PARSEME Corpus Release 1.3

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Menghan Jiang, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Alexandra Butler, Marie Candito, Apolonija Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaité, Simon Krek, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze, Abigail Walsh

PARSEME

MWE 2023 workshop @ EACL 2023, Dubrovnik, Croatia

PARSEME corpus

- Collective effort towards **annotation guidelines** for verbal MWEs (VMWEs)
    - Unified across many languages from various genera
- 5 VMWE major **categories**, annotation **decision diagrams**, multilingual **examples**
- **Corpora** in **26 languages** annotated according to these guidelines
- **Shared tasks** on automatic identification of verbal MWEs
- **Editions** 1.0 (2017), 1.1 (2018) and 1.2 (2020):
    - Guidelines, corpora and shared task tightly **intertwined**
    - Overlapping but varying language lists
    - **Morpho-syntactic** annotation - manual vs. automatic, heterogeneous sources
    - Increasing compatibility with **Universal Dependencies (UD)**

VMWE categories in PARSEME

- Universal
  - VID (verbal idiom) e.g. (de) ***schwarz fahren*** (lit. ‘black drive’) ‘take a ride without a ticket’
  - LVC (light-verb construction)
    - LVC.full, e.g. (hr, sr) ***držati govor*** (lit. ‘hold a speech’) ‘give a talk’
    - LVC.cause, e.g. (ro) ***da bătăi de cap*** (lit. ‘give strikes of head’) ‘give a hard time’
- Quasi-universal
  - IRV (inherently reflexive verbs), e.g. (pt) ***se queixar*** ‘complain’
  - VPC (verb-particle construction)
    - VPC.full, e.g. (en) ***do in***
    - VPC.semi, e.g. (en) ***eat up***,
  - MVC (multi-verb construction), e.g. (fr) ***laisser tomber*** (lit. ‘let fall’) ‘give up’.
- Language-specific categories
  - ICV (inherently clitic verb): (it) ***smetterla*** (lit. ‘quit it’) ‘knock it off’
- Experimental
  - IAV (inherently adpositional verbs), e.g. (es) ***entender de*** *algo* (lit. ‘understand of something’) ‘know about something’

PARSEME corpus - objectives for edition 1.3

- Gather all past **26 languages** in the same release
- Cover **new languages**
- Achieve full **UD compatibility**
- **Detach** the corpus releases from shared tasks
- Define a process of **continuous improvement** and systematic releasing (following the UD model)

New languages

- Arabic
  - Examples added to the guidelines
  - Covered in previous annotation campaigns but the corpus itself is not available
  - New corpus created from scratch
  - Built upon the Prague Arabic Dependency Treebank (PADT) (Hajic et al., 2004)
  - 7,500 sentences; 4,700 VMWEs
  - Single annotator per sentence; double-annotated fraction for IAA calculation
- Serbian
  - Examples added to the guidelines
  - Morphosyntactic layers generated with UDPipe (Straka, 2018)
  - 3,586 sentences; 1,300 VMWEs
  - Single annotator per sentence

PARSEME

Enlarged corpora

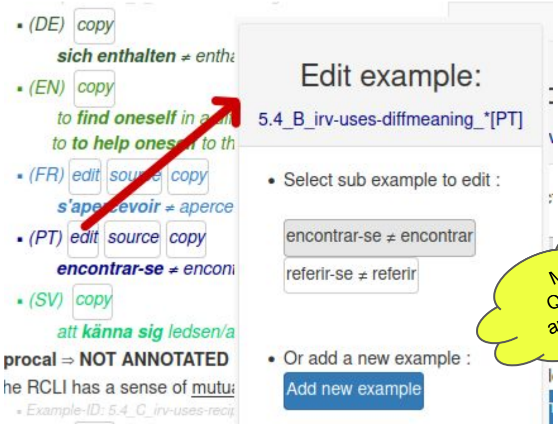
- Greek
  - 5,000 new sentences; also informal register
- Swedish
  - 1,700 new sentences; full parallelism with the UD Talbanken treebank
  - Extensive use of the consistency checker
- Chinese
  - 9,000 new sentences
  - Double annotation + adjudication for each sentence

上海深圳

Other enhancements

- Quality enhancements in particular languages
  - Croatian - alignment with gold UD annotations
  - Romanian - new category annotated (IAVs)
  - English and Polish - thorough consistency and quality checks
  - Irish - controversial category (IRV) removed
  - Turkish - manual revision of morphosyntax
  - Czech and Maltese - partial upgrade from version 1.0 to 1.3
- Full UD compatibility:
  - 11 languages: synchronisation of manual UD layers with UD release 2.11
  - 16 languages: re-generation of automatic morphosyntactic layers with UDPipe 2.10
  - All 26 corpora now use **UD 2 tagsets**
- Corpus re-split
  - Adopting shared task 1.2 strategy (controlled number of unseen VMWEs in test and dev)

Enhanced infrastructure



- (DE) copy
  sich enthalten ≠ entha
- (EN) copy
  to **find oneself** in a
  to **to help ones** to th
- (FR) edit source copy
  s'apercevoir ≠ aperce
- (PT) edit source copy
  encontrar-se ≠ encon
- (SV) copy
  att **känna sig** ledsen/a

procal ⇒ NOT ANNOTATED

he RCLI has a sense of mutu

- Example-ID: 5.4_C_irv-uses-recip

Edit example:

5.4_B_irv-uses-diffmeaning_*[PT]

- Select sub example to edit :

encontrar-se ≠ encontrar

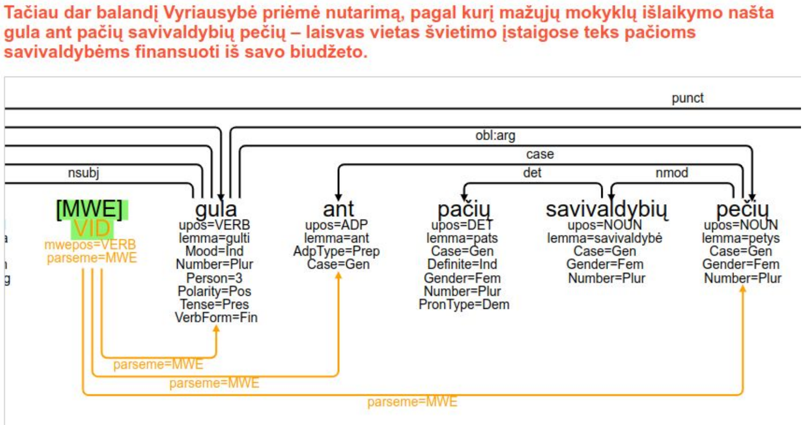
referir-se ≠ referir

- Or add a new example :

Add new example

Many thanks to Quentin Barrouyer and Baptiste Souche

- Annotation guidelines
  - Easier edition of multilingual examples
  - New examples added (2,000 in total in edition 1.3)
- Versioning via a common Gitlab project
- Rich Wiki documentation of the corpora, procedures and tools
- Grew-match corpus browser - one instance per corpus version (Guillaume, 2021)

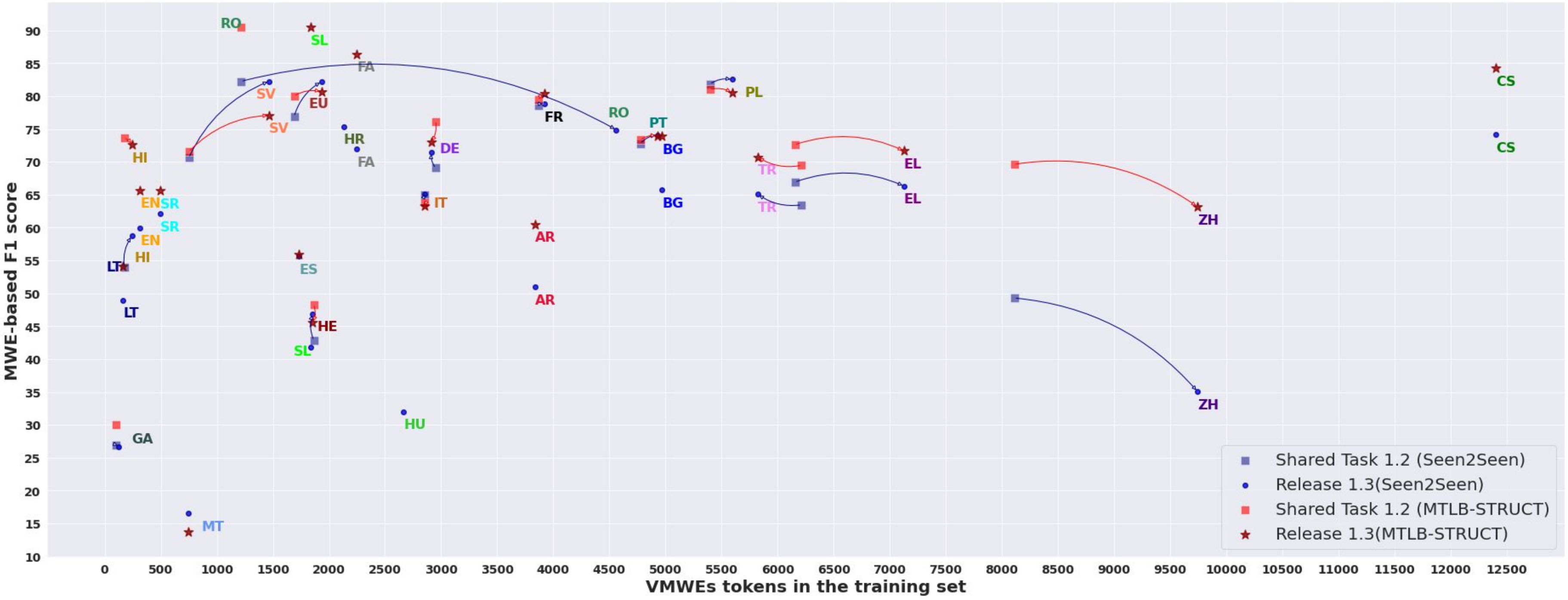


Tačiau dar balandį Vyriausybė priėmė nutarimą, pagal kurį mažųjų mokyklų išlaikymo našta gula ant pačių savivaldybių pečių – laisvas vietas švietimo įstaigose teks pačioms savivaldybėms finansuoti iš savo biudžeto.

VMWE identification

- Task: automatically annotating VMWE occurrences in running text
  - Addressed by the PARSEME shared tasks 1.0, 1.1, 1.2
- Critical hardness of the items unseen in training data
- 2 winner systems of the PARSEME shared task 1.2
  - Seen2Seen (Pasquer et al., 2020) - rule-based, fully interpretable, fast training
  - MTLB-STRUCT (Taslimipoor et al., 2020) - BERT-based, single- or multi-tasking, long training
- Both re-trained on the 1.3 release (after re-split)

Corpus sizes vs. system results



Conclusions and future work

- 1.3 release corpus with all past 26 languages
- 9 million tokens; 455,000 sentences; 127,000 VMWE annotations
- Full UD compatibility
- Universality confirmed for VIDs and LVC.full
- VMWE identification remains challenging despite larger and better corpora
- Next steps:
  - Stronger automation in the spirit of CI/CD (ongoing)
  - Extending the guidelines to other MWE categories
  - Stronger convergence with UD