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
agata.savary@universite-paris-saclay.fr

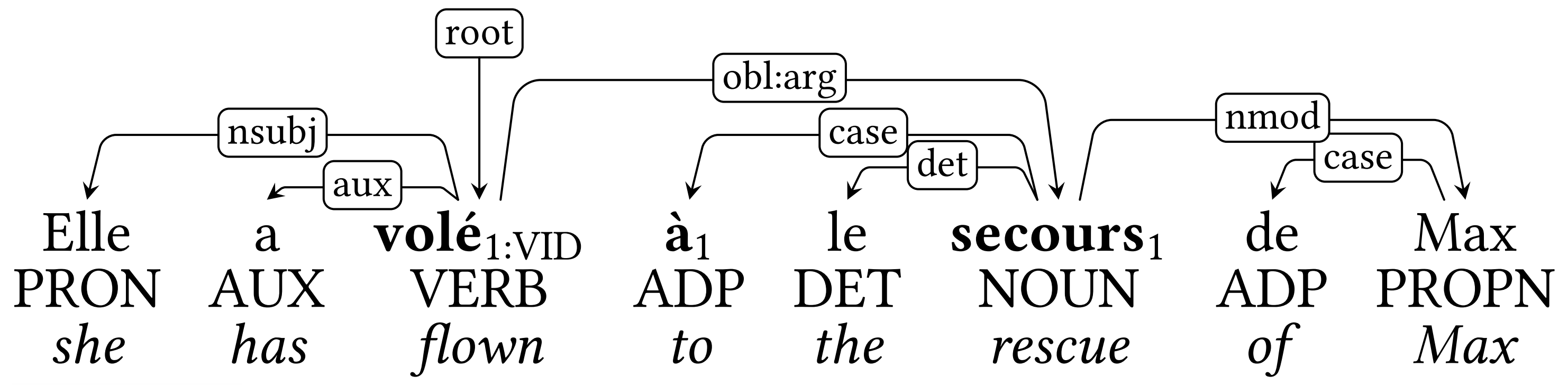
NEJLT 9(1), 2023

GURT 2023
SYNTAXFEST
Washington, D.C.

UD 🤝 PARSEME

(dependency syntax) (multiword expressions)

130+ languages — goal of universality — 26 languages



Annotation of a French verbal idiom (VID) with discontinuity

```
# 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    _ _ *
```

PARSEME's .cupt format: UD CoNLL-U + a layer for MWEs

This paper: opportunities & challenges for unifying these frameworks. We offer short-, medium-, and long-term recommendations.

Challenge #1: Scope of what is identified as a *multiword expression* (MWE)

• In MWE community, defined in terms of morphosyntactic and/or semantic *idiosyncrasy*. PARSEME has developed rigorous crosslinguistic guidelines + corpora for categories of verbal MWEs: **Inherently Reflexive Verbs**, **Verb-Particle Constructions**, **Multi-Verb Constructions**, **Light Verb Constructions**, **Verbal Idioms**.

• In UD guidelines, used loosely as a cover term for **fixed**, **flat**, **compound** relations (+ in some languages, subtypes like **compound:lvc**, **expl:pv**). But not all compounds are idiosyncratic.

➡ ***short term:*** dispense with the casual use of "MWE" in the UD guidelines

➡ ***medium term:*** extend PARSEME work beyond verbal MWEs to include nominal MWEs, multiword connectives, etc.; consider relationship to named entities

➡ ***long term:*** UD: better guidelines for *productive grammatical subsystems* like templatic named entities, numbers, measurements, dates; PARSEME: partially productive constructions (as in Construction Grammar)

Challenge #2: UD tokenization is sometimes too coarse to capture idiomatic combinations (e.g., synthetic compounds).

Haupt**rolle spielen**
head.role play
'to play the leading role'

언어**에** **대해** 읽다
language:**POSTP** **about** read
'read **about** languages'

German & Korean examples where the ideal MWE annotation is more granular than UD syntactic words

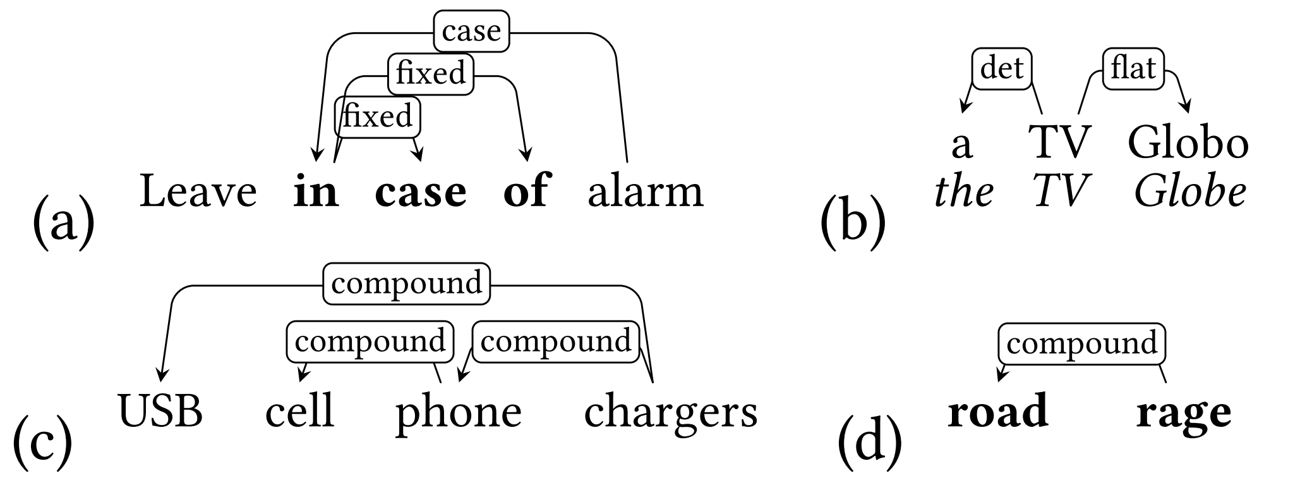
➡ ***short term:*** indicate subword character spans in PARSEME annotation

➡ ***long term:*** implement a finer-grained notion of word in UD. Splitting synthetic compounds would also disambiguate cases like Swedish *bildrulle*: *bil+drulle* 'car maniac (bad driver)' vs. *bild+rulle* 'picture roll (roll of film)'.

Challenge #3: Idiosyncrasy at the *lexical type* level is not always reflected at the *token (occurrence)* level.

• MWEs can have regular syntax, even if the meaning is idiomatic and the variability of the type is restricted (fossilization).

• UD mostly targets token-level analysis, and is agnostic to type-level variability or meaning. But this is muddled by labels like **fixed**, **compound:lvc**, **expl:pv**, **compound:prt** vs. **advmod**.



Regular vs. irregular syntax in UD (English & Portuguese)

➡ ***medium term:*** disentangle things like **:lvc** and **:pv**, which are MWE classifications, from the syntax by moving them to an MWE layer; address inconsistencies in some of the other deprels

➡ ***medium term:*** merge **fixed** and **flat** under a new label, **headless**?

➡ ***long term:*** link token occurrences in corpora to entries in a lexicon