NoVRol: A semantic role lexicon of Norwegian verbs

Henrik Torgersen[1], **Erlend Ø. Ravnanger**[1], **Lars Hellan**[2], **Dag T. T. Haug**[1]
[1] University of Oslo, [2] The Norwegian University of Science and Technology
hatorger@uio.no, erlenora@student.iln.uio.no, lars.hellan@ntnu.no, daghaug@uio.no

Abstract

In this paper, we describe NoVRol, a semantic role lexicon of Norwegian verbs. We start from the NorVal valency lexicon, which describes the syntactic frames of 7.400 verbs. We then enrich each of these frames by annotating, based on the VerbNet annotation scheme, each argument of the verb with the semantic role that it gets. We also encode the syntactic roles of the arguments based on the UD annotation scheme. Our resource will faciliate future research on Norwegian verbs, and can at a future stage be expanded to a full VerbNet.

Keywords: VerbNet, argument structure, semantic roles

1. Introduction

Semantic Role Labeling (SRL) is the task of identifying *Who did what to whom?*, i.e. what roles each of the argument entities bear in the event described by a predicate. Traditionally used for semantic representations, precise search, and in questions-answering systems, SRL has found new applications in the neural age, e.g., for image captioning (Chen et al., 2021) and computer vision (Sadhu et al., 2021), where it serves to structure the computer's interpretation of video. At the same time, the mapping from syntactic structure to semantic roles has also attracted considerable interest in theoretical linguistics with important contributions such as Fillmore (1968) and Levin (1993).

However, for Norwegian – otherwise a relatively well-resourced language – there are no datasets available that can support such research, whether practically or theoretically oriented. In this paper, we report on NoVRol, a resource which links the syntactic and semantic patterns of ca. 7.400 Norwegian verbs. For the semantic role annotation, we draw on the annotation standard of the English VerbNet (Schuler, 2005), with some modifications. For the syntactic side, we use the valency lexicon developed by Hellan (2022, 2023). In addition, we map these syntactic patterns to Universal Dependencies (UD, de Marneffe et al. 2021), thereby adding an important, lexical semantic resource to UD. UD currently containts more than 200 treebanks in more than 100 languages and has become the de facto standard for syntactic annotation and parsing. It is therefore a natural starting point for multilingual semantic parsing and many recent efforts in this direction have drawn on UD (Reddy et al., 2017; Poelman et al., 2022; Findlay et al., 2023).

We believe NovRol will be an important resource for future work in Norwegian NLP and linguistics. Moreover, because we follow the VerbNet annotation standard, we can expand the resource to a full VerbNet in future research by adding other information found in VerbNets such as selectional restrictions and event structure/logical form

The structure of this paper is as follows: in Section 2 we discuss related work on VerbNets and on the Norwegian valency lexicon. In Section 3 we describe the annotation procedure. Section 4 then discusses how our work fit in the broader picture of lexical resources for UD. Section 5 provides statistics about the data set, and Section 6 concludes and offers perspectives for further research.

2. Related work

2.1. Other VerbNets

The first VerbNet was developed for English (Schuler, 2005). It contains for each verb the semantic roles, selectional restrictions, syntactic frames and a semantic representation, as well as links to other lexical resources such as WordNet, PropBank and FrameNet. Also, verbs in VerbNet are organized in classes based on their valency alternation patterns, originally following the classes from Levin (1993) and later extended with more classes. The English VerbNet is therefore a comprehensive resource for the exploration of English verbs and their valency patterns. It has for example been used for the study of caused motion constructions (Hwang and Palmer, 2015). It has also been used in applications for word sense disambiguation, figurative language detection and it forms the basis for the semantic roles used in the Discourse Representation Structures of the Groningen Meaning Bank (Abzianidze et al., 2017). The latter was a particularly important motivation for our work, which is part of a project on UD-based semantic parsing.

There have been several efforts to create VerbNets for other languages, the most complete ones probably being those for Arabic (Mousser, 2010) and French (Pradet et al., 2014). Both of these started from the information in the English VerbNet and transfered this to the target languages semi-automatically. That is, they build on the idea that verb classes can be reliably identified across languages (see Majewska and Korhonen (2023) for a recent survey of this kind of work). This allowed for the relatively quick creation of rich resources with information comparable to that available in the original English VerbNet.

Our own approach was different, both because the goals were more modest – the immediate goal being a standard for semantic roles of Norwegian verbs for use in semantic parsing – and because Norwegian already has a rich resource for verbal valency, NorVal. It was therefore more natural to start from this Norwegian-specific resource and add information about semantic roles based on the English VerbNet, even if this meant that we gave up on structuring the resource around valency classes as in the English VerbNet and also do not provide much of the other information such as semantic structure or selectional restrictions. Some of this information is available in NorVal and can be more properly integrated in this resource to yield a richer VerbNet. We will come back to these opportunities later.

2.2. NorVal

NorVal (Hellan, 2022, 2023)[1] is a resource representing valency properties of 7,400 Norwegian verbs, theoretically based on the formal model outlined in Hellan (2019), and developed in parallel to a computational grammar of Norwegian, *NorSource*,[2] from which the verb inventory and many of the formal specifications have been ported.

The resource identifies 340 types of valency frames covering the valency properties of the verbs, and identifies for each verb lexeme which valency frames it can take. A compact notation system called *Construction Labeling* (abbreviated 'CL'), is used for classifying the frame types. More than half of the verbs take more than one frame, and the construct ⟨Verb, Valency frame taken by the verb⟩ is called a 'lexically instantiated Frame Type', abbreviated *lexval*. In the overall system there are currently 17,200 lexvals distributed over the 7,400 lexemes. Each lexval is illustrated by a 'Minimal Sentence' instantiating the lexval. A set of lexvals belonging to the same lexeme is called a *valpod*. To illustrate these constructs and their notation, (1-b) is the CL representation of the construction type: 'Expletive subject – direct object - extraposed declarative clause', exemplified by the verb *ane* ('dawn on') in (1-a):

(1) a. Det aner dem at krisen
it.expl dawns them that crisis.def
kommer
comes
'they have a hunch that the crisis is coming'
b. *trExpnSu-expnDECL*

The part 'trExpnSu' of this label is called the 'global label' of the lexval, indicating the valency frame as a whole (viz., *transitive with an 'extraposed' clause linked to subject position*), and the part 'expnDECL' is called an 'argument label' as it specifies one of the arguments.

The full set of constructions in which *ane* can be used, i.e. its valpod, is shown in Table 1. A valpod is verb-specific, but if one abstracts away the lexical item, one gets what may be called a *valpod type*, characterized by the set of frame types; such sets may be compared across the lexemes, and may be expected to provide a step toward a modeling of the notion of *verb classes* in VerbNet, based on defining valpod types across verb lexemes where a high degree of overlap in the members constituting a given set of valpods will qualify the lexemes characterized by these valpods for membership in a verb class.

NorVal provides syntactic frames for verb lexemes. Homonyms are distinguished in the verb list by hyphenated numbers, so that, e.g., *koste-1* represents the lexeme with meaning 'cost' and *koste-2* represents the lexeme with meaning 'brush'. Sub-senses of lexemes, on the other hand, are not originally recognized, but with the role annotation of this project, many cases are represented through added lexvals. Many aspects of what may be called 'basic logical form' are reflected in the frame type labels, such as causativity, semantic government, and infinitival control, and, most relevant to semantic role labeling, *participant* status, with semantic role features for *directionality* and *locativity*.[3] For example, the construction in (2-a) has the CL formula in (2-b).

(2) a. katten smyger seg langs muren
cat.def slithers refl along wall.def
The cat slithers along the wall.
b. *tr-obRefl-obDir*

This illustrates how a role specification is made by

[1] https://github.com/Regdili-NTNU/NorSource/tree/master/NorVal_files

[2] Hellan and Bruland (2015), Beermann and Hellan (2004), https://regdili.hf.ntnu.no/linguisticAce/parse

[3] These features are illustrated in the parse interface for NorSource at https://regdili.hf.ntnu.no/linguisticAce/parse, using the 'Minimal Recursion Structure' ('MRS') (Copestake, 2002) formalism for semantic representation; for a sentence like *Gutten løper til huset 'the boy runs to the house'*, for instance, a feature 'role' indicates the variable representing *huset* as 'endpoint', reflecting the combination of specifications of the verb and the preposition.

lexvals	explanation
ane intr	intransitive
ane tr	transitive
ane tr-obDECL	declarative complement
ane tr-obINTERR	interrogative complement
ane tr-suDECL	declarative subject complement
ane tr-suINTERR	interrogative subject complement
ane trExpnSu-expnDECL	transitive with expletive subject and extraposed declarative complement
ane trExpnSu-expnINTERRwh	transitive with expletive subject and extraposed wh-interrogative complement

Table 1: valpod for *ane*

appending the role indicator (*Dir*) to the argument label (*ob*), indicating that the object plays a directional role. The system also defines labels like *suAg* (subject agent), *suTh* (subject theme) and *obTh* (object theme),[4] and therewith valpods such as (3).

(3) a. *<V intr-suTh, V tr-suAg-obTh, ...>*
b. *<V intr-suAg, V tr-suAg-obTh, ...>*

The constellation in (3-a) could be used to characterize transitivity alternations like those found with verbs like *break*, as in *he broke the glass* vs. *the glass broke*, and the one in (3-b) to characterize alternations residing in constructions of 'object implicitation' like in *he is eating* vs. *he eats the bread*. While NoVRol uses a different notation, it provides a full scale encoding of roles for most aspects of verb semantics. Thus, two-membered valpods alone obtain for 1,500 verbs in NorVal, and many of them could be characterized as either of the options in (3). An assembly of valpods so annotated would throw interesting light on how common either of these types of transitivity alternations are in a representative valency inventory of a language. This illustrates how semantic role annotation, as undertaken in this project, provides an interesting addition to the specification inventory of NorVal.

3. Annotation

NoVRol includes every lexval in the NorVal database. Each verb and its arguments, as indicated in its lexvals, was annotated semantically according to the annotation guidelines for the English VerbNet.[5] The valpod for *ane* from Table 1 is shown annotated in Table 2.

We see that sometimes a single lexval needs to be assigned multiple semantic frames. For example, *ane tr(ansitive)* can take both an experiencer subject and a stimulus object and the inverse mapping. This is a special case because there is no associated meaning difference; in many other cases, the verb meaning changes slightly. For example, the verb *fortelle*, just like English 'tell' has among its syntactic frames one where it takes a subject, an object and a complement clause, but semantically, these can be agent–recipient–topic ('He told us that…') or pivot–experiencer-topic ('This tells us that…). Such multiple semantic frames are a major source of interannotation disagreement, as we will see below.

This yields a database of verb classes according to semantic roles, but without the in-depth listing of syntactic configurations or event structure specification provided by the English VerbNet. These are both aspects that can be added at a future stage. For the purpose of VerbNet as a lexical resource for a syntactic parser, this strategy has the advantage of allowing for the quick annotation of a large number of verbs. A test set of 800 (ca. 5% of total) verbs was reserved for evaluating inter-annotator agreement. In addition to the role annotation, we also give the Universal Dependencies labels for the different arguments. This section outlines how the annotation was done and comments on certain aspects of the results: differences between English and Norwegian; semantically ambiguous slots; inter-annotator agreement and the advantages and drawbacks of the annotation strategy.

3.1. Guidelines for annotation

The annotation process is split in two parts: semantic role assignment and assignment of Universal Dependencies Relations. Semantic role assignment in NoVRol is based on the annotation guidelines for the English Verbnet. In addition to annotating the verbs based on the guidelines, Norwegian verbs were compared with English translations and the semantics of their assigned VN classes to verify semantic similarity. In cases of inter-annotator disagreement, English VN classes were consulted for semantic properties to disam-

[4] This system for semantic annotation is extensively used in a resource for the West African language Ga, described in (Hellan, 2023) along with situation type labels. An issue for the annotation in that project was that many labels that had been used in similar applications for English were not adequate for Ga. We have not encountered similar issues in the present context, but, as a reviewer points out, this is an essential concern to keep in mind when classification systems in this area are borrowed from one language to another.

[5] https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf

lexvals	roles	UD
ane intr	experiencer	`nsubj`
ane tr	experiencer–stimulus	`nsubj--obj`
ane tr	stimulus–experiencer	`nsubj--obj`
ane tr-obDECL	experiencer–stimulus	`nsubj--ccomp`
ane tr-obINTERR	experiencer–stimulus	`nsubj--ccomp`
ane tr-suDECL	stimulus–experiencer	`csubj--obj`
ane tr-suINTERR	stimulus–experiencer	`csubj--obj`
ane trExpnSu-expnDECL	formal–experiencer–stimulus	`expl--obj-csubj`
ane trExpnSu-expnINTERRwh	formal–experiencer–stimulus	`expl--obj-csubj`

Table 2: Valpod for *ane* annotated for semantic roles and UD frames

biguate semantic role assignment. For example, the annotation of *hånflire* 'smirk', was annotated respectively as <agent, patient> (following the English class `bully-59.5`) and <agent, stimulus> (following `nonverbal_expression-40.2`). Only the latter class allows an interpretation where the verb is a reaction to a stimulus, which aligns with the usage of *hånflire*. The annotation <agent, stimulus> was chosen.

3.2. Annotation differences between Norwegian and English

Certain aspects of the English VerbNet do not straightforwardly align with Norwegian, or contain certain inconsistencies that this project dealt with. This section discusses three such examples.

Reflexives The annotation in NorVal pertains to syntactic properties exclusively, and not the possible status of *seg* as a semantic argument, i.e., a role-bearer; thus, *seg* in *skamme seg* 'be ashamed' is counted as an object on syntactic grounds, but would by most linguists be regarded as semantically empty. These are annotated as *null-role* in NoVRol.

One standard criterion for deciding the status as role-bearing vs. empty is substitutivity, i.e., whether another expression could be used in the place of *seg*. For example, *seg* in *skamme seg* cannot be replaced by another NP.

Another, less clear-cut, criterion is whether the situation type expressed by the construction 'feels' as expressing a participant corresponding to the position of *seg*. For *skamme seg*, this criterion matches the criterion of substitutivity. In contrast, the situation expressed in *Jon vasker seg* 'Jon washes himself' might be perceived as having just a single participant performing some activity, and thus implying no extra role status corresponding to *seg*; however, the object position is here fully substitutable by other NPs. In such cases the annotator will follow his or her intuition as to whether to assign a role or not to the reflexive.

In the English VerbNet, where the presence of light reflexives is far less prominent than in Norwegian, the annotation of reflexives in some cases makes use of the predicative relation `equals`. This relation is used in some <agent, patient> verbs, for example *dress oneself* (`dress-41.1.1`), to indicate that multiple arguments have the same referent. The predicate is absent from <agent, benefactive> verbs, e.g., *cook oneself a meal* (`preparing-26.3`), where the role annotation is the same as in the NoVRol.

Different role names The English VerbNet includes the roles *causer*, *circumstance*, *eventuality* and *subeventuality*. These roles are used in the database, but not mentioned in the documentation. *causer* has been annotated as having the possibility of being both cause and agent. *circumstance* is annoted as source. *eventuality* is annotated as theme, and *subeventuality* as co-theme. Subject expletives are given a *formal* role whereas they are just ignored in the English VerbNet. As mentioned above, light reflexives are annotated as *null-role*. These dummy roles facilitate the matching to UD syntax.

Directionals The English VerbNet contains multiple syntactico-semantic frames for structures that include directionals, whose adjunct/argument status is not clear in the literature (see for example Needham and Toivonen 2011 for discussion). One example is *pour* where the frame `pour-9.5` gives the following example: 'Maria poured water from the bowl into the cup'. In this example, there are two directionals introduced by prepositions. *The bowl* is annotated as *initial location* and *the cup* as *destination*. In NoVRol, we annotate such directionals with a lower degree of precision than other arguments, namely by the role tag *orientation*. The reason for this is that the exact role of such PPs largely depend on the semantics of the preposition itself, rather than that of the verb. Similar considerations led the Groningen Meaning Bank (Bos, 2013) to annotate the semantic role on the preposition itself.

Also, most verbs that can take directionals can take destination, source and path specifications, or any combinations thereof, yielding six different frames. Because directional adverbials are often interchangeable and combinable, this annotation shortcut is a more efficient way to preserve the information. In an SRL system, this information could then be used in combination with a lexicon of preposition senses to derive the actual semantic role in context. Moreover, the NorVal lexvals *suDir*, *obDir* and *PresntDir* tell us whether the direction specified is that of the subject, the object or the logical subject in a presentation construction, enabling a detailed semantic representation of the event structure. The task will remain challenging, however, as there are many ambiguous cases. For example, the verb *hoie* 'scream/yell' contains the lexval *intr-suDir*, which maps to the semantic tag *orientation*, which is ambiguous between different directionals, which could be realized by the preposition *etter* 'after, (here) at'. However, *hoie* also has an entry as a phrasal verb with the preposition *etter* 'scream/yell for', in which case the object of the preposition is invariably understood as a *topic*. Therefore only contextual knowledge can disambiguate examples like (4).

(4) De hoiet etter en lege
they screamed after a doctor
'They screamed at/for a doctor'

However, when the verb does not have a non-directional frame with a preposition that can introduce a direction, the *orientation* role makes it possible to retrieve the semantics of directionals.

3.3. Inter-annotator agreement

To evaluate the annotation quality, we set aside a test set of 800 lexvals, roughly 5% of the total lexval database size. These verbs were annotated by both annotators without discussion between them. All instances where the semantic frames differed in at least one semantic role were counted as disagreement. This could happen if the two annotators had assigned a different role to one of the arguments, irrespective of the number of semantic roles they agreed on. Another frequent error source are ambiguous verbs where the annotators had annotated two different frames, which were eventually both regarded as correct. Our metric is therefore relatively harsh, and the inter-annotator agreement rate was 0.58 measured using Cohen's kappa, which is relatively low. We nevertheless think the annotation is of high quality, as a closer analysis of the annotation mismatches reveals.

Of the 339 annotation mismatches in the test set, 41% of all mismatches were associated with verbs with multiple senses. Annotators had assigned different semantic rolesets, but the assigned rolesets were all valid. The verb *senke*, for example, may mean both 'sink' (<agent, patient> following the English class `other_cos-45.4`) and 'lower' (<agent, theme> – `put_direction-9.4`). Similarly, the verb *overtrekke* may mean both 'withdraw too much' and 'coat', fitting both `funnel-9.3` and `spray-9.7`.

In the remaining cases, different verb sense interpretations could not account for annotation mismatches. 55% of the remaining mismatches were yet categorized within the same macro-roles outlined in the VerbNet guidelines[6]. For example, the complement of the verb *overutstyre* 'overequip' was annotated respectively as *destination* and *recipient*, both members of the macro role *place*.

We conclude that most of the errors involve either annotators missing out on frames that should be present, in which case they can be added later, or they disagree on the exact role but agree on the macro-role, which means that even the wrong annotation is not too far off.

3.4. Annotating UD syntax

General strategy In addition to the semantic annotation, the verbs in the dataset were annotated for syntactic relations based on the UD scheme. This annotation was done for the 340 distinct valency frames in NorVal. Whenever possible, the annotation in the Norwegian UD treebank was consulted. Although most verbs in NorVal are not represented in the treebank, it was possible to find at least one verb from a particular frame most of the time. In doing this, we only paid attention to the syntactic labels assigned to the (heads of the) arguments. So for example, both interrogative and declarative complement clauses get the label `ccomp` in UD, and therefore the two NorVal frames *trExpnSu-expnDECL* and *trExpnSu-expnINTERRwh* get mapped to the single UD frame `expl--obj-csubj`. Similarly, UD does not distinguish subject and object control infinitives, while these are distinguished in the NorVal frames. As a result, the 340 NorVal frames are reduced to 64 UD frames, which therefore contain less information. However, while this is a lossy many-to-one mapping, the NoVRol does contain information about what NorVal frame the UD frame came from, making it possible at a later stage to extract more information and enrich the UD frames.

The UD frames of verbs are ordered by a hierarchy loosely following the Norwegian word order, as in (5).[7]

(5) `subj` ≺ `iobj` ≺ `obj` ≺ `advmod` ≺ `obl` ≺ `xcomp/advcl` ≺ `ccomp`

`subj` is not a UD relation, but a cover term for `nsubj`, `csubj` and `expl`, which in Norwegian is generally subject expletives. One exception to the

[6] https://verbs.colorado.edu/verb-index/VerbNet_Guidelines.pdf, p. 18

[7] See below for why some apparent adjunct functions are included in the valency.

above hierarchy happens when expletives cooccur with a displaced subject, which is called a 'logical subject' in traditional Norwegian grammar and is labelled c/nsubj in UD, although it occurs in object position (6).

(6) Det vil tilflyte oss penger
expl will flow us money
'There will flow money to us'

Such cases get the UD frame expl—obj—nsubj. Finally, the syntactic annotation was aligned to the semantics by arranging syntactic functions and semantic roles in the same order so that the mapping from function to role is transparent.

Adjuncts and obligatory arguments It is a common pattern for infinitival clauses in Norwegian to be introduced by prepositions, as in (7).

(7) Han ba dem om å gå
he asked them about to go
'He asked them to go'

Such infinitival clauses are treated as adverbial clauses (advcl) in the Norwegian UD treebank. This label suggests that they do not belong to verb's valency frame at all, but are adjuncts. This is clearly not the case, however. A related problem arises with nominal arguments, since UD does not distinguish arguments and adjuncts, but lump non-core (not subject or object) dependents as obliques. These will be given a semantic role in our annotation if and only if they are considered arguments in NorVal and appear in the frames there. This means that when our lexicon is used in conjunction with a UD parse, one cannot know a priori whether an advcl or obl dependent will be assigned a semantic role or not. We see no way around this problem as long as UD does not distinguish arguments and adjuncts, since it is not practicable to list adjunct roles in a verb-based lexicon.

4. Lexical resources for UD

The UD initiative – and dependency treebanks in general – have historically been connected with the success of data-driven dependency parsing, which by its very nature required the annotation of running text rather than lexical resources. Dependents are annotated ”as they occur” and there is no attempt to extract more systematic patterns, unlike grammar-based parsers based on Head-Driven Phrase Structure Grammar (HPSG), Lexical-Functional Grammar (LFG) and Combinatory Categorial Grammar (CCG), which are typically based on rich lexicons. This move vastly improved robustness, but currently the very success of dependency parsing is sparking new interest in dependency grammar as a theory, which from its origins in Tesniere (1959) was always interested phenomena such as valency. We believe the time has therefore come to enrich UD with lexical resources.

Some moves in this direction are already seen within UD itself. For example, the UD validator relies on a list of auxiliary verbs which are actually annotated with a simple semantics, where they are marked as either Copula, Perfect, Past, Future, Passive, Conditional, Necessitative, Potential, Desiderative, Other or Undocumented auxiliaries. High-level information like this may be all that is possible to achieve at a universal level, although one can hope that it can be extended to other functional categories such as determiners, negators and subordinators.

More realistically, though, the creation of lexical resources will happen at a language-specific level and link up to the UD scheme. This is how we see the present contribution. However, rather than extracting information from a UD treebank and systematize and curate it to produce a lexicon, we have taken the information from resources built around the Norwegian HPSG grammar, which has been developed over two decades. Such resources, which have been handcrafted for many languages, but are often tied to specific linguistic formalisms (often LFG, HPSG or CCG) and even specific computational implementations of those formalism, contain a wealth of information that can be useful also in a dependency grammar context if it is made accessible in more theory-neutral forms as free-standing resources, alongside their function inside more closed systems such as computational grammmars. In particular, such handcrafted lexical resources contain a lot of information about the long tail of rare items: as stated above, NorVal contains ca. 7,400 verbs. By comparison, the first 10M tokens of the NoWaC corpus[8] contains 5,465 distinct verbs, the first 100M contains 6,929, and only the full corpus of 687M tokens surpasses NorVal and has 7,706 verbs.

5. Dataset statistics

The annotated verb set yields a database where syntactic features are given semantic tags. This section outlines the characteristics of the verb classes, their size, content and relations to syntax.

5.1. Number of classes

In our annotation, each lexval has been associated with a set of semantic roles, one role for each of the semantic arguments expressed in the frame. Such a set we may refer to as a *roleSet*; for each lexval, we may refer to its roleSet as a *lexvalRoleset*, and a roleSet abstracted away from its lexvalRoleset may be called a *roleSetType*. Across all the annotated lexvals, 250 roleSetTypes are used, and

[8] Norwegian Web as Corpus, Guevara (2010)

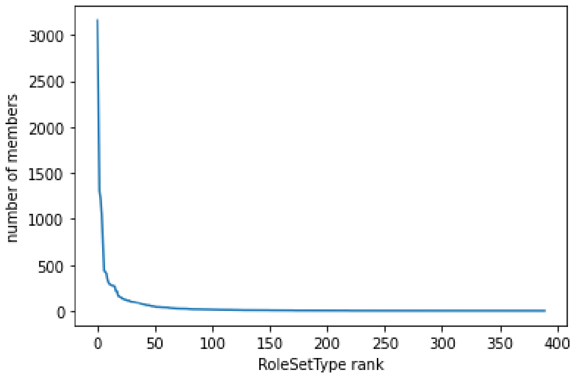

Figure 1: RoleSetType rank by members

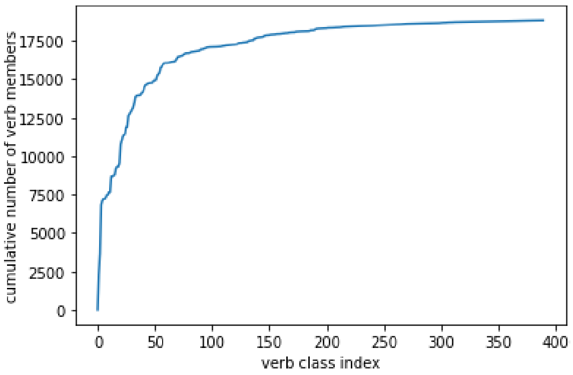

Figure 2: Cumulative members of by verb class index

we may define the notion *classes of lexvals* according to which roleSetTypes are aligned with the lexvals. Semantic role order is preserved – verbs annotated for the same semantic roles in different order, e.g., *fear* and *scare*, are members of different roleSetTypes. Derivatively we may speak of *classes of verbs* according to the *verb lexemes* represented in these classes of lexvals.

We name such classes of lexvals or verbs after the verb lexeme of the alphabetically first lexval where the roleSetType is found, for instance as in `abonnere`: <*agent*, *theme*> . This way of naming classes resembles a bit what is done for 'verb classes' in VerbNet. But note that in VerbNet 'verb class' is constituted by a combination of semantic and syntactic features, where the semantic features comprise not only roles but also logical form and elements of conceptual semantics, whereas our classes are defined by roles alone, hence the name roleSetTypes.

The number of members in each roleSetType by their rank is shown in Figure 1 and shows a Zipfian distribution. The cumulative distribution is shown in Figure 2. The three most common, `abonnere` ('subscribe'): <*agent*, *theme*>, `abbreviere` ('abbreviate'): <*agent*, *patient*> and `abdisere` ('abdicate'): <*agent*>, occur in in respectively 3,163, 2,344 and 1,307 lexvals. The first of these classes can broadly be described as representing agentive, bivalent verbs whose second argument does not undergo a change of state, as in (8).

(8) de hamstrer matvarer
they hoard foodstuffs
'they hoard foodstuffs'

The second most common roleSetType represents agentive bivalent verbs whose second arguments are internally changed – the referent of the object of the verb *abbreviere* ('abbreviate') is made shorter. The third class is the class of agentive intransitives, e.g., *abdisere* ('abdicate').

On the tail end of the frequency list, there are 12 roleSetTypes with 3 members each, 31 with 2 and 70 with 1. The reason for the large number of roleSetTypes with one member is found in the source syntactic annotation. The roleSetTypes `trives` <*experiencer*, *location*>, for example, represents one verb: *trives* 'thrive', annotated for location (9).

(9) deltagerne trives her
participants.def thrive here
'the participants are thriving here'

The number of rare roleSetTypes follows from the NorVal tagging, which for *trives* is *intrObl-oblLoc*: an intransitive verb that selects for an oblique locative. Of 30 verbs with this syntactic tag in NorVal, *trives* is the only one that takes an experiencer subject. Note crucially that the verb *trives*, without a locative, is also a member of the larger roleSetType `ane` <experiencer>, with 94 members, among them *lide* 'suffer' and *koble av* 'relax'. The large number of classes, then, needs to be seen in relation with the syntactic tagging of arguments given in NorVal.

5.2. Class granularity

As already said, our annotation yields a database where separate verbs are semantically tagged only for semantic roles. This contrast with the English VerbNet, where verbs such as *hold* and *neglect*, although annotated using the same semantic roles, belong to different classes based on semantic definitions: the class `hold-15.1` is defined semantically as *contact*, while `neglect-75.1` is defined as ¬*handle*. Our annotation thereby results in larger classes – verbs that would belong to different classes in a semantically richer classification, end up in the same class. Verbs like *antenne* 'ignite' and *vie* 'marry', for example, both end up in `abbreviere` <agent, patient>.

For the purpose of using the database as a lexical resource for UD graphs, the low semantic granularity is not an issue. The current stage of

semantic role	freq.	semantic role	freq.
agent	10,691	*topic*	946
theme	6,792	*recipient*	600
patient	3,376	*destination*	570
null-role	1,667	*orientation*	521
experiencer	1,226	*pivot*	451
stimulus	1,134	*formal*	399

Table 3: The 10 most frequent semantic roles

the database, however, is a suitable point of departure for adding more detailed semantic definitions, as is done in the English VerbNet, and for specifying valid syntactic alternations for different verb frames. *vie*, for example, may be followed by the segment <*til* co-patient> 'marry x to y'. This is not possible for *antenne*. As a syntactico-semantic resource, syntactic subcategorization of the semantic classes stands out as a central future endeavour for creating a full Norwegian VerbNet. However, the current stage of the database provides ample opportunities for examining syntactico-semantic phenomena. Some key statistics and possible usage domains are given below.

The distribution of the 10 most frequently annotated semantic roles is given in Table 3. *null-role* is annotated for reflexive pronouns lacking semantic participant status. The role *formal* represents syntactically required but semantically vacuous pronouns, as found for instance with weather-related verbs (Bolinger, 1973).

In total, 31,351 semantic role tokens were annotated for 18,830 sense-distinct lexvals (i.e., among the 17,200 lexvals in NorVal, the syntactic frame in many cases hosts more than one sense in terms of semantic roles, bringing the number of role-annotated lexvals up to 18,830). On average, each lexval frame contains 1.7 semantic roles (1.9 if null-roles and formal subjects are not counted). The database allows for queries about the co-occurrence of semantic roles in Norwegian: out of a total of 6,792 instances of the role *theme*, 141 are followed by a *co-theme*, tentatively illustrating the structural frequency of themes co-occuring with an equally salient undergoer. Out of a total of 1,342 instances of the role *experiencer*, 1,020 co-occur in structures with a stimulus, 666 of which precede the experiencer role (10) and 354 of which surface after the experiencer (11).

(10) vi$_{stim}$ avskrekker villsvinene$_{exp}$
we scare.off boars.def
'we scare off the boars'

(11) vi$_{exp}$ frykter [at huset bygges]$_{stim}$
we fear that house build.pass
'we fear the building of the house'

The database further has the potential to be used for research in lexical semantics, for example for the question of what kind of verbs combine with formal subjects in Norwegian compared to other Germanic languages. A query that looks for formal subjects *formal* followed a *results* role yields a semantic structure in Norwegian (12) that is not found for English in the English Verbnet.

(12) det$_{formal}$ slår om til [å regne]$_{result}$
it changes to rain
'it is (the weather) changing to rain'

5.3. Semantic roles and UD

As described in section 3.4, the NorVal frames were mapped to UD frames, and the semantic roles were aligned with UD functions as was shown in Table 2. In general, the mapping from VerbNet to UD is many-to-one – different semantic functions maps to a single syntactic annotation. For example, both benefactive objects (*he defended **them***) and objects of verbs of breaking (*she destroyed **the vase***) reduce to a single UD relation *obj*.

However, we also find – albeit to a lesser extent – one-to-many mappings from semantics to syntax. This is because semantic rolesets that are annotated for the same role are distinguished into multiple syntactic frames based on whether the semantic role is represented by a clausal or nominal element. The two semantically identical objects in (13) are assigned different syntactic relations in UD, respectively *obj* and *ccomp*.

(13) I accepted {it / that they wrote novels}

The 250 semantic classes (i.e., roleSetTypes) map to 63 UD configurations at the syntactic level. The most common UD configuration is *nsubj-obj* – structures with a nominal subject and object – with 7,226 roleSet tokens. The second most common is the class of argument structures with a single nominal argument – *nsubj* – with 2,602 member frames.

The mapping from semantic frames to syntactic structures is an overall reductive process. Looking at single frames, however, these often increase. Both of the semantic frames <*agent theme*> and <*experiencer stimulus*>, when following the syntactic conventions in UD, map to five syntactic frames: *nsubj-advcl*, *nsubj-ccomp*, *nsubj-obj*, *nsubj-obl* and *nsubj-xcomp*. The verb *frykte* 'fear' selects for three of the syntactic structures (14), while the phrasal verb *fortvile over* 'despair about' showcases the remaining two (15).

(14) vi$_{nsubj}$ frykter {dem$_{obj}$ / [at huset
we fear them that house.def
bygges]$_{ccomp}$ / [å tape]$_{xcomp}$}
build.pass to lose

'we fear them / that the house is built / to lose'

(15) han$_{nsubj}$ fortviler over {[vår skjebne]$_{obl}$
he despairs about our destiny
/ [hva som må gjøre]$_{advcl}$}
what that must done.pass
'he despairs about our destiny / what must be done'

6. Conclusion/Outlook

We have presented NoVRol, a semantic role lexicon of Norwegian verbs. We started from the NorVal valency lexicon and identified the semantic roles that the verbs in this database assign to their arguments, based on the VerbNet annotation guidelines. In the next step, we encoded the verbs' valency frames in UD, allowing for an easy mapping from UD functions to semantic roles that could be used, e.g., in semantic role labeling of running text.

Going beyond the current annotation, we believe there are also several exiciting avenues for further development of the resource. Integrating the detailed syntactic information from the NorVal frames, ideally in the same format as in the English VerbNet, would enable the creation of a much more detailed verb class system. This would make cross-linguistic studies of argument structure easier given the common annotation framework. Such a resource could in turn enable more research into regularities in the syntax-semantics mapping. Moreover, it would then also be possible to create detailed semantic representations of event structure. This could be exploited in semantic parsing, which was indeed the motivating application for our work.

Data availability

The dataset is available at `https://github.com/Universal-NLU/NoVRol` under the CC BY-SA 4.0 license.

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Dorothee Beermann and Lars Hellan. 2004. A treatment of directionals in two implemented hpsg grammars. In *Proceedings of the HPSG04 Conference*. CSLI Stanford.

Dwight Bolinger. 1973. Ambient it is meaningful too. *Journal of Linguistics*, 9(2):261–270.

Johan Bos. 2013. The Groningen meaning bank. In *Proceedings of the Joint Symposium on Semantic Processing. Textual Inference and Structures in Corpora*, page 2, Trento, Italy.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856.

Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications.

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

Charles J Fillmore. 1968. *The Case for Case*, pages 1–88. Holt, Rinehart, and Winston, New York.

Jamie Y. Findlay, Saeedeh Salimifar, Ahmet Yıldırım, and Dag T. T. Haug. 2023. Rule-based semantic interpretation for Universal Dependencies. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 47–57, Washington, D.C. Association for Computational Linguistics.

Emiliano Raul Guevara. 2010. NoWaC: a large web-based corpus for Norwegian. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop*, pages 1–7, NAACL-HLT, Los Angeles. Association for Computational Linguistics.

Lars Hellan. 2019. Construction-based compositional grammar. *Journal of Logic Language and Information*, 28:101–130.

Lars Hellan. 2022. A valence catalogue for norwegian. In *Natural Language Processing in Artificial Intelligence*, pages 49–104, Cham. Springer International Publishing.

Lars Hellan. 2023. A unified cluster of valence resources. In *Logic and Algorithms in Computational Linguistics 2021*, pages 311–347, Cham. Springer International Publishing.

Lars Hellan and Tore Bruland. 2015. A cluster of applications around a deep grammar. In *Proceedings from The Language Technology Conference, LTC2015, Poznan*, pages 503–508.

Jena D. Hwang and Martha Palmer. 2015. Identification of caused motion construction. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 51–60, Denver, Colorado. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Olga Majewska and Anna Korhonen. 2023. Verb classification across languages. *Annual Review of Linguistics*, 9(1):313–333.

Jaouad Mousser. 2010. A large coverage verb taxonomy for Arabic. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Stephanie Needham and Ida Toivonen. 2011. Derived arguments. In *Proceedings of the LFG11 Conference*, pages 401–421. CSLI Stanford.

Wessel Poelman, Rik van Noord, and Johan Bos. 2022. Transparent semantic parsing with Universal Dependencies using graph transformations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4186–4192, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Quentin Pradet, Laurence Danlos, and Gaël de Chalendar. 2014. Adapting verbnet to french using existing resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. Universal semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. 2021. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

Lucien Tesniere. 1959. *Éléments de syntaxe structurale*. Klincksieck, Paris.