Canonical Form for Light Verb Constructions

Jan Odijk
Utrecht University / Utrecht, the Netherlands
j.odijk@uu.nl

Relevant UniDive working groups: WG2

1 Introduction

We describe the canonical form (in the sense of (Odijk et al., 2024b; Odijk and Kroon, 2024)) for light verb constructions as they are being developed in UniDive WG2 Task 2.4. We tentatively propose a classification of light verbs. This classification makes it possible to have multiple light verbs, each with its own meaning, with the same complement. It also plays a role in regulating which verbs can combine with a predicative complement. The classification has been applied to a large Dutch MWE lexicon (Odijk, 2023) and is currently being applied to the examples of the Parseme guidelines[1] for multiple languages.[2]

2 Light verb constructions

We start with the description of light verb constructions from the Parseme guidelines,[3] which we summarize here. Light verb constructions are formed by a verb and a complement. The complement refers to an event or a state and takes arguments. The verb v is "light" in that it contributes to the meaning of the whole only by bearing morphological features.

We adopt this description, though we suggest that the semantic contribution of the verb is sometimes more than just morphological features. For example, the Dutch examples noun *poging* 'attempt' can combine with three different light verbs, each having a clearly different meaning, i.e. *doen* 'do' ('make an attempt'), *wagen* 'dare' ('make a risky attempt') and *ondernemen* 'undertake' ('make a large-scale attempt').

[1] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=home

[2] https://drive.google.com/file/d/1D7b_FJ7774qs3xOKCMKyE2y_Xfv9wtCw/view?usp=drive_link

[3] https://parsemefr.lis-lab.fr/parseme-st-guidelines/2.0/?page=050_Tests_for_VERBAL_MWEs/020_Light-verb_constructions__LB_LVC_RB_

In addition, a characteristic of light verb constructions is that the arguments of the complement can be and often are realized as complements of the verb.

The Parseme guidelines are silent on how to describe that some choices of light verb are ill-formed. For example, in Dutch one must say *een poging doen* ('do an attempt') but not **een poging maken* ('make an attempt'), but in English it is the opposite. How can we describe that *maken* is wrong here? We cannot forbid just any verb other than *doen*, since *wagen* and *ondernemen* are allowed, as are verbs that are no light verbs at all such as *becommentariëren* 'comment' , *beëindigen* 'end', *bekijken* 'watch', *beschrijven* 'describe', *evalueren* 'evaluate', *frustreren* 'frustrate', *onderzoeken* 'investigate', *verijdelen* 'thwart', and undoubtedly many others.

3 Light verb classes

We propose that there are a number of classes of light verbs. A phrase can form a light verb construction with a specific verb and the phrase as the verb's complement if this is explicitly listed in the lexicon, and if the verb occurs in the class associated with the phrase's semantic type. If a verb from this class is not listed as an option in the lexicon for the phrase, then it is ill-formed as a light verb for the phrase. Verbs that are not in the class associated with the phrase's semantic type can combine freely with the complement provided that semantic selection restrictions allow this: a phrase headed by the word *poging* 'attempt' is of semantic type activity. A verb to form a light verb construction must then be selected from the DO-class of light verbs. This DO-class contains at least the verbs *doen*, *maken*, *wagen*, and *ondernemen*. We can list the light verb constructions that can be formed with *poging* in their canonical forms in the lexicon as follows:

- iemand 0zal 0een L:poging DO:doen
- iemand 0zal 0een L:poging DO:ondernemen

- iemand 0zal 0een L:poging DO:wagen

The word *poging* is marked here with *L:* as having a meaning that it can also have outside this MWE. The complement in a LVC usually has a meaning that it also has outside of the LVC. The article *een* is present because *poging* is a count noun and requires a determiner to get a well-formed expression, but the determiner is not a component of the MWE. For this reason *een* is preceded by *0*. The verbs are marked with *DO:* to indicate that they form a light verb construction with the complement, which is semantically an activity. The verb *maken* is part of the DO-class of light verbs but not listed in the lexicon and, under the assumption that the lexicon is complete, therefore ill-formed as a light verb for *poging*. Verbs that are not in the DO-class, such as the ones listed above can combine with a phrase headed by *poging* as its direct object if semantic selection restrictions allow this. This canonical form enables the automatic generation of a query to search for instantiations of the MWE in treebanks, and allows for free modification of *poging*, including relativisation (as in 1), as shown by (Odijk et al., 2024a; Odijk and Bouma, 2025):

(1) De poging die hij gedaan had was mislukt
the attempt that he done had was failed
'The attempt he made failed'

A light verb imposes semantic and syntactic restrictions on its complement. The semantic type of this complement can be an activity, specific subcases of activities, properties, or states. Syntactically, the complement can be a direct object or a complementive (predicative or locational-directional complement). Examples:

activity/NP *een poging doen* (an attempt do) 'make an attempt';

notallowedactivity/NP *een misdaad plegen* (a crime commit) 'commit a crime', see (Odijk, 2023), Appendix D;

presentationactivity/NP *een lezing geven* (a lecture give) 'give a lecture'

bodily touch activity/NP *iemand een kus geven* (someone a kiss give) 'give someone a kiss', see (Odijk, 2023), Appendix A;

state/NP e.g. *honger hebben* (hunger have) 'be hungry'. See (Odijk, 2023), Appendix C

state/PP e.g., *in de war zijn* (in the WAR be) 'be confused', in de war zitten (in the WAR sit) 'be entangled'. See (Odijk, 2023), Appendix B.

We tentatively distinguish the light verb classes listed in Table 1, with example verbs from Dutch (and the English translations are typical examples for English). Additionally, we mark the complement of bodily touch activity LVCs (e.g., *kus* 'kiss' in *iemand een kus geven* (someone a kiss give) 'give someone a kiss' with the code *LBT*. Many more classes are distinguished than the Parseme/UniDive classes for LVCs, especially to get a clearer picture of LVCs, but the Parseme/UniDive classes (*LVC.Full* and *LVC.Cause*) can be derived from them automatically.

4 DUCAME enriched with LVC annotations

We are in the process of creating a new version of the DUCAME resource.[4] The current version of this resource is being produced and contains a list of canonical forms for (currently) more than 13k Dutch MWEs. This new version will contain annotations for the light verb constructions as described in this paper. Table 2 lists the frequency of each LVC class in the DUCAME resource (as measured on 2026-03-27). We will also present frequencies of the verbs used within each class.

By making these annotations and analysing them we can investigate whether the approach taken is feasible, whether it covers the relevant facts, and whether adaptations or extensions are needed. We will report on our initial findings.

References

Jan Odijk. 2023. A canonical form for Dutch multiword expressions version 2. Unpublished manuscript, Utrecht University.

Jan Odijk and Gosse Bouma. 2025. MWE-Annotator: Automatic identification of MWEs in Dutch. In *Proceedings of the Third General UniDive Meeting*, Budapest, Hungary. UniDive.

Jan Odijk and Martin Kroon. 2024. A canonical form for flexible multiword expressions. In *Proceedings of*

[4] https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

Code	Interpretation	Typical Dutch example
DO	carry out an activity	*doen* (do)
BE	have a property or be in a certain state	*zijn* (be), *hebben* (have)
BC	get a property or get into a state	*worden* (become), *raken* (get)
ST	keep a property or stay in a state	*blijven* (stay), *houden* (keep)
CBE	(causative BE): make someone/something have a property or be in a state.	*hebben* (have)
CBC	(causative BC): make someone/something get a property or get into a state	*krijgen* (get)
CST	(causative ST) make someone/something keep a property or stay in a state	*houden* (keep)
GV	activity directed at a second participant	*geven* (give)
GT	activity originating from a second participant	*krijgen* (get)

Table 1: Table for Light verb classes, with typical Dutch examples. The English translations are also typical examples of these classes for English.

Code	Count
DO:	874
BE:	521
BC:	396
GV:	225
CBC:	175
LBT:	163
GT:	138
CST:	77
CBE:	71
ST:	69
Total	**2709**

Table 2: Counts of occurrences of LVC classes in the DUCAME resource (measured 2026-03-27)

the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 91–101, Torino, Italia. ELRA and ICCL.

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. 2024a. MWE-finder: A demonstration. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12027–12031, Torino, Italia. ELRA and ICCL.

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. 2024b. 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.