Towards a Dutch Parseme corpus

Gosse Bouma
CLCG
University of Groningen
g.bouma@rug.nl

Jan Odijk
ILS
Utrecht University
j.odijk@uu.nl

Carole Tiberius
Dutch Language Institute,
Leiden
carole.tiberius@ivdnt.org

Relevant UniDive working groups: WG1

1 Introduction

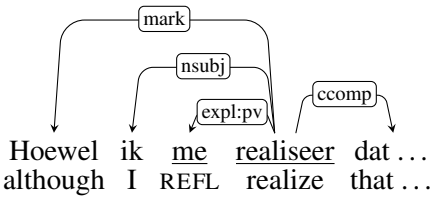
This abstract describes our work (in progress) on creating a Dutch corpus of verbal multiword expressions following the guidelines of the PARSEME[1] project. Savary et al. (2023) discuss the relationship between the syntactic dependency annotation of the Universal Dependencies initiative (De Marneffe et al., 2021) and annotation in PARSEME. Existing PARSEME corpora are extensions of UD annotation, where an additional layer of annotation is added to the morpho-syntactic annotations provided by UD. We follow this line of research by implementing conversion rules that, given a Dutch sentence annotated with UD attributes, adds the relevant PARSEME annotation. For cases where the UD annotation is insufficient to identify verbal MWEs, we consult a lexicon with detailed information on idiomatic expressions.

The output, while generally accurate, needs to be manually verified to ensure that (1) all words that are part of the MWE are included in the annotation, (2) the correct MWE (sub-)class has been assigned, and (3) no relevant expressions are missed.

2 Conversion from UD

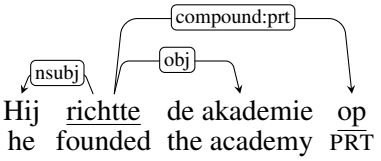
We use the Grew-match search engine to explore to what extent the various PARSEME classes can be identified in the Dutch UD (Version 2.12) corpora. The actual conversion rules are implemented in a Jupyter notebook[2], using the Python PARSEME library[3] for reading in the UD data and outputting the enriched data in `cupt` format.

The IRV class is used to label verbs that occur with an inherent (non-referential) reflexive (e.g. ***zich vergissen*** 'to be mistaken', ***zich realiseren***, 'to realise'). In UD, the reflexive pronoun dependents of such verbs are labeled with the dependency relation `expl:pv`:



Thus, instances can be identified reliably from the UD annotation.[4]

The VPC class labels verbs with a particle dependent (e.g. ***op richten***, 'to found'), a frequent phenomenon in Germanic languages. The UD dependency relation `compound:prt` identifies such particles:



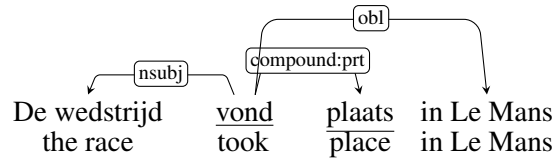
We restrict the VPC class to `compound:prt` dependents that are single tokens with POS ADP, ADJ or ADV[5] as in the Dutch corpora `compound:prt` is also used to identify a wider range of expressions that form an idiomatic expression with a verb (as explained below). Verb-particle combinations can also be written as a single word (e.g. ***afkorten*** 'to abbreviate', lemma `af_korten`). Such occurrences can be identified by inspecting the lemma of the verb, which is generally of the form `prefix_stem`, i.e. the presence of an underscore in the lemma indicates that this is a particle verb.

PARSEME distinguishes between VPC.full and VPC.semi, VPCs where the meaning of the constructions is fully non-compositional or semi compositional, respectively. This is a semantic distinction and manual checking is required to label VPCs with the correct subclass. As approximately 75% of the VPCs in the English and German PARSEME corpora are VPC.full, we assign VPC.full as default.

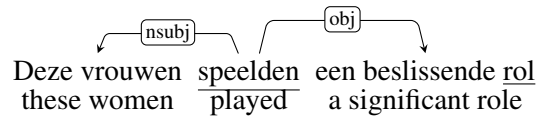
Light verb constructions (LVC) consist of a semantically light verb and a predicative noun (e.g.

[1] https://parsemefr.lis-lab.fr/doku.php
[2] https://github.com/gossebouma/Parseme-NL
[3] https://gitlab.com/parseme/cuptlib
[4] https://universal.grew.fr/?custom=6540f79d081ba
[5] https://universal.grew.fr/?custom=6543b87bf2cd2

zitting hebben 'to be a member', ***plaats vinden***, 'to take place'). In Dutch UD, a subset of these can be identified by searching for nouns that are `compound:prt` dependents of a verb and that do not have dependents of their own:[6]

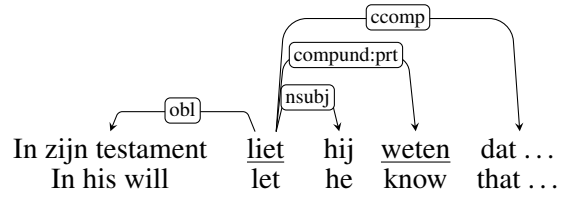


The PARSEME guidelines do not restrict LVC to a small, closed, set of verbal stems, and indeed, the EN, FR and DE corpora contain LVCs with 30-90 different stems, and thus we do impose no additional constraints on the verbal stem. Another group consists of verbs that occur with a predicative noun where the dependency label is `obj`, e.g. ***een rol spelen***, 'play a role':



As `obj` is used for direct objects of a verb in general, the label by itself is not informative. The Alpino-lexicon[7] (van Noord, 2006) contains over 300 entries with detailed information on verbs that occur with specific object nouns (often because the construction is somewhat syntactically irregular as well). We check this lexicon to annotate a subset of verb-object occurrences as LVC as well. The division into the semantic subclasses LVC.full and LVC.cause cannot be made automatically, and thus requires manual checking.

Multi-verb constructions (MVC) are idiomatic combinations of two verbs, such as ***laten vallen*** lit. 'to let fall' and ***laten weten***, 'to state, to declare':



In Dutch UD, they can be identified by searching for verbs with a verbal `compound:prt` dependent.[8]

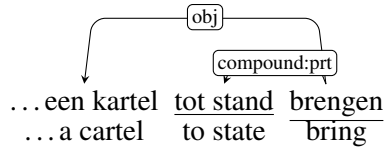
[6] https://universal.grew.fr/?custom=65410400cd225

[7] https://www.let.rug.nl/vannoord/alp/Alpino/

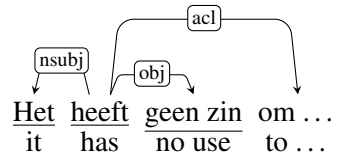
[8] https://universal.grew.fr/?custom=6541429f56a3d

Verbal idiomatic expressions (VID) are a diverse set of remaining cases. We can distinguish between three cases: (a) one where a `compound:prt` is a complex phrase itself[9] (i.e. it has `fixed` dependents or has other internal structure as in ***tot stand brengen***, 'to create') (b) idiomatic expressions involving a subject, (non-predicative) object, or oblique dependent of a verb that is listed with this pattern in the Alpino-lexicon, e.g. ***het heeft geen zin***, 'it is useless', and (c) cases where two or more phenomena co-occur. The latter happens for instance when a verb has both a verbal particle and a predicative noun as dependents. In that case we assign the class VID instead of assigning two overlapping classes LVC and VPC. Other interactions occur when a multi-verb expression involves a verb with a particle or when a verb with a particle also has an inherent reflexive as dependent.

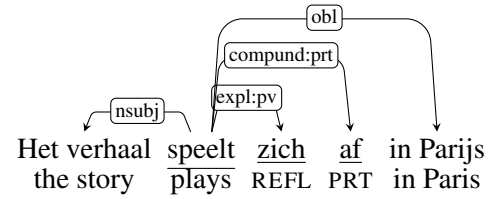
a.



b.



c.



3 Results

Table 1 gives an overview of the distribution of MWE classes after automatic conversion. The distribution is rather skewed, with the majority class consisting of verb-particle (VPC) constructions. Ignoring the subclass label, the VPC as well as the IRV cases can be identified rather easily and accurately in the UD annotation, and we assume that these need hardly manual verification.

A PARSEME class that has not been included in the conversion are the inherently adpositional verbs (IAV). This class is described as optional

[9] https://universal.grew.fr/?custom=654141a700068

and experimental in the guidelines[10]. In the Dutch UD corpora, all adpositional dependents of a verb are labeled `obl` or `compount:prt`. It should be noted, however, that the treebanks from with the Dutch UD corpora are derived, Alpino[11] and LassySmall[12], distinguish between `pc` and `mod` prepositional phrases modifying a verb, where the first correspond to the IAV cases. Thus, if a future UD release decides to preserve this distinction[13] it should be relatively straightforward to incorporate this class as well.

Class	Alpino	LassySmall
VPC.full	2088	670
VID	616	145
LVC.full	571	139
IRV	332	125
MVC	65	6
Total	3672	1085

Table 1: Statistics for various MWE classes after automatic annotation of the Dutch UD corpora

4 Future Work

We plan to use the FLAT annotation tool for manual verification of the automatically converted data. We will develop language specific guidelines for the distinction between VPC.full and VPC.semi as well as LVC.full and LVC.cause. VIDs also require special attention, as the question what lexical items to include cannot always be decided automatically, and also, many VIDs are a mix of two phenomena, such as particle-verbs with an inherent reflexive. As these are relatively frequent in the Dutch data, it might be worth reconsidering the policy of annotating these as a single construction instead of as an overlapping combination of two. After manual verification, the corpus can be used among others to assess the performance of the MWE-finder and the associated list of MWEs contained in DUCAME (Odijk et al., to appear).

[10] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/

[11] https://www.let.rug.nl/~vannoord/trees/

[12] https://taalmaterialen.ivdnt.org/download/lassy-klein-corpus7/

[13] i.e. using https://universaldependencies.org/u/dep/obl-arg.html

References

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

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. to appear. MWE-finder: Querying for multiword expressions in large Dutch text corpora. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources. Linguistic, Lexicographic and Computational perspectives*, Phraseology and Multiword Expressions. Language Science Press, Berlin.

Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions. *Northern European Journal of Language Technology*, 9(1).

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.