PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, Joakim Nivre

NEJLT paper (2023)
https://doi.org/10.3384/nejlt.2000-1533.2023.4453

PARSEME in a nutshell

- Unified multilingual guidelines for verbal MWEs

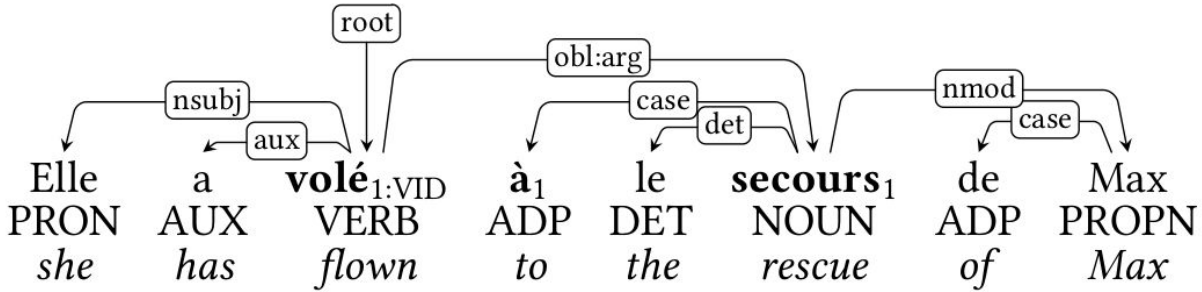
Elle	a	**volé**	**à**	le	**secours**	de	Max
She	have.3SG	fly.PTCP	to	the	rescue	of	Max
		1.VID	1		1		

- Annotated corpora in 26 languages
- 160 collaborators

PARSEME

UD in a nutshell

- Unified multilingual guidelines for morphosyntax (POS, morphological features, syntactic dependencies)



- 100 annotated corpora in 100 languages
- 300 collaborators

UD - PARSEME common objective: universality

- Cross-linguistically consistent and applicable language descriptions
- Similar phenomena – represented in a unified way
- Language-specific categories, relations and guidelines are allowed

State of affairs

- UD and PARSEME: common goals, but independent initiatives
  - Inconsistent terminologies
  - Competing methods
  - Divergent annotations
  - Low cross-lingual consistency

- Goals:
  - Greater **convergence** between UD and PARSEME processes and resources
  - Define a **roadmap** towards unification: short- mid- and long-term proposals
  - Keep morphosyntactic annotations as **independent** as possible from MWE annotations
    - How to distinguish morphosyntax from MWEs?

Dimensions of idiosyncrasy

MWE idiosyncrasy: Occurrences vs. Types

Occurrence idiosyncrasy

- Defective property

na oścież
on 'oścież'
'wide (open)'

um **deus nos acuda**
a god us.ACC help.IMP.2.SG
lit. 'a god-help-us' | 'a mess'

Elle **a beau** pleurer.
she has pretty.M cry.INF
lit. 'She has pretty to cry.' | 'She cries in vain.'

Type idiosyncrasy

- Restrictive property

She **knows** her **stuff**.
'She is skilled.'

#She knows my stuff

a **întoarce foaia**
to turn sheet.DEF
lit. 'to turn the sheet' | 'to become harsher'

#a întoarce foile
to turn sheet.PL.DEF
'to turn the sheets'

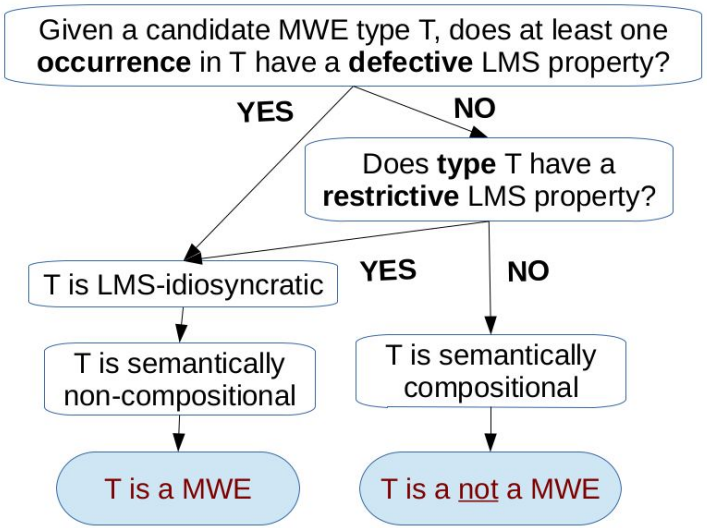
MWE idiosyncrasy: Morphosyntactic vs. Semantic

Morphosyntactic idiosyncrasy

- Previous examples

Semantic idiosyncrasy

- Prototypical of MWEs
- Hard to operationalise
- Approximated by morphosyntactic idiosyncrasy



MWE annotation in PARSEME and UD

Descriptions

UD

- segmentation, lemmas, morphology and syntax
- unitizing
- full coverage of the words in the corpus

PARSEME

- semantic (approx. by morphosyntax)
- unitizing
- sporadic
- nesting:

  [[let]$_2$ the cat [out]$_2$ of the bag]$_1$

- overlaps:

  take$_{1,2}$ a walk$_1$ and a shower$_2$

Data format

- UD: CoNLL-U format
- PARSEME: CUPT (an extended CoNLL-U file format)

Fr. *Elle a volé au secours de Max.*

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC PARSEME:MWE
1     Elle       il       PRON   _  Gender=Fem|Number=Sing|Person=3        3  nsubj  _ _  *
2     a          avoir    AUX    _  Mood=Ind|Number=Sing|Person=3|…       3  aux    _ _  *
3     volé       voler    VERB   _  Gender=Masc|Number=Sing|Tense=Past|…  0  root   _ _  1:VID
4-5   au         _        _      _  _                                     _  _      _ _  *
4     à          à        ADP    _  _                                     6  case   _ _  1
5     le         le       DET    _  Definite=Def|Gender=Masc|Number=Sing|… 6  det    _ _  *
6     secours    secours  NOUN   _  Gender=Masc|Number=Sing               3  obl:arg _ _  1
7     de         de       ADP    _  _                                     8  case   _ _  *
8     Max        Max      PROPN  _  _                                     6  nmod   _ _  *
```

Words and tokens

- Word – token:
  - Word = token: Fr. *Elle*
  - More words = one token (multiword token): Fr. *au* (*à le*)
  - One word = more tokens (multitoken word): 20_000
- Word – basic notion for UD and PARSEME:
  - UD: basic unit of analysis; PARSEME defines a MWE as containing at least 2 words
  - PARSEME relies on UD split of tokens into words
  - PARSEME covers a higher number of multiword tokens than UD => inconsistency:

2	sollst	sollen	…	*
3	**aufpassen**	aufpassen	…	1:VPC
…				
11	Haupt**rolle**	Hauptrolle	…	1:LVC.full
12	**spielen**	spielen	…	1

Morphology and syntax

UD:

- 17 universal POS tags and over 200 values for morphological features
- Dependency syntax: 37 universal syntactic relations

  + 26 subtypes thereof (language(s)-specific) – optional
  => inconsistencies among treebanks (for the same / different languages)

- Lexicalist principle: content vs. function words

PARSEME:

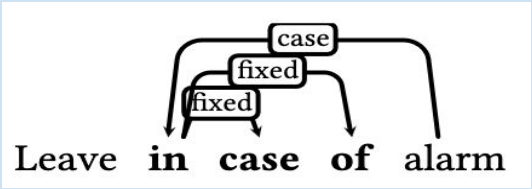
- approximates semantic compositionality by lexical and morphosyntactic flexibility tests that are driven by syntactic structure => strong dependence on the underlying syntactic framework (UD)
- Lexicalist principle => weakly connected dependency graph

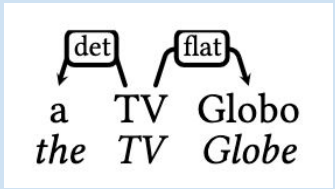
  BUT: MWEs headed by copula *be* do not obey the VMWE definition: verbal head

=> universality of UD => universality of PARSEME: all corpora in PARSEME v1.3 are UD compatible

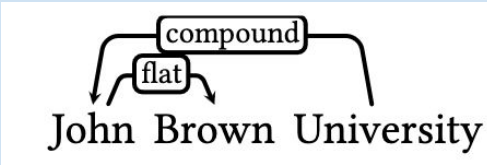
UD MWEs relations

Fixed - *grammaticalized* expressions, considered *headless* (in synchrony), mainly function words





Flat - *headless* semi-fixed expressions, like names or complex numerals



Compound - word-level compounding; *headed*

PARSEME - UD verbal MWEs

PARSEME MWE categories	**UD** MWE relations
IRV	obj or expl (with **expl:pv** for idiomatic cases)
VPC	more inclusive subrelation **compound:prt**
MVC	more inclusive subrelation **compound:mvc**
LVC	`obj` or `compound:lvc`
VID	—

Towards UD/PARSEME unification

PARSEME/UD unification roadmap

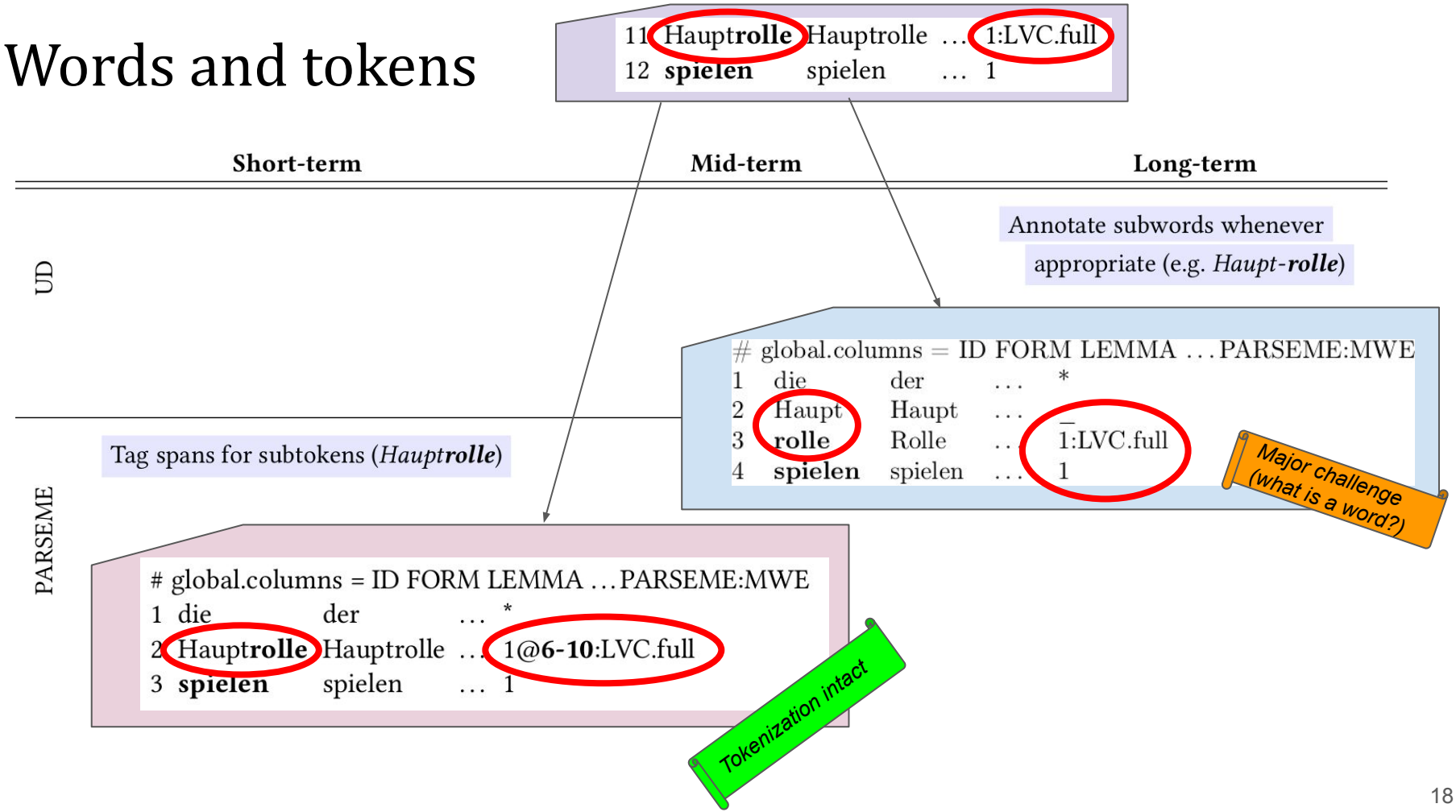
light blue - some languages

dark blue - all languages

	Short-term	Mid-term	Long-term
UD	Assume idiosyncrasy of MWEs Don't use *MWE* as umbrella term for **`fixed`**, **`compound`** and **`flat`**	Use the `.cupt` format Merge **`fixed`** with **`flat`**, maybe rename to **`headless`** Abandon **`compound:lvc`** and **`expl:pv`** In new annotations, only flag token idiosyncrasy	Annotate subwords whenever appropriate (e.g. *Haupt-**rolle***) Extend the annotation schema to subsystems
PARSEME	Tag spans for subtokens (*Haupt**rolle***) Rename VPCs to IVPCs	Guidelines for all syntactic types of MWEs, with subtypes for totally fixed MWEs Define the border between named entities and MWEs Annotate MWEs of all syntactic types Flag both token and type idiosyncrasy	Link corpora with MWE lexicons, encode MWE type properties in the lexicons Use orthogonal typology-inspired categories Extend the annotation schema to constructions

No re-annotation in UD in the first 2 stages

Words and tokens



Terminology and guidelines

Short-term

UD

Assume idiosyncrasy of MWEs
Don't use *MWE* as umbrella term for
`fixed`, **`compound`** and **`flat`**

Annotations intact

```
# global.columns = ID FORM ... HEAD DEPREL ... MWE NE
31 a      ... 32 det    ... *   *
32 TV     ... 34 nsubj  ... *   1:ORG
33 Globo  ... 32 headless ... *   1
```

Major challenge

Extend the annotation schema to subsystems

PARSEME

Rename VPCs to IVPCs

Fully automatic

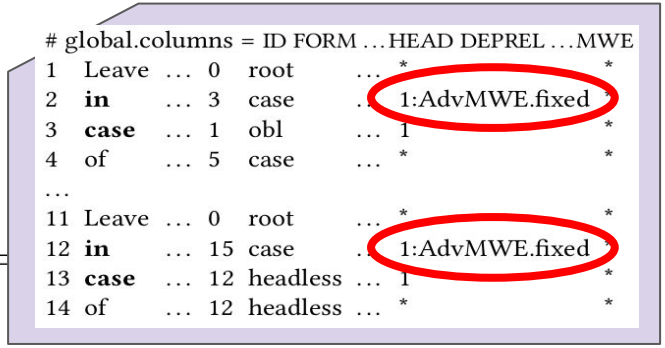
Guidelines for all syntactic types of MWEs, with subtypes for totally fixed MWEs

Define the border between named entities and MWEs

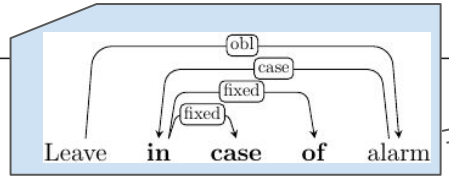
Annotate MWEs of all syntactic types

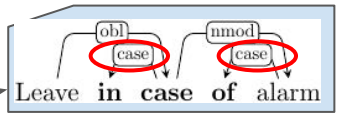
Major challenges

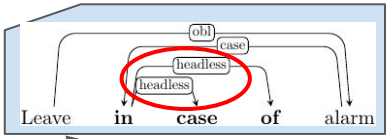
Occurrences vs. types



	Short-term	Mid-term
UD	Mostly automatic	Use the `.cupt` format Merge **`fixed`** with **`flat`**, maybe rename to **`headless`** Abandon **`compound:lvc`** and **`expl:pv`** In new annotations, only flag token idiosyncrasy
PARSEME	Consistent with past practice	Flag both token and type idiosyncrasy







Holy Grail

Link corpora with MWE lexicons,
encode MWE type properties
in the lexicons
Use orthogonal typology-inspired
categories
Extend the annotation schema
to constructions

Wrap-up

- PARSEME and UD agree on universality and diversity objectives
- Currently partial compatibility in annotation principles
- 3-step roadmap for stronger convergence
- Insight from typology experts most welcome
- Caveat: delicate balance between
  - existing/upcoming data
  - automation tools
  - willingness of contributors