The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (1.1)

C. Ramisch[1], S. Ricardo[1], A. Savary[2], V. Vincze[3], V. Barbu[4], A. Bhatia[5], M. Buljan[6], M. Candito[7], P. Gantar[8], V. Giouli[9], T. Güngör[10], A. Hawwari[11], U. Iñurrieta[12], J. Kovalevskaitė[13], S. Krek[14], T. Lichte[15], C. Liebeskind[16], J. Monti[17], C. Parra[18], B. QasemiZadeh[15], R. Ramisch[19], N. Schneider[20], I. Stoyanova[21], A. Vaidya[22], A. Walsh[18]

[1]Aix-Marseille University, **France**, [2]University of Tours, **France**, [3]University of Szeged, **Hungary**, [4]Romanian Academy, **Romania**, [5]Florida IHMC, **USA**, [6]University of Stuttgart, **Germany**, [7]Paris Diderot University, **France**, [8]Faculty of Arts, **Slovenia**, [9]Athena Research Center, **Greece**, [10]Boğaziçi University, **Turkey**, [11]George Washington University, **USA**, [12]University of the Basque Country, **Spain**, [13]Vytautas Magnus University, **Lithuania**, [14]Jožef Stefan Institute, **Slovenia**, [15]University of Düsseldorf, **Germany**, [16]Jerusalem College of Technology, **Israel**, [17]"L'Orientale" University of Naples, **Italy**, [18]Dublin City University, **Ireland**, [19]Interinstitutional Center for Computational Linguistics, **Brazil**, [20]Georgetown University, **USA**, [21]Bulgarian Academy of Sciences, **Bulgaria**, [22]IIT Delhi, **India**,

PARSEME

A multilingual shared task on MWE identification

- What is MWE identification?
  - INPUT: text
  - OUTPUT: text annotated with MWEs
- PARSEME shared task – edition 1.0 in 2017

Why focus on verbal MWEs (VMWEs)? I

- Discontinuity:
  - EN *turn the TV off*
- Variability: morphological, syntactic, lexical
  - EN *we made decisions* vs. *the decision was hard to make*
- Non-categorical nature:
  - Same surface, different syntax
    - EN *take on the task* (VPC.full) vs. *to sit on the chair*
  - Same syntax, different category
    - EN *to make a mistake* (LVC.full)
    - EN *to make a meal* of sth (VID)
- Ambiguity: idiomatic vs. literal readings
  - EN *to take the cake*

Why focus on verbal MWEs (VMWEs)? II

- Overlaps:
  - Factorization

    EN ***take*** *a* ***walk*** *and then a long* ***shower*** (coordination)
  - Nesting

    open slots: EN ***take*** *the fact that I* ***gave up*** ***into account***

    lexicalized components: EN ***let*** ***the cat*** ***out*** ***of the bag***
- Multiword tokens

  ES ***abstener/se*** (lit. *abstain self*) 'abstain'

  DE ***auf/machen*** (lit. *out/make*) 'open'
- Different languages ⇒ different behavior, linguistic traditions. . .

PARSEME shared task 1.0 at a glance

- Multilingual guidelines with examples
- Annotation methodology and teams (PARSEME)
- Corpora in 18 languages under free licenses
- Train/test corpora with 52724/9494 VMWEs
- New evaluation measures (MWE-/Token-based)
- 7 participating systems

Enhanced guidelines

- Discussion via Gitlab issues
- Main definitions remain:
  - Words and tokens
  - Lexicalized components and open slots
  - Canonical forms
- Generic decision tree based on structural tests

Decision tree

↳ Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
- ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
  - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
  - ↳ **NO** ⇒ It is not a VMWE, **exit**
- ↳ **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
  - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
    - ↳ **NO** ⇒ It is not a VMWE, **exit**
  - ↳ **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
    - ↳ **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
      - ↳ **NO** ⇒ It is not a VMWE, **exit**
    - ↳ **NO** ⇒ Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
      - ↳ **Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **IRV**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **VPC.full** or **VPC.semi**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **Verb with no lexicalized dependent** ⇒ Apply MVC-specific tests ⇒ *MVC tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **MVC**
        - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **Extended NP** ⇒ Apply LVC-specific decision tree ⇒ *LVC tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **LVC**
        - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **Another category** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**

VMWE typology I

Universal categories (all languages)

- verbal idioms (**VID**)
  - EN *to **call it a day***
- light-verb constructions (**LVCs**)
  - EN *to **give a lecture*** (LVC.full)
  - EN *to **grant rights*** (LVC.cause)

VMWE typology II

Quasi-universal categories (many languages)

- inherently reflexive verbs (**IRVs**)
  - EN *to **help oneself*** 'to take something freely'
- verb-particle constructions (**VPCs**)
  - EN *to **do in*** 'to kill' (VPC.full)
  - EN *to **eat up*** (VPC.semi)
- multi-verb constructions (**MVCs**)
  - HI ***kar le-na*** (lit. *do take.INF*) 'to do something (for one's own benefit)'

VMWE typology III

Optional/language-specific categories

- inherently clitic verbs (**LS.ICV**)

  IT ***prenderle*** (lit. *take it*) 'get beaten up'

- inherently adpositional verbs (**IAV**)

  EN *to **rely on***

Require more work to be generalized/stabilized

20 languages

Language groups

- **Balto-Slavic**: Bulgarian (BG), Croatian (HR), Lithuanian (LT), Polish (PL), Slovene (SL), ~~Czech (CZ)~~
- **Germanic**: German (DE), English (EN), ~~Swedish (SV)~~
- **Romance**: French (FR), Italian (IT), Romanian (RO), Spanish (ES), Brazilian Portuguese (PT)
- **Others**: Arabic (AR), Greek (EL), Basque (EU), Farsi (FA), Hebrew (HE), Hindi (HI), Hungarian (HU), Turkish (TR), ~~Maltese (MT)~~

Corpora

Corpus	Sent.	Tokens	VMWE
train	208,420	**4,553,431**	**59,460**
dev	31,947	**672,102**	**9,250**
test	40,471	**846,798**	**10,616**
total	280,838	**6,072,331**	**79,326**

- Varying corpus sizes per language
- No dev in EN, HI and LT
- New rules for train/dev/test split
- Morphological/syntactic information (mostly UD)

Availability

- 19 corpora released under **Creative Commons** licenses

Format

CUPT: extension of the CoNLL-U format

1	-	-	PUNCT	_	_	4	punct	_	_	*
2	si	si	SCONJ	_	_	4	mark	_	_	*
3	vous	il	PRON	_	_	4	nsubj	_	_	*
4	présentez	présenter	VERB	_	_	0	root	_	_	1:LVC.full
5	ou	ou	CCONJ	_	_	8	cc	_	_	*
6	avez	avoir	AUX	_	_	8	aux	_	_	*
7	récemment	récemment	ADV	_	_	8	advmod	_	_	*
8	présenté	présenter	VERB	_	_	4	conj	_	_	2:LVC.full
9	un	un	DET	_	_	10	det	_	_	*
10	saignement	saignement	NOUN	_	_	4	obj	_	_	1;2

Corpus quality

- Single-annotated for most languages
- Consistency checks for 19 languages
- Inter-annotator agreement on a sample
  - Around 150 to 2,500 sentences
  - Identification **IAA**: 0.227 ES $\leqslant \kappa_{span} \leqslant$ 0.984 TR
  - Categorization **IAA**: 0.573 ES $\leqslant \kappa_{cat} \leqslant$ 1.000 HU FA
  - Macro-average higher than edition 1.0
    - $\kappa_{span} = 0.58 \rightarrow \kappa_{span} = 0.691$
    - $\kappa_{cat} = 0.819 \rightarrow \kappa_{cat} = 0.836$

Shared task

Goal

Automatically identify all VMWE occurrences in running text.

Two tracks

- **Closed**: only use the provided training/dev data
- **Open**: use the provided data + any external resource
  - corpora, lexicons, grammars, language models, ...

Evaluation

- Based on **identification** only
- Categorization quality is reported on but not ranked

Evaluation measures (as in edition 1.0)

Per-language system evaluation

- Compare prediction and gold standard
- Precision, recall and F1-measure

MWE-based scores

Only predictions with the **perfect span** are considered to match

Token-based scores

- Allows **partial matches**
- We consider all partial bijections from gold to system VMWEs
- The partial bijection **maximizing the system score** is chosen

Evaluation measures (new in edition 1.1)

Cross-lingual macro-averages

- Token-based and MWE-based scores
- Phenomenon-specific scores
  - MWE-based P/R/F1 for a subset of prediction & gold standard
  - Only VMWEs that represent a given phenomenon
    - Continuous vs. discontinuous
    - Multi-token vs. single-token
    - Seen vs. unseen (wrt. training corpus)
    - Identical vs. variants (wrt. training corpus)

Submitted systems

- 12 teams (vs. 7 in edition 1.0)
  - From France (3), Germany (3), Ireland (1), Italy (1), Romania (1), Switzerland (1), Turkey (1), UK (1)
- 17 system submissions
  - 13 closed track + 4 open track
  - 16/17 submissions cover 3 or more languages
  - 11/17 submissions cover 19 languages

Techniques

Neural networks	Parsing	CRF	Stat. measures	Naive Bayes
Deep-BGT GBD-NER mumpitz SHOMA TRAPACC Veyn	Milos MWETreeC TRAVERSAL	CRF-DepTreecateg CRF-Seq-noncateg	Polirem	varIDE

Some results (more ▸ online) I

- Average MWE-based scores

submission	track	P	R	F1
TRAVERSAL	closed	67.58	44.97	**54.00**
TRAPACC-S	closed	62.28	41.40	49.74
TRAPACC	closed	55.68	44.67	49.57
CRF-Seq-nocategs	closed	56.13	39.12	46.11
SHOMA	open	66.08	51.82	**58.09**

- System strengths:
  - TRAVERSAL: Slavic and Romance languages
  - TRAPACC: German and English
  - CRF-Seq-nocategs: Hindi

Some results (more ▸ online) II

Languages

- "Easiest" languages: Hungarian (F1=90.31) and Romanian (F1=85.28) ⇒ largest training corpora
- "Hardest" languages: Hebrew (F1=23.28), English (F1=32.88) and Lithuanian (F1=32.17) ⇒ smallest training corpora
- Outlier: Hindi (F1=72.98) ⇒ MVCs are "easy"

Phenomenon-specific scores

- Discontinuous, variant and unseen VMWEs are much harder
- Variants: average recall, high precision (R_{max}=.56, P_{max}=.86)
- Unseen: low recall and precision (R_{max}=.38, P_{max}=.32)

Conclusions

- Benchmark results
- Outcomes: freely available guidelines, corpora
- 12 teams, 17 submissions, all are multilingual
- There is room for improvement
- Findings:
  - Inherently lexical nature of the phenomenon
  - Unseen VMWEs are harder to generalize over than other unseen entities (e.g. NEs)
  - VMWE identification and discovery should go hand-in-hand

Future work

- Continuous enhancement of guidelines and corpora (quality, size)
  - IAV feedback and status
  - New languages and language families
- Future shared task editions
  - New MWE categories (e.g. nominal)
  - Joint MWE identification and parsing or NER
- Synergies with other multilingual initiatives (e.g. UD)

Acknowledgements

- Funding agencies and projects:
  COST action PARSEME (IC1207)
  Czech Republic LD-PARSEME (LD14117)
  ANR PARSEME-FR (ANR-14-CERA-0001)
  Marie Skłodowska-Curie (Grant 713567)
  Science Foundation Ireland (Grant 13/RC/2106)
  Deutsche Forschungsgemeinschaft (CRC 991)
  Ministry of Human Capacities, Hungary (UNKP-17-4 New National Excellence Program)
  Slovenian Research Agency (J6-8256 project)
  DST-CSRI, Govt. of India
  Boğaziçi University Research Fund (Grant 14420)
- FLAT development
  - **Maarten van Gompel**, Radboud University, The Netherlands

Annotation teams

Balto-Slavic languages: (BG) Ivelina Stoyanova (LL), Tsvetana Dimitrova, Svetlozara Leseva, Valentina Stefanova, Maria Todorova; (HR) Maja Buljan (LL), Goranka Blagus, Ivo-Pavao Jazbec, Kristina Kocijan, Nikola Ljubešić, Ivana Matas, Jan Šnajder; (LT) Jolanta Kovalevskaitė (LL), Agnė Bielinskienė, Loic Boizou; (PL) Agata Savary (LL), Emilia Palka-Binkiewicz; (SL) Polona Gantar (LL), Simon Krek (LL), Špela Arhar Holdt, Jaka Čibej, Teja Kavčič, Taja Kuzman.

Germanic languages: (DE) Timm Lichte (LL), Rafael Ehren; (EN) Abigail Walsh (LL), Claire Bonial, Paul Cook, Kristina Geeraert, John McCrae, Nathan Schneider, Clarissa Somers.

Romance languages: (ES) Carla Parra Escartín (LL), Cristina Aceta, Héctor Martínez Alonso; (FR) Marie Candito (LL), Matthieu Constant, Carlos Ramisch, Caroline Pasquer, Yannick Parmentier, Jean-Yves Antoine, Agata Savary; (IT) Johanna Monti (LL), Valeria Caruso, Maria Pia di Buono, Antonio Pascucci, Annalisa Raffone, Anna Riccio; (RO) Verginica Barbu Mititelu (LL), Mihaela Onofrei, Mihaela Ionescu; (PT) Renata Ramisch (LL), Aline Villavicencio, Carlos Ramisch, Helena de Medeiros Caseli, Leonardo Zilio, Silvio Ricardo Cordeiro.

Other languages: (AR) Abdelati Hawwari (LL), Mona Diab, Mohamed Elbadrashiny, Rehab Ibrahim; (EU) Uxoa Inurrieta (LL), Itziar Aduriz, Ainara Estarrona, Itziar Gonzalez, Antton Gurrutxaga, Ruben Urizar; (EL) Voula Giouli (LL), Vassiliki Foufi, Aggeliki Fotopoulou, Stella Markantonatou, Stella Papadelli; (FA) Behrang QasemiZadeh (LL), Shiva Taslimipoor; (HE) Chaya Liebeskind (LL), Yaakov Ha-Cohen Kerner (LL), Hevi Elyovich, Ruth Malka; (HI) Archna Bhatia (LL), Ashwini Vaidya (LL), Kanishka Jain, Vandana Puri, Shraddha Ratori, Vishakha Shukla, Shubham Srivastava; (HU) Veronika Vincze (LL), Katalin Simkó, Viktória Kovács; (TR) Tunga Güngör (LL), Gözde Berk, Berna Erden.