Swedish Multiword Expression Corpora in PARSEME

Sara Stymne, Astrid Berntsson Ingelstam, and Eva Pettersson
Department of Linguistics and Philology
Uppsala University, Sweden
[first_name.last_name]@lingfil.uu.se

Abstract

We present the annotation of Swedish multi-word expressions under the PARSEME annotation scheme, including a new release and a historical overview of previous releases. We provide an overview of the evolution of the Swedish datasets and of inter-annotator agreement. We discuss general guidelines and the development of Swedish-specific guidelines for particle verbs and multiword tokens, as well as additional challenges for the Swedish annotation. We also conduct an initial comparison of Swedish and other Germanic languages, identifying aspects where the PARSEME guidelines require revision to ensure better consistency across languages.

1 Introduction

Multi-word expressions (MWEs) are non-compositional expressions like *hålla ett öga på* ('keep an eye on') or *ge upp* ('give up'), where the meaning of the full expression cannot be directly inferred from the meaning of its parts. MWEs are challenging for many NLP applications, due to their irregularities, and have been called "a pain in the neck for NLP" (Sag et al., 2002; Shwartz and Dagan, 2019). The processing of MWEs is challenging for classical NLP, both for MWE-centered tasks, such as MWE discovery and identification, and for handling MWEs in other tasks, such as parsing and machine translation Constant et al. (2017). MWEs remain challenging for LLMs; for instance, Miletić and Schulte im Walde (2024) find that transformer models struggle with MWEs, largely relying on surface patterns and memorized information.

PARSEME[1] is an initiative that provides universal guidelines for consistent annotation of MWEs across languages. Their first iteration, version 1, only covered verbal MWEs (VMWEs), but since version 2.0, the coverage has been extended to all types of MWEs. PARSEME is based on general MWE categories, present across languages, with an annotation framework based on decision diagrams for annotation. In some cases, language-specific rules are added as a complement.

In this paper, we focus on Swedish and describe the evolution of Swedish MWE resources in PARSEME. We give an overview of the Swedish resources in the existing PARSEME releases and describe the work on a coming release, which we refer to as release 2.1. We discuss the annotation process, inter-annotator agreement, and MWE distribution, as well as Swedish-specific guidelines and challenges, related to phenomena such as multiword tokens and particle verbs. Finally, we compare the MWE distribution for the Germanic languages available in PARSEME.

2 Related work

In this section, we describe the PARSEME framework for MWE annotation, followed by a discussion of other work on MWE resources for Swedish and other languages.

2.1 PARSEME

The PARSEME scientific network has produced guidelines and corpora for the annotation of MWEs since 2017 (Savary et al., 2017; Ramisch et al., 2018, 2020; Savary et al., 2023a, 2026). In version 1, the resources covered only verbal MWEs, but since version 2, all types of MWEs are covered.[2] The VMWE guidelines are thus tested in several iterations, whereas the extended guidelines for other MWE types are currently in their first iteration. Most PARSEME releases have been tied to a shared task on MWE identification (e.g. Scholivet et al., 2026), but version 1.3 was released inde-

---

[1] https://gitlab.com/parseme/corpora/

[2] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/

VMWE	VID	verbal idiom	*att* **träda i kraft** (lit. 'to step in force') 'to come into effect'
	LVC.full	light-verb construction; bleached verb	*att* **hålla tal** (lit. 'to hold speech') 'to make a speech'
	LVC.cause	light-verb construction; causal verb	*att* **väcka hopp** (lit. 'to wake hope') 'to inspire hope'
	IRV	inherently reflexive verb	*att* **gifta sig** (lit. 'to marry oneself') 'to get married'
	IVPC.full	idiomatic verb-particle construction	*att* **höra till** (lit. 'to hear to') 'to belong'
	IVPC.semi	semi-idiomatic verb-particle constr.	*att* **fråga ut** (lit. 'to ask out') 'to interrogate / to invite out'
	MVC	multi-verb construction	*No instances found in the Swedish data*
	IAV	inherently adpositional verb	*att* **bero på** 'to depend on' (*Optional and experimental MWE category not annotated in Swedish*)
NMWE	NID	nominal idiom	**flod\häst** (lit. 'river horse') 'hippopotamus'
	PronID	pronominal idiom	**en del** (lit. 'a part') 'some'
	NV	deverbal nominal MWE	**ut\gåva** (lit. 'out-gift') 'edition' (cf. *att* **ge ut** (lit. 'give out') 'publish')
AMWE	AdjID	adjectival idiom	**så kallad** 'so-called'
	AdvID	adverbial idiom	**över huvud taget** (lit. 'over head taken') 'at all, even'
	AV	deverbal adjectival/adverbial MWE	**igen\känd** (lit. 'again known') 'recognized' (cf. *att* **känna igen** (lit. 'know again') 'recognize')
FuncMWE	DetID	determiner idiom	**ett par** (lit. 'a pair') 'a couple of'
	AdpID	adpositional idiom	**i och med** (lit. 'in and with') 'due to, because of'
	ConjID	conjunction idiom	**såväl som** 'as well as'
	IntjID	interjection idiom	**oj då** (lit. 'oh then') 'oh dear / ouch' (*No instances found in the Swedish data*)

Table 1: PARSEME typology of MWEs with Swedish examples.

pendently. Overall, the PARSEME corpora cover 33 languages, but the language coverage varies between releases. Release 1.3 includes 26 languages, and release 2.0 includes 17 languages.

The main MWE categories are verbal, nominal, adjectival/adverbial, and functional MWEs, each of which has several subtypes. Table 1 provides an overview of all MWE categories, accompanied by Swedish examples. The verbal category has six main categories: verbal idioms, VID, light verb constructions, LVC, inherently reflexive verbs, IRV, idiomatic verb-particle constructions, IVPC[3], multi-verb constructions, MVC, and inherently adpositional verbs, IAV. IAV is an optional and experimental category that has not been included in any of the Swedish releases. Nominal MWEs are split into nominal, NID, and pronominal, PronID idioms; adjectival/adverbial MWEs into adjectival, AdjID, and adverbial, AdvID, idioms; and functional MWEs into determiner, DetID, adpositional, AdpID, conjunctional, ConjID, and interjection, IntjID, idioms. Nominal and adjectival/adverbial MWEs further have a deverbal category, for expressions that can be rephrased into a verbal expression that passes the verbal MWE tests. These are further subcategorized into the verbal subtypes, but in this paper, we group them into the main classes: deverbal nominal and deverbal adjectival/adverbial. Some meaning-preserving variants that were annotated as VMWEs prior to release 2.0, are now instead considered deverbal.

The PARSEME guidelines are organized as a decision tree, where the first question concerns the distribution of the candidate expression, specifically whether it is verbal, nominal, adjectival/adverbial, or functional. This is followed by a specific decision tree for each MWE subtype, which includes several tests to determine whether a candidate meets the MWE criteria. The goal is to annotate expressions that are non-compositional. Since it is hard to directly judge the semantic idiomacy or non-compositionality of expressions, the tests are based on the fact that non-compositionality correlates with syntactic and morphological inflexibility (Sag et al., 2002). The rules thus test aspects such as whether a regular syntactic or morphological change leads to unexpected meaning change, whether a cranberry word is present, or whether the internal syntactic structure of the expression is irregular. MWE candidates must consist of multiple tokens, but they can also be multiword tokens (MWTs), i.e., cases where one token contains several words, such as compounds (e.g. *sky\skrapa* 'skyscraper'),[4] particle verbs (e.g. *in\gripa* (lit. 'in seize') 'intervene'), and contractions (e.g. 'don't'). MWTs are common in Germanic languages, which are compounding languages, except for English.

MWEs are annotated using the FLAT annotation tool.[5] Annotators first identify MWE candidate expressions, and then label them as MWEs if they

---

[3]IVPC was called VPC in releases prior to 2.0. We will consistently use the term IVPC even for earlier releases.

[4]Vertical bars are used to indicate token boundaries in MWTs; they are not part of the Swedish orthography.

[5]https://github.com/proycon/flat

pass the tests in the PARSEME decision trees. In most cases, a file is annotated by a single annotator due to annotator availability; however, inter-annotator checks are regularly performed (Ramisch et al., 2018, 2020). To further improve the quality, the PARSEME protocol of annotation includes a consistency-check step, where all annotations are reconsidered across files (Savary et al., 2017). Inconsistencies can be identified by the use of a custom tool that highlights examples of potential MWEs across all texts, based on lemmas, by searching for all MWEs that have been annotated at least once. A single annotator can then change inconsistently annotated examples, or add/remove examples that have been missed or erroneously annotated. This process resolves many inconsistencies between annotators.

PARSEME annotations use a format called CUPT (Ramisch et al., 2018), which is an extension of the ConllU format from the Universal Dependencies (UD) initiative (Nivre et al., 2020; de Marneffe et al., 2021), adding a column with MWE annotations. UD is another cross-linguistic initiative for morphosyntactic annotation, focusing on lemmas, morphology, part-of-speech tags, and dependency annotations. PARSEME corpora, from release 1.1, include UD annotations, either by annotating MWEs on top of existing UD treebanks, as for Swedish, or by parsing other corpora into the UD format. Discussions on how to further unify PARSEME and UD are ongoing (Savary et al., 2023b).

2.2 PARSEME MWE Annotations in Other Languages

The annotation of VMWEs for the Turkish PARSEME corpus posed problems due to the agglutinative nature of the language (Ozturk et al., 2022). Specifically, the automatic lemmatization of the Turkish corpus was often incorrect, with suffixes being incorrectly or insufficiently stripped, causing inconsistencies in the finalized corpus, as a single MWE type is represented by a sequence of included lemmas.

Hadj Mohamed et al. (2025) note that since Arabic is an agglutinative language, with many MWTs, only parts of an MWT may belong to an MWE. They also note that Arabic has a high rate of discontinuous MWEs. Walsh et al. (2020) note several issues, including the difficulty of distinguishing between particle verbs and inherently adpositional verbs, where particles often are homonymous with prepositions. They note that language-specific tests for this issue are needed for Irish.

2.3 MWE Annotations Beyond PARSEME

There are many datasets available for MWEs for a single language, sometimes focusing on specific types of MWEs only, such as Estonian particle verbs (Kaalep and Muischnek, 2006) or French adverbial MWEs (Laporte et al., 2008). Datasets covering many languages are not as common. ID10M contains annotations of idioms for 10 languages, created automatically based on Wiktionary entries marked as idiomatic or literal, with matching occurrences from Wikipedia text, and a manual curation for four languages (Tedeschi et al., 2022). Other multilingual resources of MWEs are often created for specific tasks in a few languages, such as multimodal MWE comprehension in English and Turkish (Pickard et al., 2025) or idiomaticity and semantic text similarity for English, Portuguese, and Galician (Tayyar Madabushi et al., 2022).

2.3.1 Work on Swedish MWEs

The Swedish lexicon resource SALDO (Svenskt associationslexikon 2, Borin et al., 2013) contains some lexicalized multiword and compound word entries. When describing SALDO, Borin et al. (2013) discuss the frequent compounding in Swedish, and state that a single token often corresponds to an MWE in other languages. Such lexicalized compound words often have a compositional sense that differs from the literal reading; a trait that is emphasized in the PARSEME guidelines. For instance, the noun compound *husbil* (*hus+bil*) means 'camper' or 'trailer', but the literal meaning reading is 'house car'.

Kurfalı et al. (2020) form a dataset of 96 Swedish verbal, nominal, and prepositional MWEs from SALDO paired with human-judged scores for degree of compositionality. They compare the human judgment with scores from computational vector models, and find that these agree poorly. Furthermore, Tiedemann et al. (2022, 2024) attempt to relate MWEs extracted from SALDO to the language proficiency levels of the Common European Framework of Reference (CEFR). They compare difficulty rankings of MWEs by L2 learners and teachers of Swedish to frequencies in course books and learner-produced texts at different proficiency levels. Tiedemann et al. (2024) also suggest that more transparent MWEs with a lower compositionality are easier for L2 speakers to learn, whereas

less transparent MWEs are more difficult.

3 Corpus Description

In this section, we give an overview of the Swedish PARSEME releases, including our latest annotations, and discuss inter-annotator agreement.

3.1 Swedish PARSEME Releases

Swedish was present in the first PARSEME release, version 1.0 (Savary et al., 2017), with a very small dataset covering 200 sentences, sampled from the Swedish Newspaper *Göteborgsposten* (GP). Three annotators were involved in this release. Swedish was not part of release 1.1 (Ramisch et al., 2018).

From release 1.2 (Ramisch et al., 2020), Swedish was added again, now with considerably larger corpora. From this release, the data to annotate was taken from the Swedish Universal Dependencies treebanks Talbanken[6] and PUD[7], containing human-annotated morphosyntactic information and lemmas. This data was annotated for MWEs by a group of six annotators, covering the training section of Talbanken. Release 1.3 (Savary et al., 2023a) was an extension of 1.2, annotated by two annotators from release 1.2, covering all of Talbanken. Up to release 1.3, only VMWEs were annotated. The annotation group from releases 1.2 and 1.3 received initial training, followed by regular email discussions about borderline cases.

Release 2.0 covered all MWE types, and was annotated by a team of five annotators, partially overlapping with the team from version 1.2/1.3. Most of the previously annotated data from Talbanken were reannotated with the new guidelines. In addition, all data in the PUD treebank was annotated from scratch. The annotator team received initial training, followed by regular meetings to discuss annotation guidelines and tricky cases.

In this paper, we also describe a new annotation effort, which we will tentatively refer to as release 2.1, where the annotations from release 2.0 are revisited, and the remaining part of Talbanken is added. In this annotation phase, two annotators from release 2.0 participated. These annotators held discussions to resolve disagreements over the interpretation of the guidelines from version 2.0.

---

[6] https://github.com/UniversalDependencies/UD_Swedish-Talbanken/blob/master
[7] https://github.com/UniversalDependencies/UD_Swedish-PUD/tree/master

3.2 Corpus Statistics

Table 2 gives an overview of all Swedish PARSEME releases, and the distribution of VMWEs. Table 3 gives further details on the MWE distribution for the two releases that contain all MWE types. Percent distributions of the MWE categories are shown in Tables 6 and 7. Except for release 1.0, all releases are reasonably sized, covering at least 4,000 sentences, and range between 1,991 and 4,904 MWEs. Among the verbal categories, it is clear that IVPCs are the most common type, followed by LVCs and VIDs. The only rare category is LVC.cause; all other categories have close to or over 100 instances in our latest release.

Among the new MWE types added in release 2.0, AdvID is by far the most common type, followed by NID and NV. The AdvID category largely consists of prepositions followed by nouns for which a morphological inflection would lead to ungrammaticality or an unexpected change in meaning. In many cases, the noun can only take the indefinite singular form, as in (1). In other cases, the noun is always in the definite form instead, as in (2).

(1) **till exempel**
to example
'for example'

(2) **i onödan**
in non-necessity.DEF
'unnecessarily'

We note that in all releases, except 1.0, that do not cover them, multiword tokens are common. One clear distinction between the releases is the proportion of multiword tokens (%MWT). The guidelines for when to consider a token as an MWT changed between release 1.3 and 2.0, as described in Section 4.2, resulting in a decrease in the proportion of verbal MWTs from approximately 50% in releases 1.2 and 1.3 to under 30% in releases 2.0 and 2.1. This change especially affected IVPCs, with a much lower proportion in releases 2.0/2.1. This change is also reflected in the number of MWEs per sentence. The MWT rate is overall high in releases 2.0 and 2.1, though, to a high extent due to new categories with a very high proportion of MWEs, such as NID (97%), NV (100%), and AV (100%). Another notable change is the increase of ConjID from release 2.0 to 2.1, which is mainly caused by a changed decision for common correlative conjunctions like **varken ... eller** 'neither ... nor' and **både ... och** 'both ... and', which are considered as MWEs in release 2.1, but not in 2.0.

Release	Text source	Sent.	Tokens	VMWE			VID	IRV	LVC full	LVC cause	IVPC full	IVPC semi
				Total	%MWT	Per sent.						
SV 1.0	News (GP)	200	3,376	56	0	0.28	9	3	13		31	
SV 1.2	Talbanken	4,304	65,482	1,991	48.6	0.46	291	115	279	11	871	424
SV 1.3	Talbanken	6,026	96,820	3,155	51.2	0.52	441	237	417	10	1461	589
SV 2.0	Talbanken, PUD	5,553	90,392	1,779	27.9	0.32	396	186	389	19	494	294
SV 2.1	Talbanken, PUD	7,026	115,944	2,275	27.7	0.32	480	295	500	29	580	391

Table 2: Corpus overview and distribution of verbal MWEs for the Swedish PARSEME releases. Note that no distinction was made between subtypes of LVC and VPC in release 1.0.

Release	MWE			Verbal	Nominal			Adjectival/adverbial			Functional		
	Total	%MWT	Per sent.	Total	NID	PronID	NV	AdjID	AdvID	AV	AdpID	ConjID	DetID
SV 2.0	3,678	58.0	0.66	1,779	331	91	232	76	775	84	127	91	94
SV 2.1	4,904	53.7	0.70	2,275	492	111	318	98	929	116	187	222	156

Table 3: Overview and distribution of MWEs in editions that annotate all MWE types. For verbal distribution, see Table 2.

		F1-score	
	Sent.	VMWE	All MWE
SV 1.2	700	73.4	–
SV 2.0	500	46.3	47.0
SV 2.1	300	88.4	54.0

Table 4: Inter-annotator agreement as F1-score across two annotators for three versions of the Swedish PARSEME corpus.

3.3 Corpus Quality

For three of the PARSEME releases, we performed double annotation to assess inter-annotator agreement. For version 1.2, two annotators annotated the same 700 sentences for VMWEs. In version 2.0, the first version to contain MWEs beyond VMWEs, 500 sentences were annotated by a different set of two annotators from scratch. For version 2.1, we had no completely fresh data, so two annotators separately annotated on top of the VMWE annotations from version 1.3. Note, however, that the guidelines for MWTs, as well as some guideline interpretations, had changed between these two iterations, which affected a relatively large number of VMWE instances.

To calculate agreement, we followed Savary et al. (2026), and used the F1-score between the two annotations, since calculating chance agreement, which is needed for most agreement measures, is challenging in cases like these, where relatively few words in a text are annotated. Table 4 provides an overview of the results, reported both for the full annotations and for the subset of version 2 annotations that are VMWEs. We note that the agreement decreased from version 1.2 to 2.0, possibly due to the much more diverse sets of MWEs being annotated, and new MWT guidelines being developed during the course of the annotation (see Section 4.2). Between releases 2.0 and 2.1, the two annotators had a discussion about disagreements, borderline cases, and the interpretation of the guidelines in release 2.0, which led to a higher level of agreement in version 2.1. We especially note that, although the annotations for verbal MWEs were performed on version 1.3, the agreement scores are higher for VMWEs than for version 1.2.

While the agreement scores for non-verbal categories are low, the consistency of the final dataset is improved by the use of consistency checks across files (see Section 2.1), where one annotator goes through all annotations across files before each release. To give an indication of the proportion of changed decisions, Table 5 shows the F1-score of comparing the annotations before and after consistency-checking for the top categories of MWEs. For categories with subcategories, we report macro F1-scores. Overall, VMWEs show relatively few changes, with scores over 90 for all types, further supporting that the annotation of these categories is more consistent than for other categories. Nominals, conjunctions, and deverbals have considerably lower scores, between 62 and 84. For conjunctions, this can mostly be attributed to the decision to include correlative conjunctions as MWEs (see Section 3.2).

As a further point of comparison, across languages for version 2.0, the average inter-annotator agreement F1-score was 60.6, with scores ranging from 20 to 99 (Savary et al., 2026). For 7 of the 14 languages for which inter-annotator agreement was calculated, the F-score was below 65. This warrants a detailed cross-lingual investigation into agreement across PARSEME languages to deter-

VID	IRV	LVC	IVPC	AV	AdjID	AdpID	AdviD	ConjID	DetID	NID	NV
92	97	97	94	78	90	89	90	62	76	84	81

Table 5: F-scores per category when comparing corpus release 2.1 before and after consistency checks. For IVPC, LVC, and deverbal categories, results are macro-averaged across subclasses.

mine whether the widespread low agreement scores are primarily due to the new MWE categories in other languages than Swedish as well. In addition, a more detailed investigation of which categories pose problems is needed. Such an investigation could feed into a new, improved version of the PARSEME guidelines.

3.4 Releases

Versions 1.0 to 2.0 described in this paper are already released, as part of multilingual PARSEME releases. For the latest annotations, we plan to release them in two ways: as part of the next PARSEME release, 2.1 (which is currently not yet scheduled), and as part of the next UD release, by adding the PARSEME annotations to Talbanken and PUD. The release of UD version 2.18 is scheduled for May 2026. Both releases will be under permissive licenses.

4 Swedish-Specific Considerations

In this section, we discuss considerations for the Swedish PARSEME annotation, first by describing Swedish language-specific guidelines, and then by discussing issues encountered during annotation. While our discussion is based on a Swedish perspective, several issues are also relevant to other languages, and we include some comparisons to other Germanic languages.

4.1 Swedish-specific PARSEME Guidelines

The PARSEME guidelines include language-specific tests that may relate to language-specific MWE categories, more specific tests for some MWE types, or elementary language features. Swedish has language-specific tests for two issues of the latter category, particles and MWTs.

4.1.1 Particles versus Prepositions

Particle verbs or phrasal verbs are pervasive in the Germanic languages, and idiomatic uses of them are covered by the IVPC category in PARSEME. In many cases, particles are homonymous with prepositions in prepositional complements or verb prefixes, and there is thus a need to be able to distinguish these cases. Language-specific rules currently exist for Swedish, English, and German,[8] but the need is also noted for languages from other families, like Irish (Walsh et al., 2020). Example (3) shows an ambiguous sentence in Swedish, where a particle verb reading corresponds to the English translation 'visits', whereas a prepositional reading translates into 'greets'. The main way to distinguish particles and prepositions in Swedish is through stress patterns, where particle verbs have the primary stress on the particle, whereas prepositional verbs have the main stress on the verb, with an unstressed preposition (Svenonius, 2003). Thus, the stress pattern is the basis of the decision rule for Swedish in PARSEME.

(3) Hon **hälsar på** oss.
     She greets on us
     'She visits/greets us.'

This rule contrasts with the English and German rules, which are based on tests of movement and insertion. In both English and German, the particle can be placed at the end of the sentence, as in 'she takes her clients in', whereas in Swedish it cannot. Additionally, English has a test based on adjunct insertion, and German has a test for separable verb prefixes, which are not applicable to Swedish; in contrast to German, where separated and compounded forms of particle verbs are governed syntactically, Swedish particle verbs occur either in a separated or compounded form based on the expression, where some expressions can occur in both forms, but sometimes with shifted semantics between the forms (Norén, 1995). There could be a possibility of a syntactic test for Swedish, based on topicalizing the prepositional object, as in (4), which is not possible with a particle reading. However, this test can be challenging to apply in certain cases (Svenonius, 2003), so even if it were added as an additional test for particles in Swedish, we still believe the test for stress patterns is necessary.

---

[8] https://parsemefr.lis-lab.fr/
parseme-st-guidelines/2.0/?page=060_
Language-specific_tests/010_Particles,
_prepositions,_prefixes

(4) **På** oss **hälsar** hon.
On us greets she
'Us, she greets. *Us, she visits'

4.2 Multiword Tokens

To unambiguously identify multiword tokens (MWTs) is crucial for identifying MWE candidates. However, so far, Swedish is the only PARSEME language with language-specific rules for MWTs. We believe such rules should be added for other languages as well, including German and Dutch, which have a similar structure to Swedish regarding MWTs.

For Swedish, there have been two iterations of rules for MWTs, one for releases 1.2–1.3, and one from release 2.0 forward. The first set of rules was part of releases that covered only VMWEs, meaning the rules for MWTs also covered only verbal expressions. Thus, there was a first rule for testing whether the expression was verbal. The second rule, *splittable*, states that the expression is an MWT if it can be used in its split form, with the same or slightly shifted semantics. The third rule is very permissive, and states that an expression is an MWT if all of its components can be used as standalone words with the same part-of-speech as in the full token.

As discussed in Section 3.2, these rules have a considerable impact on the number of MWTs and thus also on the number of MWEs, especially for IVPCs, where the first set of MWT rules is highly permissive. They lead to many instances of MWTs that passed the MWE guidelines but did not align with the annotators' intuition. Additionally, the MWT rules needed to be extended to cover non-verbal expressions for release 2.0. Keeping these rules would have led to inconsistencies, such as: not considering the verb *för|lora* (lit. 'for LORA') 'lose' an MWT, but considering the related noun *för|lust* (lit. 'for lust') 'loss' an MWT and also an MWE, since 'lust' but not 'lora' happens to be a stand-alone word. These issues caused discussions among the annotators of the 2.0 corpus version, leading us to revise the rules before finalizing the annotations for the release.

The rules for release 2.0/2.1 first test whether the expression is a noun-noun compound (NNC). Noun-noun compounding is very frequent in Swedish, and it is ungrammatical to directly split an NNC into separate graphical tokens, while some NNCs can be split using prepositional complements. We still wanted to consider all NNCs as MWTs to achieve a better cross-lingual comparison with the many languages where NNCs are written as graphically distinct words, such as in English, and thus are considered MWE candidates in those languages. This means that an NNC like **bergs|kedja** (lit. 'mountain chain') 'mountain range' would be considered an MWT and thus it can also be an MWE as can its English counterpart.

For any remaining MWT candidates, we keep the *splittable* rule from before. We also note that many nominal, adjectival, or adverbial expressions that are deverbal could be split in their verbal form, but not in their original form. We thus add a third rule: if a non-verbal expression has a corresponding verbal form that can be split with the same or slightly changed semantics, it should be considered an MWT, but only the tests for the deverbal classes NV or AV should be considered. For examples of deverbals, see NV and AV in Table 1.

As a result of the updated MWT rules, candidates like *genom|snitt* (lit. 'through incision') 'average', *där|för* (lit. 'there fore') 'because', and *för|lust* (lit. 'for lust') 'loss' are no longer considered as MWTs, whereas all NNCs are considered as MWTs, so that those that fill the criteria for NID are annotated as such. In addition, we now cover the quite common class of deverbal MWTs.

The change in MWT guidelines affects the annotation of phrasal verbs (IVPC.full/semi), including deverbal IVPCs. Swedish phrasal verbs that have a split form, such as *att* **hälsa på** (lit. 'to greet on') 'to visit' and *att* **fråga ut** (lit. 'to ask out') 'to interrogate', can often be compounded into particle+verb. Deverbal forms of phrasal verbs are commonly compounded in such a way, for example **på|hälsad** (lit. 'on greeted') 'visited' (AV.IVPC.full) and **ut|frågning** (lit. 'out asking') 'interrogation' (NV.IVPC.semi). There are many cases of deverbal particle+verb compounds, such as **ut|bildad** (lit. 'out educated') 'educated' (from **ut|bilda** (lit. 'out educate') 'educate'), where the split verbal version is either not possible (*bilda ut*) or has a completely different sense from the compounded verb. According to the previous MWT guidelines, such cases would be annotated as deverbal IVPCs as long as the particle has primary stress. With the new guidelines, however, such words are not considered MWTs and are therefore not annotated as MWEs.

33

4.3 Issues for Swedish MWE Annotation

4.3.1 MWEs as Part of a Token

Due to frequent compounding in Swedish, the corpus contains instances in which only part of a graphical token belongs to an MWE, while other parts do not. One example is the LVC.full in (5), where only part of the compound word is a part of the LVC. In other cases, an MWE can be embedded within an MWT, such as the NV.IVPC.semi *till|syn* (lit. 'to see') 'supervision' in (6) which has the verbal correspondent *se till* (lit. 'see to') 'look after'. With the current PARSEME annotation framework, it is not possible to annotate subtokens, which would be preferable in this case, and thus, the full token is annotated as an MWE. This issue has previously been discussed with a proposed solution (Savary et al., 2023b), but it has not yet been implemented, as it requires changes to the entire annotation process, including the ability to mark subtokens in the annotation tool, process them during consistency checking, and update the CUPT format.

(5) att **ha**   vårdnads|**rätt**
     to have custody-right
     'to have right to custody'

(6) barn**tillsyn**sproblem
     barn **till|syn**s   problem
     child supervision problem
     'child supervision issue'

4.3.2 Reflexive Particle Verbs

PARSEME recognizes several subtypes of VMWEs, including idiomatic particle verbs (IVPCs) and inherently reflective verbs (IRVs). In Swedish, it is common for verbs to take both a particle and a reflexive, as in (7) and (8). However, there is no specific category for this combined category; they thus need to be annotated as verbal idioms (VIDs), which is the only possible category for verbs with more than one argument. We would advocate for the inclusion of an IVPC-IRV category for these cases in PARSEME. This issue would also need to be discussed in connection with inherently adpositional verbs (IAVs), currently experimental and not annotated for Swedish, since they can also be combined with reflexives and particles, as in *slå sig ihop med* (lit. 'hit oneself together with') 'gang up with'.

(7) att **bry sig**   **om**
     to care oneself about
     'to care about'

(8) att **ta**   **med sig**
     to take with oneself
     'to bring'

4.3.3 Splittability of Multiword Tokens

The updated guidelines for determining whether a compound word is an MWE candidate (see Section 4.2) build on splittability, but it is not always clear if a certain compound token is splittable or not. Many compounds, especially adverbial compounds, originate from a co-occurrence of two separate tokens that develop a new, compositional sense, and are accordingly written together as one token more frequently. For instance, *över|allt* (lit. 'over everything') 'everywhere' originates from the two words 'över allt', but it cannot be used in a split form anymore, and it is thus not an MWT according to release 2.0 guidelines. In other cases, as with *i|dag* (lit. 'in day') 'today' the compositional sense 'today' can correctly be written also as two separate tokens, 'i dag', which means that the form 'idag' is considered an MWT, according to the *splittable* rule.

5 Comparison of Germanic Language MWEs

To further contextualize the Swedish MWE annotations, we compare the MWE distribution across MWE types with that of other Germanic languages. Since not all languages are available in each release, we select the latest release for each Germanic language present in PARSEME: English 1.2, German 1.3, and Dutch 2.0. We compare this with the two largest Swedish releases, 1.3 and 2.1. Table 6 contains an overview of this comparison for verbal MWEs, and Table 7 contains an overview for Swedish 2.1 and Dutch 2.0, which also contain other MWE types. Some languages annotated the experimental IAV category (Dutch, 80 instances; English, 71 instances), which are not present in the other treebanks. For better comparability, these are excluded from the tables and analysis.

For all languages except English, as expected, the proportion of MWTs is quite high, around 30%. We note that our Swedish guidelines from release 2.1 give an MWT proportion that is more similar to German and Dutch than the permissive rules from release 1.3. We also note that the MWT proportion for types other than VMWEs is considerably higher for Swedish than for Dutch, with a large difference for many categories , such as NIDs (97% versus 13%) and AVs (100% versus 11%). This

	Total	%MWT	IRV	IVPC.full	IVPC.semi	LVC.full	LVC.cause	MVC	VID
Swedish 1.3	3155	51.2	7.5	46.3	18.7	13.2	0.3	0.0	14.0
Swedish 2.1	2275	27.7	13.0	25.5	17.2	22.0	1.3	0.0	21.1
Dutch 2.0	251	26.9	6.0	36.7	3.6	12.7	1.2	0.8	39.0
German 1.2	4041	30.7	8.0	43.2	4.8	7.7	0.8	0.0	35.6
English 1.3	1043	0.35	0.0	35.3	5.1	31.9	4.9	4.9	17.9

Table 6: Distribution of VMWEs for Germanic language releases in PARSEME and proportion of MWTs.

	Total	%MWT	AV	AdjID	AdpID	AdviD	ConjID	DetID	IntjID	NID	NV	Verbal
Swedish 2.1	4793	53.7	2.4	2.0	3.9	19.4	4.6	3.3	0.0	10.3	6.6	47.5
Dutch 2.0	527	25.6	2.8	1.3	3.1	20.1	1.5	4.4	0.4	6.8	4.6	54.5

Table 7: Distribution of MWEs for Germanic language releases in PARSEME and proportion of MWTs.

discrepancy calls for the synchronization of MWT guidelines across the Germanic languages.

Among the non-verbal categories, there are no major differences in distribution between Dutch and Swedish, except that ConjID and NID are more common in Swedish, which may interact with the MWT decisions. For the distribution of VMWEs, English, as expected, stands out from the other Germanic languages, with no IRVs and a high number of LVCs. Compared to Dutch and German, Swedish release 2.1 has a higher proportion of IVPC.semi and IRV, and a lower proportion of IVPC.full. We believe that this is partially due to guideline interpretations, which warrant a more in-depth comparison and discussion for the coming releases. A notable difference between the languages is that Dutch and English have annotated MVCs, whereas Swedish and German do not. This is mainly due to the annotation of expressions with (en) 'let'/ (nl) 'laten', as in *to let someone know*. While this construction exists in Swedish ('låta') and German ('lassen'), it has not been considered an MVC for those languages. The PARSEME guidelines currently have detailed language-specific MVC rules for Hindi and Chinese, with only a single rule on lexical inflexibility applicable to other languages. Our interpretation is that the *let* construction does not pass this rule and should be treated as a regular syntactic construction; however, this needs to be revisited in future guideline discussions. English also includes 'get rid' and 'cross examine' as MVCs.

6 Recommendations

Here, we summarize issues that we recommend the PARSEME community to address in order to strengthen the annotation framework.

Agreement Perform an in-depth analysis of the inter-annotator agreement across languages in release 2.0, to see if there are general difficulties across languages, and update the guidelines to address these issues.

Subtoken annotation Update the PARSEME framework so that subtoken annotation can be used for MWEs that include only a part of a graphical word (See also Savary et al., 2023b).

IVPC-IRV Create joint categories of VMWE types that can co-occur, such as reflexives and particles, and investigate how these classes interact with the IAV class.

MWTs and particles To determine whether a token is an MWT or not, and to distinguish particles from prepositions, create language-specific rules where needed, and synchronize these rules across languages.

Synchronize guidelines Synchronize the guidelines for Germanic languages, especially for IVPCs and IRVs.

MVCs Improve the language-independent guidelines for MVCs and discuss the status of the *let* construction.

7 Conclusion

We discuss the Swedish PARSEME corpus version 2.1, and give a historical overview of previous Swedish PARSEME releases. We present an overview of the Swedish annotations and discuss Swedish-specific considerations, such as the handling of particle verbs and multiword tokens. We also provide an initial comparison of the annotation of Swedish and other Germanic languages, revealing some inconsistencies. The inter-annotator agreement for Swedish is good for verbal MWEs, but quite low for other types, which were recently added to PARSEME. There is thus a need for a renewed overview of the PARSEME guidelines for Swedish as well as for other Germanic languages.

References

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47:1191–1211.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: 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.

Najet Hadj Mohamed, Cherifa Ben Khelil, Agata Savary, Iskander Keskes, Jean Yves Antoine, and Lamia Belguith Hadrich. 2025. PARSEME-AR: Arabic reference corpus for multiword expressions using PARSEME annotation guidelines. *Language Resources and Evaluation*, 59:1331–1361.

Heiki-Jaan Kaalep and Kadri Muischnek. 2006. Multiword verbs in a flective language: the case of Estonian. In *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*.

Murathan Kurfalı, Robert Östling, Johan Sjons, and Mats Wirén. 2020. A multi-word expression dataset for Swedish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4402–4409, Marseille, France. European Language Resources Association.

Eric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Proceedings of the Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51, Marrakech, Morocco.

Filip Miletić and Sabine Schulte im Walde. 2024. Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12:593–612.

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 *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Kerstin Norén. 1995. Partikelverb och lexikon: kriterier för att finna lexikaliserade partikelverb. In *Nordiske Studier i Leksikografi 3*, pages 321–330. Rapport fra Konference om leksikografi i Norden, Reykjavik, 7–10 juni 1995.

Yagmur Ozturk, Najet Hadj Mohamed, Adam Lion-Bouton, and Agata Savary. 2022. Enhancing the PARSEME Turkish Corpus of Verbal Multiword Expressions. In *18th Workshop on Multiword Expressions (MWE 2022) @LREC2022*, Marseille, France.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. SemEval-2025 task 1: AdMIRe - advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, and 6 others. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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 *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.

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, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, and 9 others. 2023a. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Agata Savary, Manon Scholivet, Carlos Ramisch, Takuya Nakamura, Eric Bilinski, Sara Stymne, Voula Giouli, Stella Markantonatou, Vasile Păiș, Maria Mitrofan, Louis Estève, Bruno Guillaume, Verginica Barbu Mititelu, Jaka Čibej, Roberto A. Díaz Hernández, Victoria Fendel, Polona Gantar, Olha Kanishcheva, Cvetana Krstev, and 9 others. 2026. PARSEME 2.0 multilingual corpus of multiword expressions. Submitted manuscript https://gitlab.com/parseme/corpora/-/blob/master/pre-prints/PARSEME_corpus_2.0_pre-print.pdf.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023b. PARSEME meets Universal Dependencies: Getting on the same page in representing multiword expressions. *Northern European Journal of Language Technology*, 9.

Manon Scholivet, Agata Savary, Carlos Ramisch, Eric Bilinski, Takuya Nakamura, Maria Carp, and Vasile Păiș. 2026. Edition 2.0 of the PARSEME shared task on multilingual identification and paraphrasing of multiword expressions. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, Rabat, Morroco. Association for Computational Linguistics.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Peter Svenonius. 2003. Swedish particles and directional prepositions. In Lars-Olof Delsing, Cecilia Falk, Gunlög Josefsson, and Halldór Á. Sigurðsson, editors, *Grammar in Focus: Festschrift for Christer Platzack 18 November 2003*, volume II, pages 343–351. Department of Scandinavian Languages, Lund University, Lund.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Therese Lindström Tiedemann, David Alfter, Yousuf Ali, Daniela Piipponen Mohammed, Beatrice Silén, and Elena Volodina. 2024. Multiword expressions in Swedish as a second language: Taxonomy, annotation, and initial results. *Multiword expressions in lexical resources*, page 309.

Therese Lindström Tiedemann, David Alfter, and Elena Volodina. 2022. CEFR-nivåer och svenska flerordssuttryck. In *Svenskan i Finland 19: Föredrag vid den nittonde sammankomsten för beskrivningen av svenskan i Finland, Vasa den 6–7 maj 2021*, pages 218–233. Svensk-österbottniska samfundet.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2020. Annotating verbal MWEs in Irish for the PARSEME shared task 1.2. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 58–65, online. Association for Computational Linguistics.