Fitting Fixed Expressions into the UD Mould: Swedish as a Use Case

Lars Ahrenberg
Department of Computer and Information Science
Linköping University
lars.ahrenberg@liu.se

Abstract

Fixed multiword expressions are common in many, if not all, natural languages. In the Universal Dependencies framework, UD, a subset of these expressions are modelled with the dependency relation *fixed*, targeting the most grammaticalized cases of functional multiword items. In this paper we perform a detailed analysis of 439 expressions modelled with *fixed* in two Swedish UD treebanks in order to reduce their numbers and fit the definition of *fixed* better. We identify a large number of dimensions of variation for fixed multiword expressions that can be used for the purpose. We also point out several problematic aspects of the current UD approach to multiword expressions and discuss different alternative solutions for modelling fixed expresions. We suggest that insights from Constructional Grammar (CxG) can help with a more systematic treatment of fixed expressions in UD.

Keywords: Multiword expressions, fixed expressions, constructions, Swedish

1. Introduction

Multiword expressions (MWEs) are ubiquitous in many, if not all, natural languages. They are usually divided into different classes with fixed, word-like expressions at one end and flexible phrase- and clause-like expressions at the other. Common English examples of these two kinds are illustrated in (1) and (2):

(1) *at first, by and large, of course*
(2) *give X the creeps, beat around the bush*

How do you search for MWEs in a treebank annotated in the Universal Dependencies (UD) framework? That would depend on the type of MWE you are interested in. UD offers three relations to represent MWEs: *compound, flat* and *fixed* (de Marneffe et al., 2021). The first is focused on compounding of nouns and other content words, the second on fixed expressions with similar behavior as function words, and the third primarily on multiword names. For definitions see Table 1. If your interest is with the flexible ones, however, you would have to use the key words of the MWE such as *creeps* or *around the bush*, as there is no particular relations devoted to them; they are annotated the same way as compositional phrases and clauses. Alternatively, you can turn to treebanks with more flexible annotations such as those developed in the PARSEME project with special annotations for verbal multiword expressions (Savary et al., 2023a).

The stated purpose of UD is to develop crosslinguistically consistent morphosyntactic annotation for as many languages as possible. The main purposes are to support research in language typology and natural-language processing, parsing in particular. Given that MWEs sometimes show deviant morphosyntactic behaviour and that the knowledge of MWEs crosslinguistically appears to be scarce (Masini, 2019) we can argue that MWEs should be given adequate representations in UD annotation. Then it is a problem that it does not cover all types of MWEs. While this problem has been recognized (Savary et al., 2023b), no solution has been agreed upon so far.

A framework that places MWEs at the center of linguistic modelling is Construction Grammar (CxG) (Fillmore et al., 1988; Booij, 2017; Hoffmann, 2022). The most radical view of CxG holds that everything in language, from morphs to sentences, are instances of form-meaning pairs of the same basic kind, called constructions. A form is a pattern of some sort and the meaning may be more or less specific. In contrast, UD only recognizes the existence of certain MWEs and by using the syntactic level of annotation it actually blurs the fact that MWEs often have a transparent syntactic structure; MWEs don't have to be syntactically deviant.

The empirical basis of the paper is a detailed analysis of the formal and structural variation in MWEs currently annotated as *fixed* in two Swedish UD treebanks. All expressions in this dataset have been annotated for the type of variation they accept, their distribution if regarded as a UD word, and for their structure. The latter aspect takes inspiration from the treatment of MWEs in Construction Grammar, in particular the idea that structures can enter into hierarchical relations. While the data is primarily taken from Swedish they illustrate general types of problems in relation to fixed MWEs. Comparisons are made with the use of *fixed* in UD treebanks for English.

Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD 2024)
@LREC-COLING-2024, pages 33–42. May 25, 2024.

Relation	Definition
compound	any kind of word-level compounding (noun compound, serial verb, phrasal verb)
fixed	fixed multiword expression; links elements of grammaticalized expressions that behave as function words or short adverbials
flat	flat multiword expression; links elements of headless semi-fixed multiword expressions like names

Table 1: Definitions of the three dependency relations used for MWEs in UD cited from (de Marneffe et al., 2021)[266]

The paper is structured as follows. The next section provides background on fixed MWEs as found in general overviews, in Usage CxG, and in UD. Section 3 presents our dataset and how it has been annotated. In Section 4 we review a number of common types of fixed MWEs found in the dataset and discuss how they can be analysed with or without the *fixed* relation. Section 5 proposes alternative ways to annotate them in UD. Section 6, finally, holds the conclusions.

2. Multiword Expressions in Different Frameworks

A common taxonomy for MWEs splits them first into lexicalized phrases and institutional phrases (Baldwin and Kim, 2010). Only the lexicalized phrases provide examples of syntactically deviant structures. They are in turn divided into fixed, semi-fixed, and syntactically flexible. This division can be seen as points on a scale from the most rigid to the fully compositional phrases (Masini, 2019). Here the focus will be on the fixed MWEs.

(Baldwin and Kim, 2010) defines fixed MWEs as expressions *'that undergo neither morphosyntactic variation nor internal modification, often due to fossilisation of what was once a compositional phrase.'* Expanding on this definition we have identified a number of ways in which a fixed MWE can vary, which is detailed in Section 3.

An interesting aspect of this definition is that it views fixed MWEs as isolated examples. Similarity of structure to other fixed MWEs seems to play little role. However, to determine whether an expression is fixed or flexible it is important to look for structural patterns that are common to sets of expressions, a key feature of Construction Grammar.

2.1. On Constructions

There are a number of variants of Construction Grammar but all of them use a notion of construction as a pairing of form and meaning. This applies to words and morphs as well as to phrases and clauses. The form level may include phonetic and/or orthographic information as well as morphological and syntactic information. Meaning may include semantic as well as pragmatic information (Hoffmann, 2022).

The morphosyntactic information is not restricted to parts-of-speech and morphological features. Depending on the scope of a construction the application of a category may be constrained in various ways, for instance to a subset of nouns or adjectives. Moreover, constructions are related to one another via inheritance links and horisontal links. In this way a phrase that seems deviant or special may be linked to a more regular pattern as a specialisation.

2.2. An Example

There is a set of Swedish time adverbials that are marked by the simultaneous occurrence of the preposition *i*, 'in' and a final suffix *-(a)s* on the following noun. The nouns are restricted to a finite number of words referring to week-days, seasons, or parts of the day. The sufix only occurs in this pattern. All expressions of the pattern are deictic and the meaning is, roughly, a reference to the most recent period of the kind signified by the noun:

i lördags	this past Saturday
i våras	this past spring
i julas	this past Christmas
i förmiddags	this past (late) morning

It is important to note that the nouns cannot be put in other nominal positions, not even as possessive modifiers. While *-s* is a genitive suffix in Swedish, the nouns in this group are seldom seen as possessive modifiers. For example, to say the equivalent of English 'the events of Saturday', in Swedish, we need to use a definite form, *lördagens händelser*, whereas an indefinite form such as **lördags händelser* on its own is out[1].

A construction in Usage Construction Grammar (Hoffmann, 2022) representing this set of time adverbials may be written as in Table 2.

Instances of this pattern that are found in Swedish UD treebanks are all annotated with the

[1] The label *kalenderplacering.genitiv*, 'calendar placement, genitive', which is found in the Swedish Constructicon (Borin et al., 2012; Lyngfelt et al., 2018) for these expressions is therefore unfortunate.

FORM:	[i NOUN$^{1}_{temp} + (a)s$]
MEANING:	this past $TIME^{1}$

Table 2: A construction in the style of a Usage CxG. The index links the noun in the FORM part to its corresponding predicate class in the MEANING part.

relation *fixed*. While there are only a finite number of them there is a clear pattern that capture their form as well as their meaning.

In a CxG patterns can be related to each other via inheritance, or as specifications of a common more general pattern. In the example we refer to more specific variables than ordinary parts-of-speech, such as week-days or seasons. This option is not available in UD, nor is UD concerned with meanings. However, a similar reasoning can be applied by relating the expression to a more general pattern captured by the part-of-speech variables ADP and NOUN. The normal relation assigned to an adposition in UD in front of a noun is *case* and the structure of the pattern can be captured as for other prepositional phrases as shown in Figure 1. Now we capture the syntactic structure of these expressions reasonably well. However, the information that we are dealing with a fixed expression has been lost. In the current UD framework we cannot say both at the same time. In the wording of (Gerdes and Kahane, 2016) the framework has created a catastrophe.

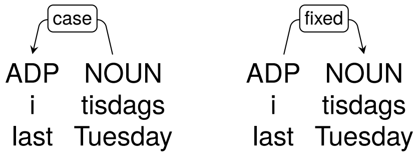

Figure 1: Two competing analyses of a fixed MWE, one as syntactically transparent and another as fixed.

Moreover, the pattern is similar to that of an adverbial expression consisting of a preposition and a non-inflected noun such as *på lördag* 'on Saturday', and *i morgon*, 'tomorrow'. Yet another similar structure employs rest morphemes such as *i går*, 'yesterday' and *i fjol*, 'last year'. Generalising further we can observe that other parts-of-speech such as adjectives can follow a preposition in expressions such as *inom kort*, 'shortly'. In UD we could view all of these as specializations of a common general pattern, ADP + ANY[2].

2.3. More on *fixed* in UD-treebanks

As stated in the introduction, *fixed* is only one of the three relations used for MWEs in UD. These relations have different properties, however. The *compound*-relation can go both to the left and the right and be embedded under a different *compound*-relation. This is not the case for *fixed* and *flat*; they are headless in principle but have the leftmost part as the head by default. Moreover, a dependent of *fixed* or *flat* can't have dependents of its own. Another UD relation with the same property is *goeswith*, which is primarily used for tokens that have been split accidentally. Structurally *fixed, flat* and *goeswith* can all be regarded as the same relation, just labelled differently for complementary information.

A special feature of *fixed*, according to its description on the UD web[3], is that it should be restricted to the most grammaticalized cases and be treated as a closed class. It is recommended that language-specific documentation is developed where the expressions for which *fixed* is applied are listed. The main reason for this is to enforce annotation consistency across treebanks in a way that can be validated automatically. This is definitely a worthy aim as the variation in its use is quite considerable. See Table 3 for figures on *fixed* in a sample of UD Treebanks, version 2.13. It can be noted that there are differences even for treebanks sharing the same language. In fact, some treebanks not shown in the table, like the Norwegian ones and UD_German-HDT do not use *fixed* at all. This shows that recommendations are motivated. It is likely that the differences are not due to language differences but to different annotation principles.

There are published lists only for a few languages, including English and Finnish. The English list has some 40 items, Finnish has around 90. The number of fixed expressions in the largest Finnish treebank is larger, however.

The idea to restrict fixed MWEs in UD to a smaller group raises the question how well it aligns with the notion of a fixed MWE as characterized in general works on the topic such as (Baldwin and Kim, 2010). Is it actually possible to find general criteria that could restrict the application of *fixed* in a principled way? This is investigated in Section 4.

3. Dataset and annotation

The main empirical data for the analysis are taken from the two Swedish UD treebanks UD_Swedish-Talbanken and UD_Swedish-Lines of version 2.13. In addition, we have looked at the list of proposed

[2] Instead of ANY we could specify a disjunction of UPOS categories.

[3] https://universaldependencies.org/u/dep/fixed.html

Treebank	Listed	In TB	%
UD_Dutch-Alpino	-	1161	2.75
UD_English-EWT	44	40	0.50
UD_English-GUM	44	44	0.64
UD_English-LinES	44	117	1.06
UD_Finnish-FTB	90	198	0.66
UD_Finnish-FTB	90	27	0.37
UD_French-Rhapsodie	-	70	2.62
UD_French-Sequoia	-	82	1.45
UD_Icelandic-IcePaHC	-	20	0.14
UD_Icelandic-Modern	-	2	0.05
UD_Italian-ISDT	-	79	0.66
UD_Italian-TWITTIRO	-	23	0.55
UD_Swedish-LinES	-	117	1.59
UD_Swedish-Talbanken	-	392	3.12

Table 3: Usage of *fixed* in a sample of UD treebanks. The column **In TB** shows the number of different types of MWE that are found in the treebank, while the column **%** shows the percentage of all tokens in the treebanks that carry *fixed* as their dependency.

English fixed expressions[4].

Together the two Swedish treebanks have 439 different MWEs annotated with *fixed*. Of these 71 are common to both treebanks, and 216 are hapaxes. For a few common MWEs, such as *som om*, 'as if', and *mer än*, 'more than' the two treebanks have made opposite decisions. Yet, the large majority satisfies the loose criterion of being multiword sequences that behave as function words, adverbs, or are special in some other way. As the treebanks are not very big we can safely assume that there are many more expressions that satisfy the same tolerant criteria as those in the treebanks. To compare, Wikipedia has 649 expressions listed under the label Swedish idioms and a recent dictionary of Swedish idioms (Luthman, 2020) contains 5000 items, although the majority of these are flexible.

Starting with the properties listed in the definition above (Baldwin and Kim, 2010) other properties were added as cases were found. Previous work on idioms in Swedish such as (Anward and Linell, 1976; Sköldberg, 2004) have largely focused on flexible idioms, but they define various criteria for recognizing MWEs including fixed expressions that we have considered. The expressions in the dataset have also been checked against larger Swedish corpora and concordances generated from the Korp interface[5] on news media. In the end we came up with 13 different properties as listed below. The first two relate to the expression's function and pattern, while the rest focus on some aspect of variation.

- **UPOS tag:** Part-of-speech if regarded as a single UD word, using the UPOS set of tags.
- **Syntactic pattern:** The syntactic pattern is expressed in terms of UPOS tags and regarded as the best generalisation of a more specific CxG pattern
- **Morpheme status:** Takes the values Roots, Inflected, Foreign, Abbr(eviation) and Special where Special includes rest morphemes and rare (obsolete) inflections.
- **Inflection variation:** Does any part of the expression allow inflectional variants? Yes or No.
- **Internal modification:** Does any part allow one or more modifiers? Yes or No.
- **Synonyms:** Is it possible to replace any part with synonyms? Yes or No.
- **Iterability:** Can a part be repeated? This is rare but occurs for several expressions that signify repeated events: *om och om (och om) igen*, 'again and again (and again)' Yes or No.
- **Order change:** Can the order among parts be different? Yes or No.
- **Optional part:** Is any part optional, or can an optional part be added? The answer is Yes or No and an example is *under det (att)*, 'while'.
- **Separability:** Can (or must) some part be separated from the rest by other material? Possible values are No, Obligatory, and Optional.
- **Idiom part:** Does the expression mainly occur as part of a longer idiom, in the treebank and generally? If so the value is Yes, otherwise No.
- **Abbreviation:** Does an abbreviated form exist? Yes or No.
- **Collapsibilty:** Does a single token equivalent exist? Often this is the result of omitting spaces as in *över allt : överallt*, 'everywhere'. Yes or No.

Every expression in the dataset has been described with these attributes. An illustration is given in Table 4 for the expression *i våras*[6]. Descriptions for the full dataset can be found in the supplementary material.

4. Types of fixed MWEs

Given the requirement that fixed expressions in UD should be a restricted closed class we want to

[4] https://universaldependencies.org/en/dep/fixed.html

[5] https://spraakbanken.gu.se/korp/

[6] In the expression *i fjol våras*, 'the spring of last year', we do not regard *fjol* as a modifier of *våras* but rather see it as a compound of two expressions *i fjol* and *(i) våras*

Attribute	Value	Comment
UPOS tag	ADV	
Pattern	ADP NOUN	
Morpheme status:	2:Special	*våras*
Inflection variation	No	
Modification	No	
Synonyms	No	
Iterability	No	
Order change	No	
Optional part	Yes	*i fjol våras*
Separability	No	
Idiom part	No	
Abbreviation	No	
Collapsible	No	

Table 4: Description of the Swedish expression *i våras*, 'this (past) spring' with respect to structure and variability.

reduce the number of expressions currently annotated with *fixed* in the Swedish treebanks. This entails two main things: identifying criteria that make *fixed* correspond well to a natural class of fixed expressions, and finding alternative dependency analyses for those expressions that are removed.

There are many different types of expressions in the dataset and the available space does not allow us to discuss all of them. We start with one type of variation that may be more common in a Swedish dataset than for other languages, the alternative renderings captured by the property of Collapsibility.

4.1. Collapsible MWEs

Swedish language planning authorities are generally quite tolerant towards variation in written Swedish. As a result many multiword expressions have alternative renderings as single tokens or, in case of three-part expressions, two tokens. As UD maintains that tokenisation should follow the orthographic rendering as far as possible, in particular that in-token spaces should be avoided, these expressions pose a special challenge.

In the dataset we find 75 collapsible MWES, which is about 17% of all. The large majority of them has an alternative rendering by omitting spaces. Examples are *till buds :: tillbuds*, 'at hand', *i dag :: idag*, 'today', *över huvud taget :: överhuvudtaget :: överhuvud taget*, 'actually'. The share of a certain rendering differs with individual expressions. We have investigated their distribution in two subsets of the Swedish Gigaword Corpus (Rødven Eide et al., 2016), news and fiction. The numbers support a division into three different groups, one where the the MWE rendering is much more common, one where the spaceless rendering is much more common, and one where the two renderings are about equally common. However, the relevance of this variation lies not so much in the exact proportions but that both renderings occur. A treebank should as far as possible assign the same analysis to both alternatives; they contain the same lexemes, but are just written differently. If spoken they would come out identical. Compare the two renderings below of the same sentence:

(3) *Hon kan när som helst komma i kapp*
(4) *Hon kan närsomhelst komma ikapp*
'She may catch up at any moment'

Given the aversion against token internal spaces in UD one option is to regard the multipart variants as basic and treat the shorter variants as multiword tokens. This solution aligns well with the long-term proposal for modelling synthetic compounds in UD put forward by (Savary et al., 2023b). A drawback is of course that this solution is sofar unseen in any Swedish treebank. Conversely, the existece of the single-token forms may be taken as an argument that they are perceived as single lexemes.

Using multiword tokens for the tokenisation of sentence (4) we would get the tokenisation in Table 5.

1	Hon	hon
2	kan	kunna
3-5	närsomhelst	_
3	när	när
4	som	som
5	helst	helst
6	komma	komma
7-8	ikapp	_
7	i	i
8	kapp	kapp

Table 5: Proposed tokenisation for single token alternatives to Swedish fixed MWEs.

4.2. Syntactic alternatives to *fixed*

For many of our expressions in the dataset we can find patterns that are shared with other expressions, as in Section 2.2. We may distinguish self-contained patterns from patterns with outward-looking parts. In the first type all included words except one have their head within the pattern. They are easy to provide a syntactic analysis for. With outward-looking parts two words have their heads outside of the pattern. Usually one of them is the last token which may be a preposition, subjunction or conjunction.

Self-contained expressions. The most common type of self-contained fixed expression in the dataset consists of a preposition followed by an uninflected noun. There are 66 such **prepositional**

phrases with examples such as *i dag*, 'today', *i allmänhet*, 'in general'. Other two-part expressions beginning with a preposition has a noun in definite form as head, *på vippen*, 'on the verge', an adjective, *på nytt*, 'anew', or a pronoun, *före detta*, 'ex-'. For some the UPOS is even hard to determine *på sistone*, 'lately', *på glänt*, 'slightly open', as the token is invariable and only occurs in this special expression. In addition there are three-part expressions with a nominal head of some sort. Taken together prepositional phrases account for almost 40% of the expressions in the dataset.

The syntactic structure of these prepositional phrases need not deviate from compositional phrases of the same patterns, see Figure 2. The fact that the correct UPOS tag for some words may be hard to determine does not prevent the assignment of an appropriate structure either. Moreover, the treatment of prepositions would actually be more consistent if they always are assigned the relation *case* when followed by a candidate head word.

We note that no more than four of the English MWEs in the list of English fixed MWEs are prepositional phrases, (*in order, of course, in case, at least*) and see this as support for treating prepositional phrases as non-fixed in the general case.

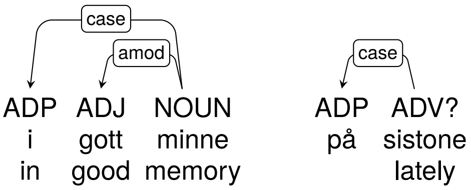

Figure 2: Syntactic dependency analysis for expressions beginning with a preposition.

Coordinations can be handled in the same way as prepositional phrases, since their syntactic structure is transparent when a coordinating conjunction is present. The most common type coordinates two adverbs but Swedish also shows instances of coordinated prepositions. Both structures can be viewed as specializations of a more general pattern for coordinations that need not require the two conjuncts to have the same part-of-speech. Thus, a fixed MWE as English *by and large* could be dealt with in the same way. The proposed structures are shown in Figure 3.

Another common type of pattern has an adverb or adjective as head modified by another adverb. Examples are *så pass (stor)*, 'that (big)' and *illa nog*, 'bad enough'. They also can be assigned the same structure as their compositional counterparts with the adverb serving as an *advmod*.

There are also expressions where an adverb seemingly modifies a preposition as in *in i*, 'into' or *fram till*, 'up to'. This is generally forbidden in the UD framework. To avoid annotating the adverb as a modifier we may regard the two parts as independently modifying the head.

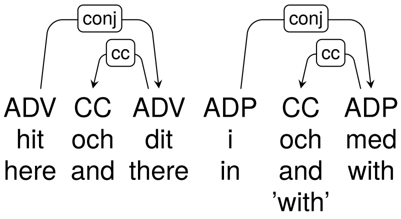

Figure 3: Syntactic dependency analysis for expressions employing coordinations.

Some of the expressions annotated with *fixed* end with a verb form of some sort most often a participle. Examples are *strängt taget*, 'actually', *allvarligt talat*, 'seriously speaking'. Regarded as verb phrases these expressions have obvious syntactic annotations: the participle is the head and the adverb an adverbial modifier. In relation to its context it may be annotated as an adverbial clause, *advcl*.

Outward-looking parts. A number of two- or three-word expressions have a last part that normally begins a phrase or clause of some sort. This applies to expressions ending in a preposition, a subjunction or one of the comparative conjunctions *än*, 'than' and *som*. 'as'.

The most common type of these are three-part sequences starting and ending with a preposition and a noun or nominal word in between. There are 48 expressions of this type in the dataset; examples are *på grund av*, 'because of', and *i samband med*, 'in connection with'.

Sometimes the final preposition introduces an optional phrase. An example is *med hjälp av*, 'with the aid of', where *med hjälp* can act as an adverbial phrase on its own. In those cases it is perfectly reasonable to view the noun in the middle as the head. See Figure 4. If the preposition is required, however, as in *på grund av*, 'because of', this solution can be questioned. We note though that in the English list of *fixed* expressions, this type of three-part expression is rare. For example, *in spite of* is not included so that *spite* comes out as the head of a noun phrase such as *in spite of the problems* giving the same structure as in Figure 4[7]

Expressions ending with a subjunction are also quite common; in the data set we find 9 ending in *att*, 'that', 2 ending in *om*, 'if', and 10 ending in *som*, 'as'. Here a different analysis may be advocated: assigning the different parts separate functions as

[7] For example, the tree with sent-id 'weblog-blogspot.com_alaindewitt_20060924104100_ENG_20060924_104100-0031 in en_ewt-ud-train.conllu

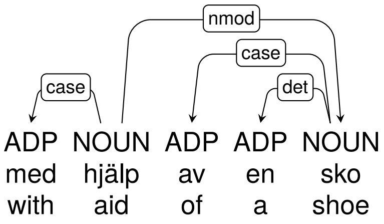

Figure 4: Syntactic relations for the three-part expression *med hjälp av*, 'with the aid of'.

mark or *case* depending on the part-of-speech. For example, in the case of *som om* and, similarly, *as if*, one may argue that each of the two parts has a function of its own. The first, *som/as* indicates that we are dealing with a comparison, the second, *om/if* that we are dealing with something unreal or assumed. In Swedish, such an analysis gains some support from the fact that the *if*-clause in certain circumstances can be replaced by a clause without the subjunction:

(5) *Han beter sig som vore han ...*
'He behaves as were he ...

(6) *Han uppför sig som om han var ...*
'He behaves as if he were ...

There are eight expressions ending with the comparative conjunction *än*, 'than'. The majority are introduced by an adjective or adverb in comparative form, such as *mer än*, 'more than', *lägre än*, 'lower than' or 'less than'. The comparatives actually all accept modifiers such as *mycket*, 'much', or *lite*, 'a little', and for this reason they may not qualify as fixed expressions. Syntactically they can be treated as other expressions with outward-looking parts, letting the conjunction find its head to the right and the whole of that complex be a dependent to the word in the comparative.

In the English treebanks the expressions *more than* and *less than* are regarded as fixed when they modify a quantity as in *more than 90 percent* but not in other contexts. This is a bit awkward as there is no difference in the possibility of adding the modifier *much*: *much more than I have* and *much more than 90 percent* sound equally well-formed.

Similar arguments apply to comparison using the conjunction *som*, 'as'. They are common both in our dataset and in the English list. But they often share a pattern as the English *as many/much/few/little as* where virtually any adjective and a number of adverbs may occur in the middle. This indicates that we are dealing with a construction that can be annotated as such with the adjective/adverb as the head.

4.3. Types based on variation

Another basis for grouping expressions is the amount of variation that they admit. For our dataset we may distinguish three groups. At one end there are expressions with no or almost no variation based on the variational attributes that may be called **rigid**. At the other end we find several expressions that allow inflectional variation, replacement with synonyms and/or internal modification. Those will be called **semi-flexible**.

Semi-flexible expressions. 57 of the expressions that are currently annotated with the relation *fixed* can actually be varied enough to be called semi-flexible. This applies to expressions with parts that can be inflected in accordance with their part of speech, be replaced by synonyms, and/or take modifiers. Expressions of this type are

- *när det gäller*, 'concerning', (inflectional alternatives *gällt, gällde*, other alternative *vad det gäller*.
- *vem som helst*, 'whoever', (modifiers *fan*, 'the devil', *av dem*, 'of them', and similarly for other expressions of the same pattern: *när som helst, var som helst*, 'whenever', 'wherever'.
- *den här*, 'this', *den där*, 'that'. with variants *de, den, det, dom* for the first part, and *här, där* for the second part. The second parts are also found after *så, sådan, sådant, sådana* giving expressions meaning 'like this' or 'like that'.

For these types we argue that they shouldn't be regarded as fixed MWEs at all because of the amount of variation they accept. Instead syntactic analyses need to be found.

Rigid expressions There are 96 expressions in the dataset that show no variation at all. By including those that are collapsible and/or have an abbreviated form we reach 146 expressions. The most common are *som om*, 'as if', *så att*, 'so that', *i dag*, 'today', *därför att*, 'because', *på grund av*, 'because of', *för att*', '(in order) to', *i stället*', 'instead', *till exempel*', 'for example', all of which occur more than 30 times in the treebanks. We note that in case the English counterparts are MWEs they are listed as *fixed* for English[8]. Rigidity may thus be regarded as a characteristic property of expressions to be annotated as *fixed*.

Also included in this group are expressions from other languages and abbreviations. They are not so numerous but illustrate general types of interest.

[8] In the case of *in order to*, however, only *order* is taken as a dependent of *in*, while *to* finds its head in a verb to the right

There are expressions of Latin origin such as *a priori* and *vice versa* and one of English origin, *to date*. Abbreviations include short forms of academic degrees such as *med lic*, 'licentiate in medicin' and common phenomena in academic prose, such as *a. a.*, short for 'anfört arbete', and a counterpart to the Latin 'op. cit.'.

In UD foreign material may be annotated in different ways. If regarded as a borrowing it should be given a suitable UPOS tag and different parts be connected via the relation *flat* (sic!). If regarded as truly foreign each part should have the UPOS X, and, in addition, carry the feature information FOREIGN=Yes. The parts should again be connected via *flat*. With one exception, the expression *ad calendas graecas*, the examples in the dataset are sufficiently common in Swedish to be regarded as borrowings. Depending on their status as functional (*vice versa*) or not (*ad hoc* they could fit either *fixed* or *flat*.

Abbreviations should be marked by the feature Abbr=Yes. The UPOS tag should reflect the part-of-speech of the abbreviated word. The expanded versions of our two examples both consist of an adjective and a noun so the dependency analysis could use the *amod*-relation rather than *fixed*. See Figure 5.

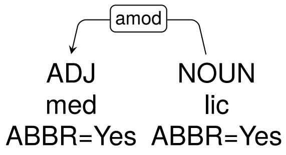

Figure 5: Dependency analysis of the abbreviated title *med lic*, 'licentiate in medicine'.

4.4. Candidates for the *fixed* list

. A large number of MWEs currently marked as *fixed*can be excluded as candidates for the list of *fixed* expressions on the basis of their morphosyntactic variation. With a fairly strict criterion on rigidity, not excluding MWEs that are collapsible or can be abbreviated, there are 146 items left. By considering that *fixed* should be restricted to items with function word distribution another seven can also be removed, leaving 139. This is still a large number, however, especially considering that the treebanks only cover a subset of the Swedish MWEs. On the other hand, many of them have a transparent syntactic structure; being self-contained expressions of the kinds described in Section 4.2. By consistently preferring a headed structure when the MWE satisfies such a pattern the numbers can be reduced further. Other types that may be excluded are those where different parts of the MWE can be separately annotated with a dependency to an outside head as was argued in the case of *som om*, 'as if' and as is done in English treebanks with many MWEs of the form 'ADP NOUN ADP'.

As UD is reluctant to see function words as heads the most likely MWEs to put on the list of items annotated with *fixed* are two-word MWEs ending in a preposition or a subjunction. Examples of the first kind are such *in i*, 'into' and *rent av*, 'actually' and of the second *så att*, 'so that', *för att*, '(in order) to', and *ifråga om*, 'as regards'. Another set of likely candidates come from adverbial and prepositional MWEs where the head word is not an adverb or a preposition as for *tack vare*, 'because of', *till synes*, 'seemingly'.

5. Alternative annotations of fixed expressions in UD

The current UD guidelines on fixed expressions hide their, in many cases, apparent syntactic structure. (Gerdes and Kahane, 2016) have pointed out this as a 'catastrophe' problem and makes a proposal to subcategorize syntactic dependencies with a special identifier such as *mwe*. A disadvantage of this solution is that it will profilerate the *mwe* subcategory in the trees. Moreover it annotates the property of being a multiword expression at a single level to the exclusion of other properties that an MWE may have. The proposal in (Kahane et al., 2017) to insert extra lines for fixed expressions such as *top of the range*, which may carry a dependency relation of its own seems more accurate for capturing the lexical character of fixed expressions.

An alternative is to unify the shallow headless relations to one, say *flat*[9], and treat a property such as fixedness with a feature in the same way as is done with foreignness. This would make the annotation similar to that for split words, where the relation *goeswith* is used in tandem with the feature Typo=Yes[10]. The features for a fixed MWE could then be applied to its head and be interpreted as including the dependents by default.

This solution would also solve the problem of choosing between *fixed* and *flat*. As shown above the properties of phrases as being fixed, abbreviated, or from a different language sometimes converge. An expression such as *vice versa* could actually be annotated as foreign and fixed at the same time. Then the *fixed* is in conflict with *flat* which is recommended for foreign material. Annotating these properties at the level of features allows them to be combined.

[9] A similar proposal is made in (Savary et al., 2023b) using the label *headless*.

[10] https://universaldependencies.org/u/dep/goeswith.html

A third more radical alternative is not to deal with fixed expressions at all in the current UD format. While there is a need to mark headlessness in the syntactic trees, it is evident that not all kinds of MWEs can be handled as part of UD dependency trees. It is also evident that the current feature annotation is insufficient. It is restricted to words and thus cannot cover subtrees with one feature. The CUPT format (CoNLL-U Plus Format) as used by the PARSEME:MWE framework for annotating verbal MWEs allows more complex feature annotation and may be used for many types of MWEs including fixed expressions. This seems to be the future that is also envisioned by (Savary et al., 2023b).

With this alternative appropriate syntactic dependencies need to be found. We have suggested that a Construction Grammar perspective on fixed MWEs is helpful for this purpose. UD has a general principle of a tight relation between UPOS categories and dependency relations. This principle could be extended to UPOS sequences that share enough common features to be related hierarchically to a dependency template as suggested in Section 2.2.

6. Conclusions

We have analysed 439 expressions currently annotated as fixed expressions in Swedish UD treebanks with the aim of producing a well-defined subset that meets UD requirements on the use of the relation *fixed*. We have found a way to reduce this set by closely studying their variational properties and the structural patterns that they share. Although we find a number of rigid MWEs, i.e., expressions admitting no or almost no variation at all, they often have a transparent syntactic structure which is not accounted for when *fixed* is used. And many of them share structure with other MWEs. These structures can be represented in more detail in Construction Grammar frameworks, as we have shown with examples. Although UD does not allow such detail we can nevertheless often generalise the structure to something that can be expressed in UD-terms. Moreover, to capture all kinds of MWEs, whether fixed or flexible, requires a more versatile format than CoNLL-U such as the CUSP-format used for annotating verbal MWEs in the PARSEME:MWE project.

Annotating fixed expressions with a specific relation as part of the dependency structure, as is currently done in UD, prevents the annotation of its syntactic structure. A better solution would be to isolate the structural properties of *fixed*, which it shares with other UD relations such as *flat* and *goeswith*, in a single relation and use features to indicate the character of the expression, something which now is done only for typos.

Another problem we discovered, which may be specific to Swedish, is the large numbers of collapsible MWEs. The best solution we could propose for these, in order to ensure that the dependency analysis would come out the same whether the MWE is collapsed or not is to make use of UD's provision of multiword tokens.

7. Acknowledgements

I am indepted to Joakim Nivre for valuable comments on an earlier version of the paper and to an anonymous reviewer for detailed critique and pointing me to the article (Savary et al., 2023b). The work has been partially funded by the Swedish Research Council through the Swe-Clarin part of the Swedish National Language Bank.

8. Optional Supplementary Materials

A spreadsheet with our analysis of 439 MWEs currently analysed as fixed in Swedish treebanks is provided as supplementary material.

8.1. Extra space for ethical considerations and limitations

This work is based on open resources and, as far as we can see, pose no ethical problems. A limitation is that it is based on treebank data from one language only and some comparisons with English data. We are certain, though, that the types of problematic multiword expressions discussed here can be found also in other UD treebanks. However, the restriction to one language means that the list of types is likely to be incomplete.

9. Bibliographical References

Jan Anward and Per Linell. 1976. Om lexikaliserade fraser i svenskan (lexicalized phrases in Swedish). *Nysvenska studier. (Studies in Modern Swedish)*, 55/56:77–119.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Boca Raton, USA,.

Geert Booij. 2017. Construction morphology. In Mark Aronoff, editor, *Oxford research encyclopedia of linguistics.* Oxford University Press, New York.

Lars Borin, Markus Forsberg, Benjamin Lyngfelt, Julia Prentice, Rudolf Rydstedt, Emma Sköldberg, and Sofia Tingsell. 2012. Growing a Swedish constructicon in lexical soil. In *Proceedings of SLTC 2012. (The Fourth Swedish Language Technology Conference)*, pages 10–11.

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

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

Kim Gerdes and Sylvain Kahane. 2016. Dependency annotation choices: Assessing theoretical and practical issues of Universal Dependencies. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 131–140, Berlin, Germany. Association for Computational Linguistics.

Thomas Hoffmann. 2022. *Construction Grammar: The Structure of English*. Cambridge University Press.

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.

Hans Luthman. 2020. *Svenska idiom: 5000 vardagsuttryck (Swedish idioms: 5000 everyday expressions)*. Folkuniversitetets förlag.

Benjamin Lyngfelt. 2021. Valens och konstruktioner - om samspelet mellan lexikon och grammatik (valency and constructions - on the interplay of lexicon and grammar. In Johan Brandtler and Mikael Kalm, editors, *Nyanser av grammatik. Gränser, mångfald, fördjupning*. Studentlitteratur.

Benjamin Lyngfelt, Linnéa Bäckström, Lars Borin, Anna Ehrlemark, and Rudolf Rydstedt. 2018. Constructicography at work: Theory meets practice in the Swedish constructicon. In B. Lyngfelt, L. Borin, K. Ohara, and T. T. Torrent, editors, *Constructicography: Constructicon development across languages*, pages 41–106. John Benjamins, Amsterdam.

Francesca Masini. 2019. Multi-word expressions and morphology. In Mark Aronoff, editor, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press,.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The Swedish Culturomics Gigaword Corpus: A one billion word Swedish reference dataset for nlp. In *Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts*, pages 3–36, Krakow, Poland. John Benjamins Publishing Company.

Ivan A. Sag, Timothy Baldwin, Frances 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, and et al. 2023a. Parseme corpus release 1.3. In *Proceedings of The 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35.

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(1).

Emma Sköldberg. 2004. *Korten på bordet: Innehålls- och uttrycksmässig variation hos svenska idiom (Cards on the table. Variations in content and expression in Swedish idioms)*. Ph.D. thesis, University of Gothenburg.