Istorijat i pregled radionice posvećene obradi polileksemskih jedinica (eng. Multiword expressions)

MWE-LEX 2020

JELENA MITROVIĆ, UNIVERZITET PASSAU

18-02-2021

MWEs and SIGLEX-MWE

https://multiword.org/

Welcome to SIGLEX-MWE section!

SIGLEX-MWE is a section of the Special Interest Group on the Lexicon (SIGLEX) of the Association for Computational Linguistics (ACL). It is dedicated to promoting scientific activity on multiword expressions (MWEs) in computational linguistics. Its main activity is the coordination of the MWE workshop series and related events such as the PARSEME shared tasks.

- MWE are „idiosyncratic interpretations that cross word boundaries“.
- idioms as „kick the bucket“
- compound nouns as „telephone box“ and „post office“
- verb-particle constructions as „look sth. up“ or proper names as „San Francisco“.

Kako je sve počelo?

2003

ACL workshop *Multiword Expressions: Analysis, Acquisition and Treatment* - Sapporo, Japan.

2004

ACL workshop *Multiword Expressions: Integrating Processing* - Barcelona, Spain.

2006

ACL workshop *Multiword Expressions: Identifying and Exploiting Underlying Properties* - Sydney, Australia. [Proceedings]

MWE-LEX 2020 (COLING) MWE-WN 2019 (ACL) LAW-MWE-CxG 2018 (COLING)

MWE 2017 (EACL) MWE 2016 (ACL) MWE 2015 (NAACL)

MWE 2014 (EACL) MWE 2013 (NAACL) STARSEM 2012

MWE 2011 (ACL) MWE 2010 (COLING) MWE 2009 (ACL)

MWE 2008 (LREC) MWE 2007 (ACL)

2020 - 16. izdanje radionice

Udruženi rad sa radionicom posvećenoj elektronskim leksikonima

- MWE-hood is a largely lexical phenomenon
- MWE lexicons are indispensable for robust MWE identification
- Partly overlapping communities/goals, but divergent practices/terms

PARSEME Shared Task 1.2 on Semi-Supervised verbal MWE Identification

Organizacija i sponzorstvo:

- Special Interest Group on the Lexicon (SIGLEX) of the ACL
- ELEXIS - European Lexicographic Infrastructure.

MWE-LEX 2020 Statistika

Research track

- 25 radova predato
- 14 dužih radova
- 11 kraćih radova
- 13 prihvaćenih radova - 7 dužih i 6 kraćih

52% ukupno prihvaćenih radova

Svi prihvaćeni radovi dostupni preko platforme Underline – rad + prezentacija (video)

Shared task track

- 9 sistema i 7 timova (4 sistema pokrivaju svih 14 jezika)
- 6 timova predalo je radove koji opisuju njihove sisteme
- svi prihvaćeni radovi su predstavljeni

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

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

Program

14:00–14:10 Welcome

14:10–14:40 **Session 1: MWE resources and linguistics (30min) -** Chairs: Jean-Pierre Colson & Antonios Anastasopoulos

14:40–14:50 Break (10min)

14:50–15:20 **Session 2: Verbal multiword expressions (30min) -** Chairs: Archna Bhatia & Carole Tiberius

15:20–15:30 Break (10min)

15:30–16:30 **Session 3: Keynote speech (60min) -** Chairs: Jelena Mitrović & Agata Savary
A close look at Generationary or: "How We Went beyond Sense Inventories and Learned to Gloss"
Roberto Navigli

16:30–16:40 Break (10min)

16:40-17:10 **Session 4: Processing multiword expressions (30min) -** Chairs: Shiva Taslimipoor & Vivian Stamou

17:10–17:20 Break (10min)

17:20-18:00 **Session 5: Shared task (40min) -** Chairs: Ashwini Vaidya & Lifeng Han

18:00–18:10 Break

18:10-19:00 **Session 6: Section reporting, panel discussion (50min) -** Chairs: Carlos Ramisch & Agata Savary

Program

Videos are online for all papers on Underline

Four 30-min Q&A sessions with 2 long and 2 short papers

- Session chairs manage who talks when
- Presenters are expected to give an "elevator pitch" of about 1 min

Invited talk and community discussion are live

Attendees can ask questions by:

- Typing the question in the chat (Q&A) at any moment
  - Questions typed in the Q&A of the video are not visible in the discussion room
- Use the "Raise hand" button and join the live session with mic/camera

The 2020 conference etiquette

- Please turn off your mic when others are speaking
- You cannot delete messages from the Q&A chat
- Asking your question live makes the workshop more lively/human :-)
- Avoid "zoom fatigue": frequent 10-min breaks to stretch, drink water, eat a fruit, ...

Session 1 - MWE resources and linguistics (30 min)

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

LONG Filling the ___-s in Finnish MWE lexicons, Frankie Robertson

---

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

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

Session 2 - Verbal multiword expressions (30 min)

LONG Polish corpus of verbal multiword expressions, Agata Savary and Jakub Waszczuk

LONG AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations, LIFENG HAN, Gareth Jones and Alan Smeaton

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

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

Invited speaker

Roberto Navigli

Sapienza University of Rome (Italy)

A closer look at Generationary or: "How We Went beyond Sense Inventories and Learned to Gloss"

Live at 15:30-16:30 (CET) on Underline

KEYNOTE SPEECH (60 min)

Generationary or: "How We Went beyond Sense Inventories and Learned to Gloss", Roberto Navigli

---

16:40–17:10 Session 3 -- Processing multiword expressions \ (30 min)

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

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

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

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

Session 4 -- Shared task (40 min)

Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions, Carlos Ramisch, Agata Savary et al.

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 Yirmibe¸so˘glu 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

MWE Section reporting, panel discussion

Organizing Committee

MWE-LEX topics

- **John McCrae**, National University of Ireland Galway (Ireland)
- **Carole Tiberius**, Dutch Language Institute in Leiden (Netherlands)

MWE-specific topics

- **Stella Markantonatou**, ILSP, R.C. "Athena" (Greece)
- **Jelena Mitrović**, University of Passau (Germany)

Shared task

- **Carlos Ramisch**, Aix-Marseille University (France)
- **Ashwini Vaidya**, Indian Institute of Technology in Delhi (India)

Publication Chairs

Publication chairs (a.k.a. SoftConf heroes)

- **Petya Osenova**, University of Sofia and Bulgarian Academy of Sciences (Bulgaria)
- **Agata Savary**, Université of Tours (France)

Zbornik radova je dostupan na ACL Anthology
https://www.aclweb.org/anthology/volumes/2020.mwe-1/

Thank You

Program committee (reviewers): Tim Baldwin, Verginica Barbu Mititelu, Archna Bhatia, Francis Bond, Tiberiu Boroș, Marie Candito, Helena Caseli, Anastasia Christofidou, Ken Church, Matthieu Constant, Paul Cook, Monika Czerepowicka, Béatrice Daille, Gerard de Melo, Thierry Declerck, Gaël Dias, Meghdad Farahmand, Christiane Fellbaum, Joaquim Ferreira da Silva, Aggeliki Fotopoulou, Francesca Frontini, Marcos Garcia, Voula Giouli, Chikara Hashimoto, Kyo Kageura, Diptesh Kanojia, Dimitris Kokkinakis, Ioannis Korkontzelos, Iztok Kosem, Cvetana Krstev, Malhar Kulkarni, Eric Laporte, Timm Lichte, Irina Lobzhanidze, Ismail el Maarouf, Yuji Matsumoto, Nurit Melnik, Elena Montiel-Ponsoda, Sanni Nimb, Haris Papageorgiou, Carla Parra Escartín, Marie-Sophie Pausé, Pavel Pecina, Scott Piao, Alain Polguère, Alexandre Rademaker, Laurent Romary, Mike Rosner, Manfred Sailer, Magali Sanches Duran, Nathan Schneider, Sabine Schulte im Walde, Kiril Simov, Ranka Stanković, Ivelina Stoyanova, Stan Szpakowicz, Shiva Taslimipoor, Arvi Tavast, Beata Trawinski, Zdeňka Urešová, Ruben Urizar, Lonneke van der Plas, Veronika Vincze, Jakub Waszczuk, Eric Wehrli, Seid Muhie Yimam.

Session chairs: Jean-Pierre Colson, Antonios Anastasopoulos, Archna Bhatia, Carole Tiberius, Jelena Mitrović, Agata Savary, Shiva Taslimipoor, Vivian Stamou, Ashwini Vaidya, Lifeng Han, Carlos Ramisch

COLING 2020 organizers, workshop chairs, volunteers, Underline staff

All authors, presenters, participants

Community Discussion

Feedback

- From the workshop attendees
- From the shared task participants

Announcements

- SIGLEX-MWE news

Future

- MWE workshop
- Parseme shared task

17th Workshop on Multiword Expressions (MWE 2021)

- Not a joint event, but joint session with Workshop on Online Abuse and Harm (WOAH)
- Collocated with ACL-IJCNLP 2021 in Bangkok, Thailand (fingers crossed!)
  - Workshop dates: August 5 – 6, 2021
  - Submission deadline April 19th 2021
  - **Prepare your submissions!**

MWE 2022 and beyond

- Continue joint workshops with other communities?
  - Universal dependencies
  - Figurative language
  - Designing meaning representations
  - VarDial - variation and dialects
  - Machine translation and/or MUMTTT

Multi-word Expressions for Abusive Speech Detection in Serbian

Ranka Stankovic[1], Jelena Mitrovic[2], Danka Jokic[1] and Cvetana Krstev[1]

Lexical Cleaning of Serbian part of Hurtlex

- Alphabet unification (Cyrillic to Latin)
- Entries with non-existing words removal with the help of SMD
- Lemma and POS correction
- Removing duplicates

2528 [1903] -> 1402 [1000]

Cleaning: 265 entries -> 156 confirmed

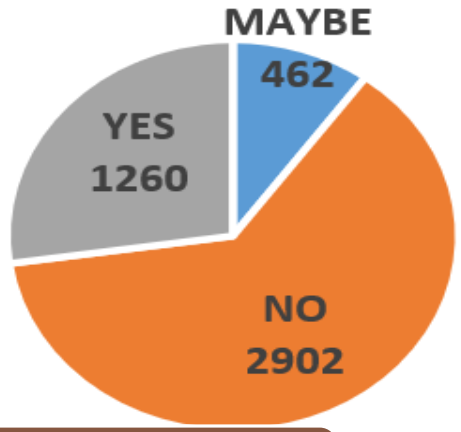
Reclassification

- Wrong category assigned
- Abusive meaning lost in translation or not abusive in Serbian
- Alignment of conservative and inclusive meaning
- Paraphrasing translation to be meaningful

The MWE Lexicon Construction Step-by-Step

Initial cleaning and classification of the abusive MWEs list

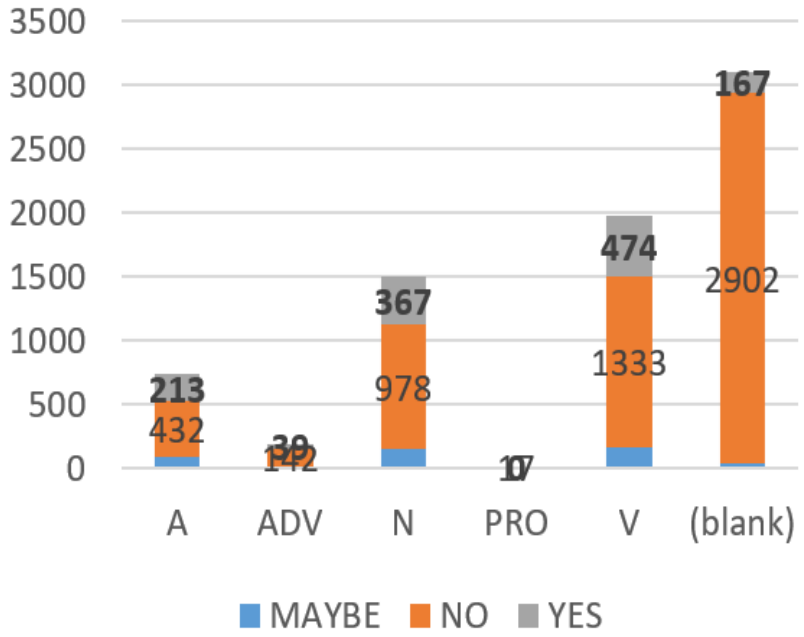
- YES - abusive speech
- MAYBE - could lead to abusive content
- NO - not abusive



Example

- for trigger “junak”(hero) **abusive MWE** junak na jeziku (scaramouch), **non-abusive MWE** ‘junak romana’(a hero of the novel)

MWEs classified as
YES, NO, MAYBE
and
POS of trigger words



The MWE Lexicon Construction Step-by-Step

	ženski petko (abusive for a man not manly enough)
noun lemma	ženski(ženski.A2:adms1g) petko(petko.N68:ms1v),NC_AXN+Hum+HRT=yes
noun form	ženskog petka,ženskog petka.N+Hum+HRT=yes:ms2v

	glup kao klada (stupid as a log)
adj lemma	glup(glup.A15:akms1g) kao klada,AC_A4X+Simile+HRT=yes
adj form	glupog kao klada,glup kao klada.A+Simile+HRT=yes:adms2g kao klada glupog,glup kao klada.A+Simile+HRT=yes:adms2g

- Evaluation done using abusive and potentially abusive MWEs on separate Twitter corpus containing 8000 tweets
- 800 hits obtained of which 80-90% indicated the abusive language

Hvala na pažnji!

Vidimo se na sledećoj MWE radionici ☺