn: Rich Bilingual English–French Collocation Resource
Beatriz Fisas, Joan Codina-Filbá, Luis Espinosa Anke and Leo Wanner

Filling the ___-s in Finnish MWE lexicons
Frankie Robertson

Hierarchy-aware Learning of Sequential Tool Usage via Semi-automatically Constructed Taxonomies
Nima Nabizadeh, Martin Heckmann and Dorothea Kolossa

Scalar vs. mereological conceptualizations of the N-BY-N and NUM-BY-NUM adverbials
Lucia Vlášková and Mojmír Dočekal

14:40–14:50 ***Break***

14:50–15:20 ***Session 2: Verbal Multiword Expressions***

Polish corpus of verbal multiword expressions
Agata Savary and Jakub Waszczuk

AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations
Lifeng Han, Gareth Jones and Alan Smeaton

Annotating Verbal MWEs in Irish for the PARSEME Shared Task 1.2
Abigail Walsh, Teresa Lynn and Jennifer Foster

VMWE discovery: a comparative analysis between Literature and Twitter Corpora
Vivian Stamou, Artemis Xylogianni, Marilena Malli, Penny Takorou and Stella Markantonatou

Sunday, December 13, 2020 (continued)

15:20–15:30 ***Break***

15:30–16:30 ***Session 3: Invited Talk***

Generationary or: "How We Went beyond Sense Inventories and Learnedto Gloss"
Roberto Navigli

16:30–16:40 ***Break***

16:40–17:10 ***Session 4: Processing Multiword Expressions***

Multi-word Expressions for Abusive Speech Detection in Serbian
Ranka Stankovic, Jelena Mitrović, Danka Jokic and Cvetana Krstev

Disambiguation of Potentially Idiomatic Expressions with Contextual Embeddings
Murathan Kurfalı and Robert Östling

Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality
Thomas Pickard

Automatic detection of unexpected/erroneous collocations in learner corpus
Jen-Yu Li and Thomas Gaillat

17:10–17:20 ***Break***

Sunday, December 13, 2020 (continued)

17:20–18:00 ***Session 5: Shared Task***

Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions
Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Gungor, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh and Hongzhi Xu

HMSid and HMSid2 at PARSEME Shared Task 2020: Computational Corpus Linguistics and unseen-in-training MWEs
Jean-Pierre Colson

Seen2Unseen at PARSEME Shared Task 2020: All Roads do not Lead to Unseen Verb-Noun VMWEs
Caroline Pasquer, Agata Savary, Carlos Ramisch and Jean-Yves Antoine

ERMI at PARSEME Shared Task 2020: Embedding-Rich Multiword Expression Identification
Zeynep Yirmibeşoğlu and Tunga Gungor

TRAVIS at PARSEME Shared Task 2020: How good is (m)BERT at seeing the unseen?
Murathan Kurfalı

MTLB-STRUCT @Parseme 2020: Capturing Unseen Multiword Expressions Using Multi-task Learning and Pre-trained Masked Language Models
Shiva Taslimipoor, Sara Bahaadini and Ekaterina Kochmar

MultiVitaminBooster and MultiVitaminRegressor at PARSEME Shared Task 2020: Combining Window- and Dependency-Based Features with Multilingual Contextualized Word Embeddings for Detecting Verbal Multiword Expressions
Sebastian Gombert and Sabine Bartsch

18:00–18:10 ***Break***

18:10–19:10 ***Session 6: Section reporting, panel discussion***