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

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

Agata Savary, University of Paris-Saclay, CNRS, LISN, France agata.savary@universite-paris-saclay.fr
Sara Stymne, Uppsala University, Sweden sara.stymne@lingfil.uu.se
Verginica Barbu Mititelu, Romanian Academy Research Institute for Artificial Intelligence, Romania vergi@racai.ro
Nathan Schneider, Georgetown University, USA nathan.schneider@georgetown.edu
Carlos Ramisch, Aix Marseille Univ, CNRS, LIS, Marseille, France carlos.ramisch@lis-lab.fr
Joakim Nivre, Uppsala University & RISE Research Institutes of Sweden, Sweden joakim.nivre@lingfil.uu.se

Abstract Multiword expressions (MWEs) are challenging and pervasive phenomena whose idiosyncratic properties show notably at the levels of lexicon, morphology, and syntax. Thus, they should best be annotated jointly with morphosyntax. In this position paper we discuss two multilingual initiatives, Universal Dependencies and PARSEME, addressing these annotation layers in cross-lingually unified ways. We compare the annotation principles of these initiatives with respect to MWEs, and we put forward a roadmap towards their gradual unification. The expected outcomes are more consistent treebanking and higher universality in modeling idiosyncrasy.

1 Introduction

Multiword expression (MWE) is an umbrella term spanning a range of linguistic phenomena whose common property is *idiosyncrasy* or, more specifically, *idiomaticity*, which may manifest in many different respects: lexical, morphological, syntactic, semantic, pragmatic, and statistical (Baldwin and Kim, 2010).

MWEs are challenging and pervasive. For instance, in an MWE-annotated corpus of French (Candito et al., 2021), over 11% of all tokens belong to MWEs. Moreover, MWEs likely exist in any natural language. Therefore, modeling idiosyncrasy in language resources and tools is a natural quest. This position paper addresses two language annotation frameworks, Universal Dependencies and PARSEME, from the point of view of MWEs.

Universal Dependencies[1] (UD; Nivre et al., 2016, 2020; de Marneffe et al., 2021) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across many languages. It is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. PARSEME[2] (Savary et al., 2018; Ramisch et al., 2020) is a scientific network which evolved from a homonymous COST action dedicated to parsing and MWEs. One of its major outcomes is a multilingual corpus annotated for verbal MWEs (VMWEs) in 26 languages by over 160 native annotators.

The common objective of UD and PARSEME is *universality*, i.e., the development of cross-linguistically consistent and applicable language descriptions. Such consistency leads to valuable

[1] https://universaldependencies.org/
[2] https://gitlab.com/parseme/corpora/

insights about linguistic phenomena (including idiosyncrasy), contributes to contrastive studies, and promotes progress in NLP across many languages. Concretely, both UD and PARSEME (i) develop cross-lingually unified and continuously enhanced annotation guidelines, (ii) annotate, enhance, and release corpora on the basis of these guidelines, and (iii) use these corpora to develop NLP tools for syntactic parsing and MWE identification.

Despite their common goals, UD and PARSEME have operated relatively independently, ending up with partly divergent and competing terminologies and methods. Some of the MWE types addressed by PARSEME, such as light-verb constructions, are annotated to some extent also within UD, but typically not consistently across languages, as we will discuss in Section 3.6. We think it is desirable to keep morphosyntactic annotations separate from MWE-related annotations.[3]

The desire for greater convergence between UD and PARSEME practices has steadily grown as the initiatives have matured. PARSEME has relied on the UD format (cf. Sec. 3.2) and data in its latest corpus releases. In August 2021, a joint Dagstuhl Seminar on *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics* brought together the two initiatives (Baldwin et al., 2021).[4] Finally, September 2022 saw the start of a new COST action entitled UniDive (*Universality, Diversity and Idiosyncrasy in Language Technology*), with UD/-PARSEME unification on the agenda.

This paper aims at providing a roadmap towards this unification. We first survey the dimensions of MWE idiosyncrasy (Sec. 2) and compare the two frameworks' annotation principles that bear on MWEs (Sec. 3). Then, we offer short-, mid- and long-term proposals for adjusting the frameworks, paving the way towards eventually unifying them (Sec. 4). Sec. 5 concludes with future perspectives.

[3]The current status led to problems for the VMWE identifiers evaluated in the PARSEME shared tasks (Ramisch et al., 2020), which were given UD morphosyntactic annotations as input, and were expected to predict VMWE annotations. Since some MWE-related phenomena currently are annotated in the morphosyntactic layer, this type of evaluation is biased (since part of the information to be predicted is already given as input).

[4]`https://www.dagstuhl.de/seminars/seminar-calendar/seminar-details/21351`

2 Dimensions of Idiosyncrasy in MWEs

MWEs deviate from compositionality norms, as seen in the examples from the PARSEME languages below. The MWE in (1) contains a cranberry word *oścież*, i.e. a token having no status of a standalone word but only occurring in a MWE.[5]

(1) **na oścież** (pl)
on 'oścież'
'wide (open)'

The MWE in (2) is exocentric, since it is a nominal phrase whose head is a finite-form verb.

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

In (3), the verb is modified by an adjective and an infinitive, which is not a regular syntactic structure.

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

In (4), the possessive *her* must agree with the subject, otherwise the MWE is understood literally, as in (5).

(4) She **knows** her **stuff**. (en)
'She is skilled.'

(5) #She knows my stuff (en)

Concrete nouns in verb-object constructions can inflect for number, but pluralizing the noun in (6) implies losing the idiomatic reading, as shown in (7).

(6) a **întoarce foaia** (ro)
to turn sheet.DEF
lit. 'to turn the sheet' | 'to become harsher'

(7) #a întoarce foile (ro)
to turn sheet.PL.DEF
'to turn the sheets'

Given these examples, MWE idiosyncrasy can be considered along two orthogonal dimensions.

[5]Examples follow the PMWE conventions (Markantonatou et al., 2021). POS and morphological features use UD. We use the IETF BCP-47 standardized language codes in all examples.

Occurrences vs. Types Some idiomatic properties of MWEs display at the level of individual occurrences of MWEs (Savary et al., 2019). Conversely, others are visible at the level of *types*, that is, sets of surface realizations of the same MWE. For instance, the cranberry word (1), irregular agreement (2), and irregular syntax (3) can be observed in every single occurrence of these MWEs. On the other hand, compulsory agreement (4) or restricted inflection (6) can only be attested while considering several possible surface realizations of the given MWE, so as to test whether different inflection, agreement or syntactic alternations do or do not preserve the idiomatic reading.

Lichte et al. (2019) propose a different but isomorphic terminology, contrasting restrictive vs. defective idiosyncrasy. A *defective* property excludes a literal interpretation of a given MWE. This is observable precisely at the level of individual MWE occurrences, as in (1)–(3). A *restrictive* property reduces the number of possible surface realizations of a given MWE relative to the corresponding literal interpretation. This amounts to idiosyncrasy at the level of MWE types, as in (4)–(6).

Morphosyntactic vs. Semantic Idiosyncrasy The idiosyncratic properties discussed above occur at the morphosyntactic level. However, the most salient property of MWEs is *semantic non-compositionality*: their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a way deemed regular (Sag et al., 2002). Examples (1)–(6) can safely be considered as semantically non-compositional.

Distinguishing morphosyntactic from semantic idiosyncrasy is a hard nut to crack. First, the borders between morphology, syntax and semantics are fuzzy. For instance, the notions of syntactic and semantic arguments are closely related in the linguistic debate about arguments vs. adjuncts (Przepiórkowski and Patejuk, 2018). Second, idiosyncratic properties in MWEs usually cross multiple layers of linguistic description. For instance, the MWE in (3) exhibits not only unusual syntax but also restricted inflection, as in (6). Third, semantic non-compositionality is hard to test directly and reliably at the level of occurrences. Nonetheless, it can be more accurately approximated by lexical and morphosyntactic inflexibility, by testing it at the level of types (Gross, 1988; Gibbs and Nayak, 1989). This again suggests that morphosyntactic and semantic idiosyncrasies are entangled.

Kahane et al. (2017) propose considering syntactic and semantic idiosyncrasy as separate dimensions. They consider: (i) regular constructions, subsystems and irregular constructions, (ii) compositional, semi-compositional and non-compositional expressions, along the syntactic and semantic axes. Various expressions are then placed in this two-dimensional space. For instance, syntactically irregular constructions can be semantically compositional, e.g. (fr) ***peser lourd*** (lit. 'to weigh heavy') 'to be very heavy'. While this classification is promising, it fails to provide an operational definition of semantic non-compositionality. In particular, assuming that formal semantics accurately approximates semantic compositionality, there can be no constructions with irregular syntax but compositional semantics.[6] Still, what we retain from Kahane et al. (2017) is the premise that syntactic and semantic properties of MWEs should be annotated at different layers as much as possible. In particular, it is useful to display regular syntax in MWEs despite their semantic idiosyncrasy.

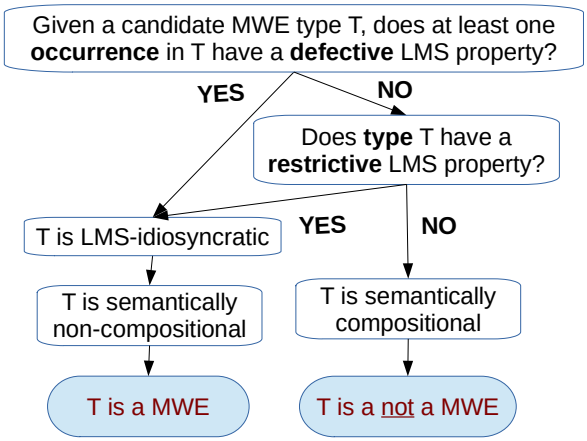

Figure 1: Implications among lexical and/or morphosyntactic (LMS) and semantic idiosyncrasy of MWE occurrences and types.

In short, we distinguish occurrence vs. type and lexical/morphosyntactic vs. semantic idiosyncrasies in MWEs, but we note that these dimensions are closely linked, as shown in Fig. 1. First, if at least one MWE occurrence is idiosyncratic,

[6] Specific compositional semantic procedures are assigned to syntactic structures deemed regular (Steedman, 2000).

then the whole type is irregular. Second, lexical and/or morphosyntactic idiosyncrasy of MWE occurrences and/or types approximates their semantic non-compositionality. Note that the choice of testing defectiveness (of an occurrence) before restrictiveness (of the whole type) is not arbitrary. First, basic observable units in an annotated corpus are occurrences (by contrast, lexicons primarily focus on types). Second, testing irregularity for an occurrence is cognitively easier than regarding the whole type. Third, the definition of a restrictive property is based on the understanding of the literal interpretation of a potential MWE. However, if a token is defective, its literal interpretation is excluded. Finally, the border between defective and restrictive properties is precisely where we would like to ultimately draw the line between UD and PARSEME annotations, i.e. only defective properties would be rendered in the UD layers.

3 Annotation Principles

This section compares the annotation principles of UD and PARSEME, focusing on MWEs.

3.1 Objectives and Principles

The common objective of UD and PARSEME is *universality*, defined as development of cross-linguistically consistent and applicable language descriptions.[7] Both initiatives aim at representing in a unified way those phenomena which are truly similar, while leaving room for language-specific categories, relations and guidelines. The utility of these descriptions is twofold – meaningful linguistic analysis and useful language processing – in both monolingual and cross-lingual settings.

UD descriptions concern several aspects of language: segmentation, lemmas, morphology and syntax. According to the annotation properties defined by Mathet et al. (2015), these descriptions include *unitizing* (identify sentence and word boundaries) and have a *full covering* (concern all words in a corpus). PARSEME descriptions are mostly semantic (even if largely approximated by morphosyntax, see below). They also require unitizing, but are *sporadic* (only focus on components of MWEs), can be

[7]This is in contrast with the quest for absolute language universals (Greenberg, 1966; Chomsky, 1975; Tallerman, 2009).

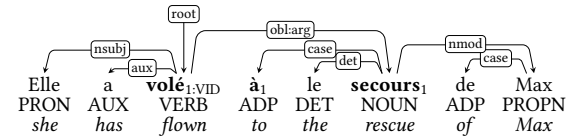

Figure 2: Sentence (8) with main annotations from UD (tree, POS tags) and PARSEME (bolding, subscripts).

nested (*[[**let**]$_2$ **the cat** [**out**]$_2$ **of the bag**]$_1$* 'reveal a secret') and exhibit *free overlap* (***take***$_{1,2}$ *a* ***walk***$_1$ *and a* ***shower***$_2$).

3.2 Notations and Formats

With respect to data formats, UD and PARSEME are largely compatible. Consider the example in sentence (8). Its main UD and PARSEME annotations are visualized in Fig. 2: parts of speech and dependencies are the UD-specific data, while MWEs (highlighted in boldface and subscripts) are tagged by PARSEME. The same example, in more detail, is presented in Fig. 3 in the tabular `.cupt` format.[8] Each word is described in a separate line, with 11 tab-separated fields, whose headings are listed in the first line of each file. The first 10 columns are those of the `.conllu` format used by UD. The 11th column (PARSEME:MWE) is used by PARSEME. Components of MWEs annotated in column 11 are shown in bold.

(8) Elle a **volé** **au** **secours** de Max (fr)
She have.3SG fly.PTCP to.the rescue of Max
'She hurried up to help Max'

3.3 Words and Tokens

Word is a fundamental notion both for UD, since its basic annotation unit is a word, and for PARSEME, since MWEs must contain at least two words. However, defining a word is one of the hardest challenges in UD, due to its fuzzy borders with morphemes on the one hand and with MWEs on the

[8]The `.cupt` format instantiates the CONLL-U Plus meta-format meant for complementing UD with additional layers: `https://universaldependencies.org/ext-format.html`

```
# 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   _ _  *
```

Figure 3: Annotation of sentence (8) as the first sentence in a corpus, in the `.cupt` format.

other. In UD, words are defined in morphosyntactic terms as units bearing morphological properties (e.g. a single POS) and entering into syntactic relations. Words do not always coincide with orthographic units called *tokens*.[9] Therefore, UD defines a 3-fold relationship between words and tokens:

- A token coincides with a word.
- Several tokens build up one *multitoken word* (MTW), as in *20␣000*.
- One *multiword token* (MWT) contains several words, as in (fr) *aux* (à+les) 'in.the'.

The words (not orthographic tokens) form the basic units of analysis and receive integer indices. MWTs are represented as spans over multiple words (e.g. 4–5 in Fig. 3), including cases where words (*à* and *le*) are not retrievable from tokens (*au*). PARSEME conforms to the same definitions of words, MWTs, and MTWs, with implications for MWEs like in Fig. 3. Only the adposition *à* 'to' belongs to the MWE;[10] the determiner *le* 'the' is excluded. This is possible in PARSEME due to splitting MWTs into words by UD.

Still, PARSEME covers a considerably higher number of MWTs than UD, especially verb-particle constructions written sometimes as 1 and sometimes as 2 tokens as in (9), and orthographically unitary (*closed* or *synthetic*) compounds as in (10).

(9) **auf-passen**, **pass auf**! (de)
on-fit.INF, fit.IMP on!
lit. 'to fit on, fit on!'
'to be careful, be careful!'

(10) Haupt**rolle spielen** (de)
head.role play
'to play the leading role'

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

Figure 4: PARSEME annotation of unsplit MWTs.

This discrepancy leads to two issues, illustrated in Fig. 4. First, the definition of a word is inconsistent: item 3 is one word for UD but two words for PARSEME. Second, in item 11 only *rolle* 'role' belongs to an MWE, since *Haupt* 'head' can be freely replaced (*Neben**rolle spielen*** 'play the secondary role'). This cannot be rendered if UD keeps compounds unsplit.

3.4 Morphology and Syntax

In UD, the morphological description of a word employs 17 universal POS tags and over 200 values for morphological features (columns 4 and 6 in Fig. 3), though explicitly admitting that some of them may not be necessary in some languages. Syntactic annotation in UD follows the dependency approach and adopts the *lexicalist* principle. Namely, words are divided into *content words* – typically verbs, nouns, adjectives or adverbs, with referential meaning – and *function words* – determiners, adpositions, auxiliaries, etc. Content words are linked by syntactic relations, while function words attach to the content words they modify. For instance, in Fig. 2, the verb is the head of the auxiliary (items 2–3) and the nouns are the heads of the prepositions (items 4–6 and 7–8) rather than vice versa. A set of 37

[9] Neither UD nor PARSEME define tokens. We see them as units stemming from segmenting raw text for annotation.

[10] As evidenced by variants like (fr) ***voler à*** *son* ***secours*** (lit. 'to.fly to his/her rescue') 'to hurry up to help him/her'

syntactic relations considered universal (column 8 in Fig. 3) is defined. More specific relations in a language are accepted as subtypes of the universal ones (e.g. **obl:arg** in line 6 in Fig. 3) and 26 such subtypes are currently found in the UD treebanks. Treebanks are not required to use language-specific extensions, even if they cover phenomena for which such extensions are defined. This leads to significant inconsistencies in the use of subrelations, even among treebanks of the same language.

PARSEME, while modeling idiosyncrasy, tries to remain as independent of a particular linguistic framework as possible. It considers, for instance, that in a prepositional phrase *a preposition directly governs a noun, or the opposite, depending on a particular linguistic theory*. However, PARSEME approximates semantic compositionality by lexical and morphosyntactic flexibility tests that are driven by syntactic structure. Thus, the main PARSEME decision diagram asks questions about the syntactic head of the candidate expression, its dependents, its morphosyntactic category, etc. This implies a strong dependence on the underlying syntactic framework, and UD provides such framework, validated across many languages.

Another advantage for PARSEME is that the lexicalism in UD helps keep the MWE definition relatively simple. Namely, MWE components more easily form a weakly connected dependency graph (Sec. 3.5) if content words head function words than vice versa (Savary and Waszczuk, 2020). One minor disadvantage from lexicalism concerns MWEs with copulas. For instance in (en) *to* ***be somebody*** 'to be important' the pronoun heads the copula *be*, which prevents PARSEME from saying that a verbal MWE is always headed by a verb.

The universality of UD thus enables universality for PARSEME, which has been increasingly relying on UD. For all 14 languages in version 1.2 of PARSEME, MWE annotations build upon UD-compatible corpora (manually annotated or automatically predicted); and among all 26 PARSEME corpora, 20 are UD-compatible.

3.5 The Notion of MWE

The way UD and PARSEME understand the notion of an MWE is the major source of apparent discrepancies between the two frameworks. UD did not attempt to formally define MWEs, using it as an umbrella term for expressions for which other syntactic relations seem useless or inconvenient. UD defines 3 dependency labels in such cases.

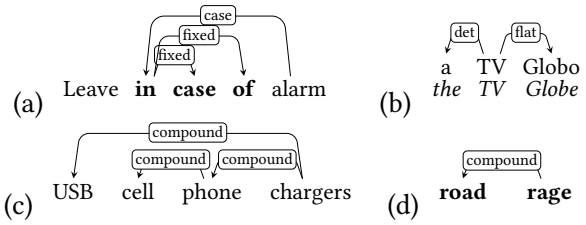

Figure 5: A complex preposition, a proper name (in Portuguese) and a nominal compound.synana

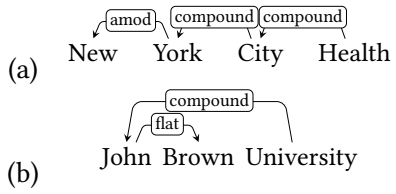

Figure 6: Complex names with mixed dependencies.

[fixed] is used for highly *grammaticalised* expressions, as in Fig. 5a, that typically behave as function words or short adverbials, i.e. belong to *closed* grammatical categories. The name of the label inspired by Sag et al. (2002) signals morphosyntactic fixedness. By convention, all parts of such an expression are attached to the leftmost component, that is, the whole is considered *headless* (even if a head might be identifiable).

[flat] is meant for *headless* semi-fixed expressions, like names or complex numerals, as in Fig. 5b. These belong to *open* categories and are subject to high *productivity*.

[compound] marks any word-level compounding, including nouns, adjectives, and verbs. Compounds are seen as *headed* expressions, i.e. modification relations are rendered, as in Fig. 5c. A compound may or may not be semantically compositional, as in Fig. 5c and 5d, respectively.

This typology concerns dependency relations, not expressions. In particular, various labels can be mixed within one expression, as shown in Fig. 6. Some UD subtypes (e.g. **compound:lvc**, **expl:pv**) are related to MWEs in PARSEME (Sec. 3.6).

For PARSEME, an MWE is a combination of words with at least two *lexicalized components* (always realized by the same lexemes) displaying lex-

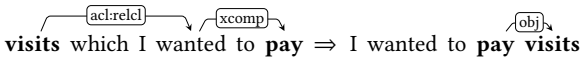

Figure 7: A VMWE candidate and its canonical form.

ical, morphological, syntactic or semantic *idiosyncrasies* (Sec. 2). Even if PARSEME's ambition is to model MWEs in general, its major efforts were put into *verbal MWEs* (VMWEs). A VMWE is defined as an MWE whose *canonical form* (least syntactically marked form keeping the idiomatic reading) is such that its syntactic head is a verb, its other lexicalized components form phrases directly dependent on this verb (the whole forms a weakly connected graph), and it passes the idiosyncrasy tests defined in the PARSEME guidelines. MWE candidate sequences must be transformed into canonical forms. For instance, the candidate on the left of Fig. 7 does not fulfill the conditions, but transforming it into the canonical form on the right restores graph connectivity and verb-headedness.

3.6 MWE Categories

PARSEME defines 3 quasi-universal categories (the first 3 below, present in many languages but not all), and 2 universal ones (the last 2 below, present in all languages under study).[11] Statistics about these annotations in the data are given in Appendix A.

Inherently Reflexive Verbs (IRV) combine a verb V and a reflexive clitic R such that (i) V never occurs without R, as in (sv) ***gifta sig*** (lit. 'get-married oneself') 'get married', or (ii) R distinctly changes the meaning or valency of V, as in (es) ***recogerse*** 'go home'/*recoger* 'gather'. They are contrasted with regular reflexives: true reflexive, reciprocal, middle passive and impersonal, e.g. (ro) *casele se vând bine* (lit. 'houses sell themselves well') 'houses sell well'. In UD, the above uses are divided into two classes, depending on if the reflexive clitic can or cannot be mapped on a semantic argument of the verb. In the former case, the "regular" dependency label corresponding to the role of the clitic is used, e.g. **obj** in (pl) *myć się* 'wash oneself'. The latter is covered by the **expl**etive label. Subrelations can further distinguish these uses, in particular, **expl:pv** covers case (i) above, signaling idiosyncrasy.

Verb-Particle Constructions (VPCs) have two subclasses in PARSEME. In fully non-compositional VPCs (VPC.full), adding the particle considerably changes the meaning of the verb, as in (sv) *Det* ***gick upp*** *för mig* (lit. 'It went up to me') 'It occurred to me'. In semi-compositional VPCs (VPC.semi), the particle adds a partly predictable but non-spatial meaning to the verb, as in (sv) ***äta upp*** (lit. 'eat up') 'finish eating'. Verb-particle combinations where the particle only adds spatial meaning are not annotated, as in (sv) *gick upp på vinden* 'went up to the attic'. In UD the subrelation **compound:prt** can be used to connect a particle to its head verb, regardless of idiomaticity, i.e. all 3 examples above fall into this category.

Multi-Verb Constructions (MVCs) are idiomatic combinations of two verbs, e.g. (fr) ***laisser tomber*** (lit. 'to let fall') 'to abandon', in particular serial verbs in Asian languages, e.g. (hi) ***kar le*** (lit. 'do take') 'do for one's own benefit'. This relates to the UD **compound:svc** subrelation, which however covers serial verbs both in idiomatic and compositional uses, e.g. (ja) *naguri korosi* (lit. 'punch kill') 'kill by punching'.

Light-Verb Constructions (LVCs) are combinations of semantically light verbs and predicative nouns expressing the semantics of the action or state. Two subcategories are defined. In LVC.full the verb's subject is the noun's semantic argument as in (sl) ***imeti predavanje*** 'give a lecture'. In LVC.cause the verb's subject is the cause or source of the noun, as in (en) ***grant right***. In UD, the same expressions are most often annotated with the "regular" **obj** dependency, even if the scope of the **compound:lvc** subrelation is similar to LVC.full.

Verbal Idioms (VID) is the most diverse category in PARSEME, gathering cases not covered by other categories. The verb's dependents are unrestricted, including subjects, as in (en) ***a little bird told*** *me*, direct objects, as in (6), etc. The verb can have several dependents, as in (en) ***cut a long story short***, or combine features from other VMWE categories, as in (sv) ***sätta sig upp*** *mot någon* (lit. 'sit oneself up against someone') 'defy someone'. A VID candidate must display lexical or morphosyntactic idiosyncrasy, as in (1)–(6). As VIDs are so diverse, there is no direct correspondence in UD. They are

[11] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=030_Categories_of_VMWEs

typically annotated as syntactically regular, possibly with subrelations for particles and reflexives when those are parts of the VID. The UD **`fixed`** relation cannot be used to signal inflexibility in VIDs since it is limited to functional MWEs.

```
# global.columns = ID FORM LEMMA …PARSEME:MWE
1 die          der          …  *
2 Haupt**rolle** Hauptrolle …  1@**6-10**:LVC.full
3 **spielen**  spielen      …  1
```

Figure 8: PARSEME tag with a sub-token span

4 Towards UD-PARSEME Unification

The discrepancies discussed above harm universality, therefore we are taking steps towards unifying UD and PARSEME. The expected advantage lies in a better parallelism in annotating syntactic vs. semantic and regular vs. idiosyncratic properties. Our intuition is that semantic non-compositionality is an intriguing phenomenon and annotators wish to signal it even when annotating morphosyntax. If an MWE has (partly) regular syntax but idiomatic semantics, and if only morphosyntactic labels are available, annotators might prefer to signal idiosyncrasy rather than regularity.[12] Another temptation is to introduce new subtypes such as **`obj:lvc`**, which could block other useful syntactic distinctions that could be encoded with subtypes (since recursive subtypes are not allowed). Adding the MWE layer to the annotation schema solves these problems.

Another motivation is that both automatic processing of MWEs and parsing benefit from solving the two tasks jointly (Constant et al., 2017; Taslimipoor et al., 2020), therefore aligning morphosyntactic and MWE annotations serves NLP. Here, we lay out a multistage unification roadmap for major issues, summarized in Appendix B.

Note that no re-annotation effort is required on the UD side in the first two stages. This is important for at least three reasons. Firstly, while for PARSEME idiosyncrasy is central, for UD it is only one of the many phenomena to be modeled. It is therefore natural for the PARSEME community to be the main responsible party for changes related to idiosyncrasy. Secondly, the UD community of treebank creators and users is very large. Any change in annotation principles, in order to be widely adopted, should minimize manual re-annotation and should be divided into small, easily achievable steps. Thirdly, as mentioned in Section 3.1, while PARSEME annotation is sporadic, UD trees fully cover the annotated text, which implies a heavier (re-)annotation workload.

[12] E.g. in the Romanian Reference Treebank, VIDs with regular syntax like ***avea loc*** 'take place' are marked as **`fixed`**.

4.1 Words and Tokens

PARSEME's notion of word is sometimes more granular than UD's (Sec. 3.3), and the segmentation of tokens into words would need to be reconciled. This is crucial for cases where MWEs cover parts of tokens, as in (10). Another such case occurs in Korean UD treebanks (Chun et al., 2018; Oh et al., 2020), where agglutinative postpositions are considered to form a syntactic word with their stem (segmented only in the lemma), as in (11). The postposition *에 (-ey)* and following word *대해 (tayhay)* together mean 'about', but because the postposition is not split, we would need to refer to a subword unit.

(11) 언어**에** **대해** 읽다 (ko)
language**:POSTP** **about** read
'read **about** languages'

Short-term Proposal We propose to supplement existing UD parses with MWE annotations, *without altering tokenization.* MWEs that encompass entire MWTs, as in (9), are already covered by PARSEME. For cases like (10) and (11), the MWE column could specify sub-token spans, as in Fig. 8.

Long-term Proposal Parts of unsplit tokens participating separately in MWEs suggest a deficiency in UD's implementation of MWTs. We propose that, ultimately, UD syntactically recognize synthetic compounds as productive, regardless of the MWE status. This would require UD treebanks in some languages to systematically split current compound words into MWTs, ensuring each component word has an appropriate lemma and morphological features, and adding a dependency relation (such as **`compound`** or **`compound:prt`**) between them. This could also help disambiguate the interpretation of some compounds, as in (sv) *bildrulle: bil+drulle, bild+rulle* 'car maniac (bad driver), picture roll (roll

of film)'.[13]

4.2 Terminology and Guidelines

A common understanding of MWE-related terminology is a basic requirement for UD/PARSEME convergence. This could be achieved progressively.

Short-term proposals Different interpretations of the term "multiword expression" are understandable (Sec. 3.5), since it literally means an expression containing two or more words, with no further restriction. However, the term, as understood by the MWE community (Baldwin and Kim, 2010), has an extra meaning component of idiosyncrasy (it is itself an MWE!), and we propose to adhere to this definition.[14] This would mean, for UD, *not to use the term MWE for phenomena considered regular*, replacing the MWE heading (currently describing **`compound`**, **`fixed`** and **`flat`**) with a more neutral description like "other complex constructions". This should be easy to achieve, as MWEs do not have a technical definition in UD: the term is used casually in the guidelines, but is not part of the morphological or syntactic labels or their criteria. This proposal is conservative in the sense that it does not, in principle, require modifications of the annotations.[15] On the PARSEME side, the VPC label might be renamed to IVPC (for idiomatic VPC), so as to signal that verb-particle combinations can be both regular and idiomatic, and only the latter are MWEs (Sec. 3.6). Criticism of the current VMWE guidelines (Savary and Waszczuk, 2020; Fotopoulou et al., 2021) should also be addressed.

Mid-term Proposals A major mid-term requirement for PARSEME would be to extend its terminology and guidelines to *all syntactic types* of MWEs, rather than VMWEs only, e.g. based on the foundational work by Schneider et al. (2014) and Candito et al. (2021). Challenges include defining the borders between named entities and MWEs.

Long-term Proposals Most languages contain *productive grammatical subsystems* which yield expressions with particular syntax and semantics, such as names, numbers, measurements, and dates (Kahane et al., 2017; Schneider and Zeldes, 2021). Their heavy semantic load makes them central units of interest for NLP. They partially overlap with regular syntax and MWEs, e.g. (pl) *Małgorzata Kowalska* is a name with a regular noun-adjective structure, and (en) *"Always **Look On The Bright Side** Of Life"* is a title containing a VID. However, they also follow specific patterns, such as defective number agreement in (en) *two million*, and nesting (Fig. 6). They call for normalisation standards like TimeML and AMR (Pustejovsky et al., 2003; Banarescu et al., 2013). Annotating subsystems jointly with UD and PARSEME would require new instantiations of CONLL-U Plus, with extra columns, such as 'NE' in Fig. 9. Other initiatives are making progress towards adding entity and coreference layers to UD (Nedoluzhko et al., 2022).

4.3 Occurrence vs. Type Encoding

We suggest unification steps towards a better account of the type/occurrence nature of idiosyncrasies.

Mid-term Proposals As soon as PARSEME extends its guidelines to all syntactic MWE types, they should be applied to all PARSEME corpora. The general principle would be:

- UD layers only account for lexical/morphosyntactic idiosyncrasy of MWE *occurrences*, such as irregular syntax in (3). Grammatically regular MWE occurrences would receive "ordinary" annotation, regardless of semantics.
- The PARSEME layer would signal *any* kind of semantic *idiosyncrasy*, i.e. it would flag each expression which is lexically/morphosyntactically irregular, whether at the level occurrences or of types, e.g. for all examples in Sec. 2.

This would require a systematic use of the `.cupt` format to jointly represent all dimensions of idiosyncrasy. This would also question the utility of UD's **`fixed`** label, since fixedness is a property of types rather than occurrences. Maybe this label could be merged with **`flat`** and both renamed to **`headless`** to avoid confusion with previous interpretations.

[13] Tokenization issues occur not only in compounds. Agglutinative languages may adopt different word segmentation strategies, in spite of similar structure (Han et al., 2020). This must also be addressed (Tyers et al., 2021) but goes beyond this paper's scope.

[14] Even if "idiomatic MWE" would be more precise.

[15] One exception, in English, would be to abandon the semantic **`compound:prt`** vs. **`advmod`** distinction in VPCs.

```
# global.columns = ID FORM ...HEAD DEPREL ...MWE NE
1  Leave ... 0  root    ... *                 *
2  in    ... 3  case    ... 1:AdvMWE.fixed *
3  case  ... 1  obl     ... 1                 *
4  of    ... 5  case    ... *                 *
...
11 Leave ... 0  root    ... *                 *
12 in    ... 15 case    ... 1:AdvMWE.fixed *
13 case  ... 12 headless ... 1                 *
14 of    ... 12 headless ... *                 *

31 a     ... 32 det     ... *                 *
32 TV    ... 34 nsubj   ... *                 1:ORG
33 Globo ... 32 headless ... *                 1
```

Figure 9: Two possible annotations for a multiword preposition; and a headless organization name.

The PARSEME layer might deal with signaling total (rather than partial) fixedness if needed. The example in Fig. 5a would be annotated as in Fig. 9, depending if it is seen as analysable (lines 1–4) or **`headless`** (lines 11–14). The example in Fig. 5b would also be **`headless`** (lines 31–33), with a possible named entity type (column 12), if a subsystem layer is added to the schema (Sec. 4.2).

These would be major changes, and authors of some treebanks might not be sufficiently interested in idiomaticity to accept the addition of a column. In this case, the previous distinction between **`fixed`** and **`flat`** should be kept to distinguish grammaticalized and productive headlessness. Subrelations such as **`compound:prt`** and **`compound:svc`** should probably be kept but used more consistently, since they are orthogonal to idiosyncrasy. Subrelations **`:lvc`** and **`:pv`** are superfluous: we propose to abandon them and use the 11th column instead.

Long-term Proposals Most optimally, the occurrence-type dichotomy of idiosyncrasy could be modeled in a framework in which corpus and lexicon are interlinked. A corpus would document occurrences, i.e. MWE occurrences would only be annotated for individual properties (including occurrence-wise idiosyncrasy such as irregular syntax). The lexicon would describe types, i.e. all occurrences of the same MWE would be linked to a lexicon entry representing its type and storing its type-wise properties such as categories (LVC, VID, etc.) and a meaning. A similar schema was implemented by Bejcek and Stranák (2010). An MWE lexicon entry could also contain other type-specific properties such as canonical forms (lemmas), flexibility and agreement constraints, as in (4–6), and links to ontologies (Hajnicz and Bartosiak, 2019).

Finally, an MWE lexicon could be more compliant with a typological perspective. PARSEME's current MWE typology is driven by annotation needs, i.e. new categories are introduced if specific tests are needed to identify some MWE in texts. An orthogonal, more typologically-driven categorization could use cross-linguistic constructions and language-specific structural types (Koptjevskaja-Tamm, 2002).

4.4 Data Quality

Both UD and PARSEME provide contributors with automatic data quality checkers. These should be unified and extended. PARSEME might enhance its validator to check compliance with guidelines (e.g. a verb in an LVC must have a single lexicalised dependent), and should integrate it with the UD validator, which runs automatically when a new version of a treebank is pushed to the GitHub repository. UD might develop tools inspired by PARSEME's consistency checks, in which a "vertical" view of the corpus groups annotations of the same MWE. This might help overcome inconsistencies within a treebank or within treebanks for the same language, e.g. due to the optionality of subrelations.

5 Conclusions and Future Work

We have compared how UD and PARSEME capture linguistic idiosyncrasy. Since PARSEME largely agrees with UD's objectives, it increasingly follows UD on data formats, morphology, (regular) syntax and tokenisation.

We are optimistic about UD and PARSEME joining forces for compatible encoding of regular and idiosyncratic phenomena, as detailed in our roadmap proposal. In the long run, these efforts might benefit from more typological insights. Also, extending the annotation schema to large classes of constructions would enable an even more comprehensive account of idiosyncrasy. The implementation of these suggestions will depend, however, on a delicate balance between existing and upcoming data, automation tools, and—above all—on availability and willingness of contributors.

Acknowledgements

This work started out as a discussion at the Dagstuhl Seminar 21351: Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics.[16] We thank all the participants for inspiration. We are grateful to Schloss Dagstuhl, the Leibniz Center for Informatics, as well as the organizers of the event, for bringing us together.

We are also building upon the efforts of UniDive, the CA21167 COST Action: Universality, diversity and idiosyncrasy in language technology.[17] At the time of drafting this paper, UniDive was at the proposal stage and paved the way towards many of the ideas presented here.

Our work has also been partly funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

We thank Jena Hwang for providing the Korean example (11).

References

Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7):89–138.

Timothy Baldwin and Su Nam Kim. 2010. Multiword Expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second edition*, pages 267–292. CRC Press, Boca Raton.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Eduard Bejcek and Pavel Stranák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1-2):7–21.

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Bruno Guillaume, Yannick Parmentier, and Silvio Cordeiro. 2021. A French corpus annotated for multiword expressions and named entities. *Journal of Language Modelling*, 8(2):415–479.

Noam Chomsky. 1975. *Reflections on Language*. Pantheon, New York.

Jayeol Chun, Na-Rae Han, Jena D. Hwang, and Jinho D. Choi. 2018. Building Universal Dependency treebanks in Korean. In *Proc. of LREC*, pages 2194–2202, Miyazaki, Japan.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Aggeliki Fotopoulou, Eric Laporte, and Takuya Nakamura. 2021. Where Do Aspectual Variants of Light Verb Constructions Belong? In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 2–12, Online. Association for Computational Linguistics.

Raymond W. Gibbs and Nandini P. Nayak. 1989. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive Psychology*, 21:100–138.

Joseph H. Greenberg, editor. 1966. *Universals of language*, 2nd edition. MIT Press, Cambridge, MA.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages*, 23(90):57–72.

Elżbieta Hajnicz and Tomasz Bartosiak. 2019. Connections between the semantic layer of *Walenty* valency dictionary and PlWordNet. In *Proceedings of the 10th Global WordNet Conference (GWC 2019)*, pages 99–107, Wrocław. Oficyna Wydawnicza Politechniki Wrocławskiej.

[16] https://drops.dagstuhl.de/opus/volltexte/2021/15591/

[17] https://www.cost.eu/actions/CA21167/

Ji Yoon Han, Tae Hwan Oh, Lee Jin, and Hansaem Kim. 2020. Annotation issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 99–108, Barcelona, Spain (Online). Association for Computational Linguistics.

Sylvain Kahane, Marine Courtin, and Kim Gerdes. 2017. Multi-word annotation in syntactic treebanks - Propositions for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 181–189, Prague, Czech Republic.

Maria Koptjevskaja-Tamm. 2002. Adnominal possession in the European languages: form and function. *STUF - Language Typology and Universals*, 55(2):141–172.

Timm Lichte, Simon Petitjean, Agata Savary, and Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and parsing of multiword expressions: Current trends*, pages 1–33. Language Science Press, Berlin.

Stella Markantonatou, Carlos Ramisch, Victoria Rosén, Mike Rosner, Manfred Sailer, Agata Savary, and Veronika Vincze. 2021. PMWE conventions for examples containing multiword expressions. Technical report, Phraseology and Multiword Expressions – book series at Language Science Press.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proc. of LREC*, Marseille, France.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: a multilingual treebank collection. In *Proc. of LREC*, pages 1659–1666, Portorož, Slovenia.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proc. of LREC*, pages 4027–4036, Marseille, France.

Tae Hwan Oh, Ji Yoon Han, Hyonsu Choe, Seokwon Park, Han He, Jinho D. Choi, Na-Rae Han, Jena D. Hwang, and Hansaem Kim. 2020. Analysis of the Penn Korean Universal Dependency Treebank (PKT-UD): Manual revision to build robust parsing model in Korean. In *Proc. of IWPT*, pages 122–131, Online.

Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3837–3852, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *IWCS-5, Fifth International Workshop on Computational Semantics*, Tilburg, Netherlands.

Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In *Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, Barcelona, Spain (Online).

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on*

Intelligent Text Processing and Computational Linguistics (CICLing-2002), pages 1–15, Mexico City, Mexico.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91, Florence, Italy. Association for Computational Linguistics.

Agata Savary and Jakub Waszczuk. 2020. Polish corpus of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 32–43, online. Association for Computational Linguistics.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proc. of LREC*, pages 455–461, Reykjavík, Iceland.

Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proc. of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 160–172, Sofia, Bulgaria.

Mark Steedman. 2000. *The Syntactic Process*. MIT Press.

Maggie Tallerman. 2009. If language is a jungle, why are we all cultivating the same plot? *Behavioral and Brain Sciences*, 32:469–470.

Shiva Taslimipoor, Sara Bahaadini, and Ekaterina Kochmar. 2020. MTLB-STRUCT @PARSEME 2020: Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 142–148, online. Association for Computational Linguistics.

Francis Tyers, Ekaterina Vylomova, Daniel Zeman, and Tim Zingler. 2021. *Working Group 1 (What counts as a word?)*, chapter 4.1. Volume 11 of (Baldwin et al., 2021).

A Statistics of the Use of MWE-related Labels in the UD and PARSEME Corpora

Table 1 shows the statistics and comments about the use of MWE-related labels in the UD treebanks in version 2.9 (with 131 treebanks in total).

Table 2 documents the number of PARSEME languages in which the MWE labels are used.

B Roadmap for UD-PARSEME Unification

Table 3 summarises the proposals from Section 4.

Label	Treebanks	Comments
`fixed`	109	Limited to functional MWEs
`flat`	119	Productive headless constructions
`compound`	99	Productive headed compounds. 10 additional treebanks have the compound relation, but always with a subtype.
`expl:pv`	20	Inconsistent use in Spanish-AnCora vs. Spanish-GSD, French-GSD vs. other French treebanks
`compound:prt`	32	In English **`compound:prt`** is used when the particle is not spacial, and **`advmod`** otherwise. The same distinction is suggested in the universal guidelines. Inconsistently used in Persian-Seraji vs. Persian-PerDT
`compound:svc`	8	
`compound:lvc`	11	Most often commuted for **`obj`**. Inconsistently used in Turkish-BOUN and Turkish-IMST vs. all other Turkish treebanks

Table 1: Use of MWE-related labels in the UD treebanks in version 2.9 (with 131 treebanks in total)

Label	Corpora	Comments
IRV	8	
VPC.full	6	Greek and Hebrew use only VPC.full
VPC.semi	5	Chinese uses only VPC.semi
MVC	7	
LVC.full	14	Hindi allows adjectives in place of nouns
LVC.cause	13	Not in Turkish
VID	14	

Table 2: Use of MWE-related labels in the PARSEME corpora in version 1.2 (with 14 languages in total)

	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

Table 3: Roadmap for the UD-PARSEME unification. Actions with white background require no manual (re-)annotation. Actions highlighted in blue will require major annotation effort: those in dark blue apply to all languages, whereas those in light blue (concerning subword-level annotations) apply to a subset of languages.