MWE-Finder: An evaluation through three case studies

Martin Kroon
Institute for Language Sciences
University of Utrecht, the Netherlands
m.s.kroon@uu.nl

Jan Odijk
Institute for Language Sciences
University of Utrecht, the Netherlands
j.odijk@uu.nl

Abstract

In this paper we showcase and evaluate MWE-Finder, a system that allows users to search for occurrences of an MWE in a large Dutch text corpus. To this end, we conduct three small case studies, and discuss the results in detail. We make use of the MWEs *0geen *+haan zal naar iets kraaien* 'no one will say anything about something', *iemand zal 0dat *+varken wassen* 'someone will deal with that problem' and *iemand zal iemand het hemd van het lijf vragen* 'someone will want to know all the ins and outs of something from someone', which are all in canonical form following Odijk (2023) and Odijk and Kroon (2024).

The results show that MWE-Finder is very accurate in retrieving the target MWEs, reaching an accuracy of 93.7%, and an F_1-score of 95.2%. The case studies additionally lay bare points of improvement of MWE-Finder, specifically concerning the enrichment of syntactic parses by making the object relation explicit in certain constructions.

1 Introduction

Many multiword expressions (MWEs) are flexible in the sense that their components can have different forms, can occur in different orders, or may not be contiguous, with other words appearing between elements of the MWE. This makes searching for such MWEs in large text corpora difficult. To this end, Odijk et al. (2024) developed MWE-Finder for Dutch.

MWE-Finder is a system that allows users to search for occurrences of an MWE in a large Dutch text corpus. It automatically generates three queries based on an input in canonical form. These three queries are increasingly less strict, and allow the user to investigate potential variation easily. MWE-Finder is implemented in the newest version (v5) of GrETEL (Augustinus et al., 2012), and ships with several pre-configured corpora, e.g. SONAR (Oostdijk et al., 2013), parts of Lassy-Large, Lassy-Small (van Noord et al., 2013), the Spoken Dutch Corpus (Oostdijk et al., 2002), and Mediargus.[1] It also offers the user over 11k Dutch MWEs in a canonical form from the DUCAME resource (Odijk, 2023).[2] MWE-Finder is particularly effective, as it takes into account the syntactic structure of the MWEs and the sentences over which it queries using the Alpino parser (van Noord, 2006).

MWE-Finder is intended as a research tool for any linguist or lexicographer interested in research into multiword expressions, in particular *flexible* multiword expressions. It is therefore a natural part of the CLARIN research infrastructure, and it is factually part of it because it is embedded in the CLARIN web application GrETEL. The three queries can lay bare potential variation, leading to the updating of the canonical form in a lexicon for MWEs such as DUCAME, and a more complete and accurate description of the MWE. The system does *identification* (in the sense of (Constant et al., 2017)) of candidate occurrences of MWEs. Though it cannot determine whether an expression is used literally or as an MWE, most occurrences are found to be instances of the MWE, compliant with the observations by other researchers (Savary et al., 2019).

[1] A large treebank with Flemish newspaper text created by Kris Heylen from KU Leuven in 2009.

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

Whereas Odijk et al. (2024) describe how MWE-Finder works in detail and explain certain development choices, this paper acts as a showcase of how MWE-Finder can be used to determine properties of Dutch MWEs on the basis of large corpus data and as an evaluation of MWE-Finder's performance when querying for MWEs, identifying potential points of improvement for the system. To this end, we conduct three small case studies, and discuss the results in detail. We make use of the following Dutch MWEs, which are in canonical form:

(1) 0geen *+haan zal naar iets kraaien
no rooster will to something crow
'no one will say anything about something'

(2) iemand zal 0dat *+varken wassen
someone will that pig wash
'someone will deal with that problem'

(3) iemand zal iemand het hemd van het lijf vragen
someone will someone the shirt from the body ask
'someone will want to know all the ins and outs of something from someone'

The organisation of this paper is as follows. We begin with a brief introduction of the notion multiword expression (Section 2), followed by a dicussion of the notion *canonical form* for an MWE (Section 3). We continue with a brief description of MWE-Finder in Section 4. Section 5 introduces general characteristics of the evaluation performed here. In Subsections 5.1 through 5.3 we discuss the results from the small case studies involving the MWEs listed above. We briefly discuss related work in Section 6 and we conclude in Section 7.

2 Multiword expressions

An MWE is a word combination with linguistic properties that cannot be predicted from the properties of the individual words or the way they have been combined by the rules of grammar (Odijk, 2013).[3] A word combination can, for example, have an unpredictable meaning (*de boeken neerleggen*, lit. 'to put down the books', meaning 'to declare oneself bankrupt'), an unpredictable form (e.g. *ter plaatse* 'on location', with idiosyncratic use of *ter* and *e*-suffix on the noun),[4] or it can have only limited usage (e.g. *met vriendelijke groet* 'kind regards', used as the closing of a letter). In a translation context, it can have an unpredictable translation (*dikke darm*, lit. 'thick intestine', 'large intestine'), etc.

Note that it is not always easy to determine whether a combination of words is an MWE, because we do not always know the exact properties of the individual component words or what the grammar rules of a language are exactly. So this may require a substantial amount of research.

Words of an MWE need not always be fixed. This can be illustrated with the Dutch MWE *de boeken neerleggen* 'to declare oneself bankrupt'. The verb *neerleggen* in (4) can occur in all of its inflectional variants (e.g., past participle in (4a), infinitive in (4b), and past tense singular in (4c) and (4d)), and with the separable particle *neer* attached to it (4a, 4b) or separated (4c, 4d). MWEs do not necessarily consist of words that are adjacent, and the words making up an MWE need not always occur in the same order. This expression allows a canonical order with contiguous elements (as in (4a)), but it also allows other words to intervene between its components (as in (4b)), as well as permutations of its component words (as in (4c)), and combinations of permutations and intervention by other words that are not components of the MWE (as in (4d)):

(4) a. Saab heeft gisteren **de boeken neergelegd**.
Saab has yesterday the books down.laid
'Saab declared itself bankrupt yesterday.'

b. Ik dacht dat Saab gisteren **de boeken** wilde **neerleggen**.
I thought that Saab yesterday the books wanted down.lay
'I thought Saab wanted to declare itself bankrupt yesterday.'

[3] For a similar but slightly different definition, see (Sag et al., 2001).

[4] *Ter plaatse* actually concerns a fossilization of an old dative form, which is no longer productive in Dutch.

c. Saab **legde de boeken neer**.
Saab laid the books down
'Saab declared itself bankrupt.'

d. Saab **legde** gisteren **de boeken neer**.
Saab laid yesterday the books down
'Saab declared itself bankrupt yesterday.'

In addition, certain MWEs allow for (and require) controlled variation in lexical item choice, e.g. in expressions containing bound anaphora such as *zijn geduld verliezen* 'to lose one's temper', where the possessive pronoun varies depending on the subject (cf. *Ik verloor mijn/*jouw geduld; jij verloor *mijn/jouw geduld*, etc.), exactly as the English expression *to lose one's temper*. Of course, not every MWE allows all of these options, and not all permutations of the components of an MWE are well-formed (e.g. one cannot have **Saab heeft neergelegd boeken de.* lit. 'Saab has downlaid books the.').

This flexible nature of such MWEs makes it difficult to reliably search for such expressions in text corpora. Standard search engines such as Google do not enable the user to systematically search for different word forms of the same lemma. Search applications such as OpenSoNaR (de Does et al., 2017; van de Camp et al., 2017) or Nederlab (Brugman et al., 2016) can do this, but it is difficult to formulate a query allowing different orders and interspersed irrelevant words, and the results of such a query will be unreliable. At best, one will find all instances but at the same time also many cases where all the component words occur but do not make up an MWE. One should be able to search for flexible MWEs in such a way that their grammatical structure is taken into account. This can be done in a treebank, and MWE-Finder enables searching for MWEs in a treebank.

3 Canonical Form

A canonical form for an MWE is a unique representation for a set of variants of this MWE that differ only in grammatical properties. A canonical form for MWEs is necessary because many MWEs are flexible, i.e., their component words can occur in different forms, in different orders, or need not always be adjacent. Odijk and Kroon (2024) describe related work on canonical forms for MWEs. The canonical forms assumed here have several unique features: (1) they are defined by very detailed requirements (Odijk, 2023; Odijk & Kroon, 2024); (2) they form well-formed utterances of the language; (3) they can be enriched with annotations; (4) the definition of canonical form has been tested on more than 11k Dutch MWEs in the DUCAME resource; and (5) there are explicit rules on how one can generalise from the canonical form to other forms, which have been implemented in MWE-Finder by a mechanism for the automatic generation of multiple queries for searching for the MWE in large text corpora.

4 MWE-Finder

MWE-Finder enables a user to search for occurrences of an MWE in a treebank based on an example MWE. The example MWE must be in canonical form. The canonical form for MWEs has been defined in (Odijk, 2023) and (Odijk & Kroon, 2024). Canonical forms of MWEs can contain annotations to describe properties of their component words. We will not say anything about this here, but we do provide a list of the most important annotations in Table 1.

MWE-Finder is available in a first version since the end of 2022 as part of the web application GrETEL 5.[5] Thanks to this integration, MWE-Finder has access to all GrETEL features, and supports all treebanks that are included in GrETEL as well as its possibility of uploading one's own text corpora. MWE-Finder partially mimics the structure of GrETEL's main *query-by-example* search functionality. It distinguishes the following steps: *Canonical Form* (cf. GrETEL's *Example* step), *Treebanks*, *Results*, and *Analysis*.

Just like GrETEL, MWE-Finder enables the user to enter an MWE example, though it must be in its canonical form. The user thereby implicitly formulates a hypothesis about the properties of this MWE. The annotations on the example (or their absence) specify how the system should generalise from this example, so these annotations can be seen as a different way of implementing GrETEL's *Matrix*.

[5] https://gretel5.hum.uu.nl

notation	interpretation
**word*	*word* is modifiable/determinable
+word	*word* is inflectable
=word	*word* must occur in the MWE as given
!word	*word* is not modifiable/determinable
dd:[*word*]	*word* must be a definite determiner
<text>	*text* is interpreted as a freely replaceable argument
0word	*word* is not part of the MWE

Table 1: Notational devices for annotating a canonical form. The code + can also be combined with * or *!* (in any order).

The MWEs contained within DUCAME[6] have been included in a drop-down list and are directly searchable within MWE-Finder. The user can also enter a new MWE, provided that it complies with the conventions for MWE canonical forms.

After the MWE has been selected or entered, the system automatically generates three queries to search for occurrences of this MWE in a treebank. The three queries correspond to different levels of agreement between the MWE and the sentences of the corpora. They are the *major lemma query*, the *near-miss query*, and the *MWE query*. The query generation process has been explained in detail in (Odijk et al., 2024).

The MWE query (MEQ) searches for sentences that contain an occurrence of the MWE. The near-miss query (NMQ) searches for sentences in which the major words[7] of the MWE occur in the grammatical configuration in which they occur in the MWE. It allows the presence of determiners and modifiers that are not expected on the basis of the MWE's canonical form. The results of the NMQ are a superset of the results of the MEQ. MWE-Finder enables the user to inspect the difference between the results of the NMQ and the results of the MEQ. The Major Lemma Query (MLQ) searches for sentences in which the major words of the MWE occur in any grammatical configuration. The results of the MLQ are a superset of the results of the NMQ.

Next, the user can select the treebank that the query should be applied to. Once selected, the application switches to the *Results* view where query results are displayed as they arrive from the server. In that view, the user can also switch between the different queries for the chosen MWE or choose to exclude results of finer-grained queries. It is also possible to inspect or manually change the automatically generated XPath queries and retrieve new results. In the *Results* view, users can also look at the parse trees for results or toggle extra context (one preceding sentence, one following sentence) to better analyze the occurrences found, just like in GrETEL.

Finally, there is the analysis step, which we will not describe here and for which we kindly refer to (Odijk et al., 2024).

Odijk et al. (2024) illustrated MWE-Finder using the example MWE *de dans ontspringen* (lit. 'the dance escape', 'to escape the nasty consequences'). Here we will carry out small case studies for three different MWEs to show the potential of MWE-Finder and to identify opportunities for improving it.

5 Evaluating MWE-Finder

In this paper we evaluate MWE-Finder by assessing the performance of three different MWEs. In such an evaluation multiple factors play a role:

- The quality of the Alpino parses;
- The quality of the canonical form;

[6] DUtch CAnonical form Multiword Expressions, https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP.

[7] The major words of an MWE are the content words if there is more than one, and the content and function words otherwise.

- The quality of the automatic query generation.

We focus on the quality of the automatic query generation in this paper, but the other aspects will be relevant in some cases as well.

For each generated query we have specific expectations. We expect that the MEQ finds examples that satisfy all lexical, morphological and syntactic requirements that are encoded in the canonical form. Savary et al. (2019) found that when syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96% to 98% of the cases. Therefore we expect most of the results found by MEQ to have the idiosyncratic reading of the MWE.

We expect that the results of NMQ minus the results of the MEQ do not contain instances of the MWE. If it does contain them, it could mean that the canonical form for the MWE was too strict (see an illustration of this with the MWE *de dans ontspringen* 'to escape the nasty consequences' in (Odijk et al., 2024)). In such a case, the user can adapt the canonical form and evaluate with this revised canonical form. It can also mean that Alpino parsed the sentence incorrectly, or preferred one parse over another in case of an ambiguity, or that MWE-Finder's query generation mechanism contains errors or omissions.

As for the MLQ, it is the expectation that it finds all sentences in which the lemmas of the major words of the MWE occur. This is surely the case, but we have no guarantee that Alpino assigns the right lemmas to the words of an MWE in a particular sentence. This may happen if a word is ambiguous (e.g., *bommen*) and Alpino analyses it in a particular sentence as a noun ('bombs') rather than as a verb ('to concern') as required by the MWE. Or when a combination of the verb *passen* with the word *aan* is not analysed as a verb + preposition combination ('to fit to'), as required by the MWE, but as a verb + separable particle combination ('to adapt'). We are aware that such examples exist, but we will not deal with this in this paper. We are working on an even more general query (the *Related Word Query*, RWQ) to cover such cases, and we hope to describe this in more detail in future work.

We also expect that the results of the MLQ minus the results of the NMQ do not contain instances of the MWE. The latter can be the case, however, especially when Alpino parsed the sentence incorrectly, or when MWE-Finder's query generation mechanism contains errors or omissions.

We will deal with each of these aspects in the sections on specific MWEs. We will give links to the queries for each of the examples in the tables summarising the search results. For one MWE we have also listed the queries in the Appendix.

5.1 0Geen *+haan zal naar iets kraaien

In this section we illustrate MWE-Finder with the MWE canonical form in (5), which we will call the *target MWE*.

(5) 0geen *+haan zal naar iets kraaien
no rooster will to something crow
'no one will say anything about something'

We search in the treebank for the SoNaR corpus (Oostdijk et al., 2013), known in MWE-Finder as *Sonar4*. It contains more than 40 million sentences.

The canonical form states that the word *haan* can be modified and inflected, and that the word *geen* is not a component of this MWE. The queries automatically derived from this canonical form yield the results given in (6):

(6) Results of the queries derived from *0geen *+haan zal naar iets kraaien*[8]

Query	**Matches**
MEQ	331
NMQ	379
MLQ	804

We discuss the results of the three queries in separate subsections, the MLQ results in section 5.1.1, the NMQ results in section 5.1.2, and the MEQ results in section 5.1.3.

[8]The query names in the table are links to the query results in the application. Beware that there may have been updates to MWE-Finder since the time of writing, influencing results.

5.1.1 Analysis of the MLQ results

In order to analyse the MLQ results we only consider the MLQ results without the NMQ results. The NMQ results will be analysed in section 5.1.2.

The analysis of the MLQ results minus the NMQ results (804 – 379 = 425) has been summarised in Table 2. We discuss it here in detail.

MWE?	Cause	Details	Total
no	other MWE	*iemand zal victorie kraaien*	1
		iemands haan zal koning kraaien	5
			363
	variant-P + R-pronouns	*om*	17
		over	6
		achter	1
		Total	**393**
yes	*daar*-drop	(blank)	1
	dialect		1
	no P		3
	wrong sentence		1
	wrong parse	*naar* as ADJ, no `pc`	5
		PP attachment	1
	typo		1
	R-pronoun		17
		Total	**30**
unclear	no P		2
		Total	**2**
		Grand Total	**425**

Table 2: Analysis of the MLQ results minus the NMQ results (804 – 379 = 425) for *0geen *+haan zal naar iets kraaien*.

We first focus on example sentences that do not involve the target MWE. There are 393 such cases. 363 sentences contain forms of *haan* and *kraaien* but must be interpreted literally.

There are variants with other prepositions than *naar*: *om* (17), *over* (6), *achter* (1). For the authors all of these are ill-formed but *om* and *over* are probably correct variants of local dialects, given the number of occurrences. One would expect these as a result of the NMQ, but they are not there because the NMQ allows the absence of *naar* but still requires the presence of a prepositional complement, or because they contain R-pronouns,[9] which were not treated correctly yet at the moment of the measurement. We believe that it is necessary to reformulate the procedure to generate the NMQ so that it does not require an adpositional complement if its head need not be present.

In some cases the sentences contain a different MWE in which *haan* and *kraaien* happen to occur, e.g. *iemand zal victorie kraaien* (1 occurrence), and *iemands haan zal koning kraaien* (5 occurrences), an expression that the authors do not know (but it occurs in (Kruyskamp, 1974)):[10]

(7) different MWEs:

a. De Franse haan **kraait** dus **victorie**.
the French rooster crows so victory
'Thus the French celebrate the victory.' (WR-P-E-H_part00004.data.dz:4330)

[9] A limited set of pronouns with special syntax that all happen to contain the character *r*, e.g. *er* 'there', *waar* 'where', *hier* 'here'.

[10] The actual sentences are often very long. For long sentences we selected the relevant part and used '...' to indicate that we left out a part of the full sentence.

b. ... en voor de rest van de dag **kraait jouw haan koning**.
... and for the rest of the day crows your rooster king
'... and for the rest of the day you will have your way.' (WR-P-P-B_part00129.data.dz:7676)

There are 5 occurrences without a preposition, 3 of which are intended as the target MWE. For 2 it is unclear whether they are intended as the target MWE or as the literal expression (*unclear* in Table 2). The possibility for the adposition to be absent suggests that a variant without a prepositional complement must be included in DUCAME.

There are several occurrences in which *naar* is analysed as an adjective (meaning 'nasty') instead of a preposition, and no adpositional complement (`pc`) is present (5 occurrences).

One case involves topic drop of the complement of the preposition *naar*, most probably *daar*, a construction that is not in accordance with the norms of the standard language but that occurs frequently in colloquial language. MWE-Finder cannot currently handle this construction properly, so the example is in the MLQ results rather than in the MEQ results.

(8) ... **kraait** geen **haan naar**
... crows no rooster to
'... no one will say anything about that.' (WR-U-E-A_part00227.data.dz:9310)

17 cases indeed involve the target MWE but contain R-pronouns or pronominal adverbs not dealt with correctly yet by MWE-Finder at the time, which is why we find them among the results of the MLQ.

Finally, some have not been identified as the target MWE because the Alpino parse is completely wrong, in one case because of the strong dialectical nature of the sentence, in other cases caused by typos or grammar errors.

5.1.2 Analysis of the NMQ results

We analyze the NMQ results minus the MEQ results. This involves 48 examples. Most do not involve the target MWE. There are 2 occurrences of a different MWE (*iemands haan zal victorie kraaien*), 3 examples with the preposition *over* instead of *naar*, and 42 examples involve a literal interpretation of the words *haan* and *kraaien*. Only one case does involve the target MWE, but here *naar* has been parsed as an adjective, which is why we find this example among the NMQ results rather than among the MEQ results.

5.1.3 Analysis of the MEQ results

In example (9) Alpino analysed a directional PP incorrectly as a prepositional complement to the verb *kraaien*. It is clearly not an instance of the target MWE, because the sentence does not contain a negative polarity licensor (see below).

(9) ... en er kraaide een haan .- (14:68) naar het voorportaal ...
... and there crowed a rooster .- (14:68) to the vestibule ...
'... and a rooster crowed .- (14:68) to the vestibule ...' (WR-P-P-B_part00165.data.dz:4493)

All other results found among the MEQ results are indeed instances of the target MWE. A special case to mention is (10), in which the word *haan* is the head of a noun phrase that contains a relative clause that contains the verb *kraaien* with the relative pronoun *die*. MWE-Finder can correctly handle such constructions.

(10) Geen **haan** die **ernaar kraaide**.
no rooster that there.to crowed
'(There was) no one who said anything about it.' (WR-P-E-A_part00099.data.dz:8178)

As for form variants of *haan*, we mostly find *haan*, but occasionally also the plural form *hanen*. We encountered no diminutive forms.

As far as determination is concerned, *geen* 'no' was the most frequent determiner, but the indefinite article *een* 'a' occurred approximately 20 times (including the dialectical variant *ne*) and *weinig* 'few' had 2 occurrences.

The MWE is a negative polarity construction, i.e. it requires a negative polarity licensor (NPL) in the sentence (and in the right position). The determiners *geen* 'no' and *weinig* 'few' are such NPLs. In the examples where the determiner *een* occurs we observed NPLs such as *zonder* (as in (11a)) 'without', *nauwelijks* 'hardly', *amper* 'hardly', *nooit* 'never' and yes-no questions (as in (11b)):

(11) Some examples with negative polarity licensors:

a. ... zonder dat er in het Westen een **haan naar kraaide** ...
... without that there in the West a rooster to crowed ...
'... without anyone saying anything about it in the West ...'
(WR-P-E-A_part00199.data.dz:2034)

b. ... is daar een **haan** die daar **naar kraait**?
... is there a rooster that there to crows
'... is there anyone who says anything about that?' (WR-P-E-A_part00021.data.dz:1667)

c. ... zonder dat er ook maar een diplomatieke **haan naar kraait**.
... without that there also but a diplomatic rooster to crows
'... without any diplomat saying anything about it.' (WR-P-P-H_part00272.data.dz:2312)

MWE-Finder currently does not check for the presence of NPLs, but this surely could be a useful extension of the MEQ. In fact, it is crucial if we would allow *0geen haan zal kraaien* (without the PP) as a canonical form: if there is no check for a polarity licensor any sentence with *haan* in the subject of *kraaien* will match.

Modifiers of *haan* were most often absent, but we encountered the adjectives *diplomatieke* 'diplomatic' (as in 11c), *Europese* 'European', and *rode* 'red',[11] and relative clauses.

5.2 Iemand zal 0dat *+varken wassen

For another illustration of MWE-Finder we use the MWE as in (12), which we shall again refer to as the *target MWE*.

(12) iemand zal 0dat *+varken wassen
someone will that pig wash
'someone will deal with that problem'

The canonical form of the target MWE indicates that *dat* 'that' is not part of the MWE, and that *varken* can be modified, determined and inflected.

For this example, we query the Mediargus corpus (containing around 103 million sentences), which is pre-configured within MWE-Finder. The queries automatically derived from the target MWE yield the results given in (13):

(13) Results of the queries derived from *iemand zal 0dat *+varken wassen*:

Query	**Matches**
MEQ	537
NMQ	537
MLQ	615

As can be seen in (13), the MEQ and the NMQ yield the same hits.[12] All 537 but 2 were instances of the target MWE, with the 2 sentences without the MWE requiring a literal reading of *varken* and *wassen*. The fact that the MEQ and the NMQ yield the same results, suggests that the canonical form is correct as is. Indeed, *dat* is not a part of the MWE: in many instances of the MWE *dat* is replaced by another determiner such as the definite article *het*, or no determiner is present at all. Furthermore, *varken* is often attested in inflected form (it often occurs in its diminutive form *varkentje*, but not exclusively) and modified (it is possible to have an adjective modify the noun). All this is illustrated in one single hit, given in (14).

[11] Most probably intended as 'socialist' or 'communist'.

[12] Remember that the MEQ results are a subset of the NMQ results; when the sets are of equal size, they must contain the same items.

(14) Ik denk dat jonge progressieven wel andere **varkentjes** te **wassen** hebben.
I think that young progressive POS other pigs.DIM to wash have
'I think young progressives have other things to deal with.' (DM_20041213_01.data.dz:1685)

The MLQ, however, found more than the MEQ and NMQ. 78 more sentences were retrieved containing the "major lemmas", i.e. the content words, of the target MWE, regardless of their grammatical relation. Of these 78 hits, 38 were clear instances of the target MWE, while 35 clearly were not, for instance where *varken* was not even a direct object of *wassen*. The remaining 5 hits were more difficult to judge and required closer inspection of the surrounding sentences in the corpus, a feature that MWE-Finder offers. These 5 hits were found to indeed be instances of the target MWE.

Of the 43 (= 38 + 5) misses, 20 can be attributed to a wrong parse, for instance caused by missing punctuation (e.g. (15), which would have been found if it had an extra comma between *gewassen* and *dacht*, as that would have resulted in a correct parse by Alpino), by Alpino's treatment of quotes, the words of which are all put in a flat list of nodes under a `mwu` (multiword unit) phrase (e.g. (16)), or complex, long-distance dependencies (e.g. (17), in which Alpino takes *uit dit administratieve varkentje* as PP, while *uit* should be analyzed as a particle belonging to the main verb *dagen* 'to challenge'). These parsing mistakes are highlighted because they occur more than once.[13]

(15) $ **Varkentje gewassen** dacht iedereen, behalve de bezoekers.
pig.DIM washed thought everybody except the visitors
'Problem dealt with, everybody thought, except for the visitors.'
(NB_19990308_01.data.dz:11479)

(16) $ ["We hadden twee uur nodig om dit **varkentje** te **wassen**",]$_{\texttt{mwu}}$ is de Wevelgemse trainer Marnix
we had two hour needed for this pig.DIM to wash is the Wevelgem trainer Marnix
Pattyn vol lof over de Markse tegenstander.
Pattyn full praise about the Marke opponent
'"We needed two hours to deal with them," Wevelgem trainer Marnix Pattyn praised his opponents from Marke.' (NB_19981202_01.data.dz:12777)

(17) $ Landbouwers dagen staatssecretaris van Administratieve Vereenvoudiging, Vincent Van
farmers challenge Secretary of State of administrative simplification Vincent Van
Quickenborne (VLD), [uit dit administratieve **varkentje**]$_{\text{PP}}$ te **wassen**.
Quickenborne PRT this administrative pig.DIM to wash
'Farmers challenge Secretary of State for Administrative Simplification, Vincent Van Quickenborne (VLD), to deal with this administrative problem.' (NB_20040602_01.data.dz:160)

The remaining 23 results did contain instances of the target MWE, but were missed. It concerns 7 cases of finite relative clauses, with *varken* as antecedent and the verb *wassen* inside the relative clause. While MWE-Finder can deal with relative clauses correctly (cases of relativization were found with the MEQ and NMQ), these cases were missed due to the relative pronoun *dat* being wrongly tagged as a subordinating conjunction (`vg` instead of `vnw`), such as in (18). It is interesting, however, to note that the Mediargus corpus was parsed with an older version of Alpino, and that the wrong parse cannot be recreated using the newer version embedded within MWE-Finder.[14]

(18) Woluwe is een **varkentje** dat$_{\texttt{vg}}$ jullie moeten kunnen **wassen**?
Woluwe is a pig.DIM that 2.PL must can wash
'Woluwe is something that you should be able to deal with?' (NB_20050930_01.data.dz:3975)

Another 13 can all be attributed to the object relation between *varken* and *wassen* not being explicitly labelled by Alpino, or indeed recognized at all, in certain specific non-frequent constructions. During the

[13] With the notation '$', we mean that while the sentence is grammatical (ignoring normative spelling and punctuation rules), the parse associated with the sentence is wrong.

[14] We do not count these tagging errors towards wrong parses, because in certain cases it is correct to analyse the relative pronoun as a `vg` (though not in this one), and MWE-Finder should be robust against this, which these examples showed us it is not. For instance in *de dag dat jij naar Leuven ging* 'the day you went to Louvain', *dat* cannot be a `vnw`, as it does not agree with its antecedent *dag*.

development of MWE-Finder, these constructions were not accounted for, and should be implemented in a future version. These can be broken down into:

- 4 cases of nonfinite relative clauses, with an infinitive and a zero object. In these cases the noun *varken* is modified by an infinitival clause with *wassen* accompanied by the particles *om* and *te*, in which *varken* is understood as the object of *wassen*, e.g. (19).

  (19) Een **varkentje** om te **wassen**
  a pig.DIM for to wash
  'A problem to deal with', lit. 'A piglet to wash' (DM_20060429_01.data.dz:4833)

- 2 cases of a reflexive middle construction involving the permissive verb *laten* 'to let' (Broekhuis et al., 2016), such as in (20). Note that the construction is itself embedded in a relative clause, but that is not what causes the sentence to be missed by the MEQ and NMQ.

  (20) Het probleem van Carestel is geen **varkentje** dat zich snel **laat wassen**.
  the problem of Carestel is no pig.DIM that SELF quickly lets wash
  'The problem of Carestel is not one quickly dealt with.' (DS_20031129_01.data.dz:5618)

- 2 cases of a deverbal noun causing the object *varken* to end up in a PP headed by *van* 'of', and therefore losing its object relation to the verb *wassen*, e.g. (21), which also contains the MWE *de kous zal af zijn* 'it will be settled'.

  (21) Of dachten de heren en dames van de partijbesturen dat met het **wassen** van het
  or thought the gentlemen and ladies of the party.leaderships that with the washing of the
  Antwerpse **varkentje** de kous af is?
  Antwerp pig.DIM the stocking off is
  'Or did the ladies and gentlemen of the party leaderships think that dealing with Antwerp would solve everything?' (LN_20030405_01.data.dz:937)

- 2 cases of *varkentjes* being deeply embedded, e.g. (22). The problem here is that *varkentjes* is a modifier to a quantity noun (*aantal* 'number') and is itself modified by an adjective, resulting in an NP inside an NP and the loss of the object relation between *varkentjes* and *wassen*. Alpino can analyse constructions such as these in two ways: one where it takes *aantal* as the head of the NP (which is what happens in this parse in question); the other where it takes *varkentjes* as the head of the NP with *een flink aantal* as a complex determiner, in which case MWE-Finder would have been able to find this instance of the target MWE. Whether or not Alpino disambiguated wrongly between these two analyses, MWE-Finder could be made more robust for constructions such as these, so that it can find instances with the former analysis, too.

  (22) Maar wie dit jaar Brad Pitt of Julia Roberts van nabij wil monsteren, zal eerst [een flink
  but who this year Brad Pitt or Julia Roberts from near wants inspect will first a large
  aantal [bureaucratische **varkentjes**]$_{NP}$]$_{NP}$ moeten **wassen**.
  number bureaucratic pigs.DIM must wash
  'But those who want to inspect Brad Pitt or Julia Roberts from up close this year, must deal with quite a few bureaucratic hurdles first.' (DM_20020206_01.data.dz:2169)

- 2 cases of an adjectivized past participle, e.g. (23), which were missed because Alpino does not explicitly label the object relation between the head noun and the attributive past participle.

  (23) Allicht, want Beringen, zou een snel **gewassen varkentje** worden.
  obviously for Beringen would a quickly washed pig.DIM become
  'Obviously, for Bering would quickly be dealt with.' (NB_20010417_01.data.dz:9265)

- one case of a modal infinitive, expressing some notion of ability or obligation (Broekhuis & Keizer, 2015), as in (24).

(24) ≪Nochtans moeten wij ons niet zenuwachtig maken want zowel Neerlandia als Hoboken 2000
nevertheless must we us not nervous make for both Neerlandia as Hoboken 2000
zijn te **wassen varkentjes**.
are to wash pigs.DIM

'Nevertheless, we must not get nervous, for both Neerlandia and Hoboken 2000 are to be dealt with.' (LN_20061209_01.data.dz:15625)

Finally, there were 3 cases of a passive perfect without a copula, e.g. (25). In Dutch the passive perfect uses the copula *zijn* 'to be' as auxiliary, which in specific cases may be absent. 2 out of the 3 hits were additionally embedded within an accusativus cum infinitivo (AcI) construction, e.g. (26). For us the authors, all three hits are only marginally grammatical, but they may be considered more grammatical in Belgian Dutch.[15]

(25) ? Eens dat **varkentje gewassen**, heeft Frank de handen vrij en dat zal niet alleen zijn om zich in
once that pig.DIM washed has Frank the hands free and that will not only be for SELF in
het nachtleven te werpen.
the nightlife to throw

'Once that has been dealt with, Frank will have his hands free, and that won't be just to immerse himself into the nightlife.' (DM_20050223_01.data.dz:4554)

(26) ? Bij Antonia dacht men het **varkentje gewassen**.
at Antonia thought one the pig.DIM washed

'Those at Antonia thought the problem dealt with.' (NB_19981014_01.data.dz:3343)

These 78 misses therefore do not suggest an adjustment of the canonical form, but rather point to possible points of improvement for MWE-Finder's algorithm, which will require some form of enrichment of the parses given by Alpino or an extension of the query generation mechanism. Several infrequent Dutch constructions lead to the loss of the object relation of *varken* to *wassen* in Alpino's parse, and indeed of any direct object part of a verbal MWE. The implementation of these constructions in MWE-Finder remains a *varkentje om te wassen* in itself.

5.3 Iemand zal iemand het hemd van het lijf vragen

As a final illustration we search for the MWE in (27) as target MWE:

(27) iemand zal iemand het hemd van het lijf vragen
someone will someone the shirt from the body ask

'someone will want to know all the ins and outs of something from someone'

We search in Mediargus. The results of the three queries are summarised in (28):

(28) Results of the queries derived from *iemand zal iemand het hemd van het lijf vragen*:

Query	**Matches**
MEQ	30
NMQ	32
MLQ	43

All 11 examples from the MLQ results minus the NMQ results contain an occurrence of the target MWE. 10 sentences have been parsed completely wrongly by Alpino. In one sentence Alpino produces a different parse than expected, which prevents recognition of the target MWE, but additionally, one of the MWE components is modified (29):

[15] It could be argued that (26) is missing some form of punctuation between *men* and *het*. This would render the analysis of the sentence rather different with a speech tag at the start of the sentence, and would allow MWE-Finder to find this MWE with the MEQ or NMQ. The reason to argue for this is the sentence's marginal grammaticality for the authors if it is analysed as an AcI, however more instances of an AcI embedded under the verb *denken* were attested in the corpus, suggesting that it may be considered more grammatical in Belgian Dutch and corroborating the AcI analysis. Alpino, though, remains unable to parse it as an AcI.

(29) ... **vroegen** ze Marc **het** bezwete **hemd van het lijf**.
... asked they Marc the sweating shirt from the body
'... they wanted to know all the ins and outs from sweating Marc.'
(NB_20010104_01.data.dz:5254)

It is not clear that we should change the canonical form for the target MWE based on this single example. It seems that the modifier does not really semantically modify *hemd* (this sentence can be uttered even if Marc is not wearing a shirt) but rather *Marc*.

In example (30) a variant of the expression occurs, with *de hemd* instead of *het hemd*. It is unclear to us whether this is a performance error or a real variant of the expression. The example is also in passive, and there is agreement between the indirect object and the finite verb. The latter is ill-formed according to normative grammar though it occurs often in colloquial speech. Despite the fact that Alpino cannot analyse this aspect correctly, the parse is sufficiently good to recognize this variant of the target MWE.

(30) ... werden de klanten **de hemd van het lijf gevraagd** ...
... were the customers the shirt from the body asked ...
'... they wanted to get to know all ins and outs from the customers...'
(LN_20030821_01.data.dz:10761)

Of the 2 results in the NMQ results minus the MEQ results, one sentence is parsed incorrectly, and in the other sentence (31) a variant of the expression occurs with *zijn lijf* 'his body' instead of *het lijf* 'the body'. We find this expression well-formed, so we should include it in a new version of DUCAME.

(31) Ik **vroeg** hem **het hemd van zijn lijf**.
I asked him the shirt from his body
'I wanted to know all ins and outs from him.' (NB_19990710_01.data.dz:14448)

All sentences from the MEQ results contain the target MWE. There are rather complex sentences among them, e.g. one with a large verb cluster (32a) and one in passive voice (32b).

(32) a. ... een journalist die je al zo vaak **het hemd van het lijf** heeft proberen te **vragen**.
... a journalist who you already so often the shirt from the body has tried to ask
'... a journalist that has already often tried to get to know all ins and outs from you.'
(DM_19980206_01.data.dz:305)

b. ... telkens weer wordt je **het hemd van het lijf gevraagd**.
... each time again is you the shirt from the body asked
'... again and again they want to get to know all ins and outs from you.'
(DS_20010905_01.data.dz:4124)

6 Related Work

There has been abundant research in developing automatic tools to reliably search for occurrences of MWEs (MWE Identification in the sense of (Constant et al., 2017). An excellent survey of recent work on MWE identification is (Ramisch et al., 2023). It mostly describes and evaluates work done in the context of the 2016 SEMEVAL DiMSUM Shared Task[16] (Schneider et al., 2016) or PARSEME Shared Tasks (Ramisch et al., 2018, 2020; Savary et al., 2017).

We want to highlight here the commonalities and differences of the approach taken in this paper in comparison to most other work.

First of all, all works on MWE identification describe software tools to annotate text for MWEs. None of the works describe a software application with a dedicated user interface and an identified target user group. MWE-Finder is a web application with a user interface targeted at linguists and lexicographers that do research on MWEs or try to describe them.

Second, none of the works on MWE identification target the Dutch language. To our knowledge, MWE-Finder is the first piece of software to identify Dutch MWEs. So far, linguists and lexicographers

[16] https://dimsum16.github.io/

		found by MEQ?				
		Yes	**No**	**Total**	**Acc.**	93.7%
	Yes	895	87	982	**Prec.**	99.7%
contains MWE?	**No**	3	475	478	**Rec.**	91.1%
	Unclear	0	2	2	$\mathbf{F}_1$	95.2%
	Total	898	562	1462		

Table 3: Results of the MEQs for the three case-study MWEs. Note that **No** also includes literal readings of the target MWEs.

investigating Dutch MWEs defined queries in OpenSoNaR or similar interfaces (Nederlab (Brugman et al., 2016), Corpus of Contemporary Dutch[17]) to search for the lemma's of an MWE in each other's neighborhood within a sentence, basically equivalent to the major lemma query of MWE-Finder. These corpus search applications provide no special facilities for searching for MWEs.

Though MWE-Finder has been developed for Dutch, the general approach is not restricted to the Dutch language. However, this is not the place to discuss this in detail. Instead we refer to (Odijk et al., 2024), which describes what would be needed to create an MWE-Finder for other languages.

Ramisch et al. (2023, p. 106) classify approaches in a number of categories (paradigms): syntactic parsing, compositionality prediction of MWE candidates, and sequence annotation. Our work falls in the syntactic parsing paradigm, and is in this respect comparable to the work by Nagy T. and Vincze (2014) focusing on English verb particle constructions and Constant and Nivre (2016), who developed a general strategy for MWE identification and tested it on French and English data.

Many works focus on specific MWE subclasses, e.g. the PARSEME-related work focused on verbal MWEs, Nagy T. and Vincze (2014) focused on verb particle constructions. MWE-Finder aims to cover all MWE classes, especially the most difficult class of flexible MWEs.

Finally, in MWE-Finder currently nothing special is done to distinguish the MWE-interpretation of an expression from a literal interpretation of the expression.

7 Discussion and conclusion

In this paper we have showcased MWE-Finder's performance through the use of three small case studies of Dutch MWEs.

We observed that MWE-Finder is very accurate in retrieving the target MWE using the MEQ. Of all sentences in the corpus containing at least the relevant lemmas of the content words (i.e., the results of the MLQs), the MEQ correctly found 895 of them to contain the target MWE while it also correctly identified 475 not to contain the target MWE. Out of the total 1462 (remember that there were 2 hits for which it remained unclear whether they did or did not contain the MWE; see Table 2), this results in an accuracy for the MEQ of 93.7%. In terms of precision and recall, the MEQ shows a precision for these three case studies of 99.7%, finding only 3 sentences that did not concern an instance of the target MWE. Recall is slightly lower at 91.1%, missing 87 cases of the target MWE. An F_1-score is then calculated as 95.2%, see Table 3. Note that these numbers also include literal readings, and our numbers therefore nicely reflect (Savary et al., 2019), which found that when syntactic conditions necessary for an idiomatic reading are fulfilled, this reading occurs in 96% to 98% of the cases.

Most false negatives, i.e. the instances of the target MWEs that the MEQ missed, can be attributed to wrong parses, an issue that will likely remain to impair the results of MWE-Finder as it depends on Alpino and the cleanliness of the data. While not many, the false positives, i.e. the sentences that the MEQ did retrieve despite them not containing the target MWE, can be attributed to the fact that MWE-Finder was not designed to distinguish between a metaphorical and literal reading of the components of the MWE.

The results of the NMQ showed its effectiveness in guiding the user in the formulation of a better,

[17] https://ivdnt.org/corpora-lexica/corpus-hedendaags-nederlands/

more complete description of the MWE in a canonical form. For instance, it suggested the addition of the canonical form of *iemand zal iemand het hemd van het lijf vragen*, given that it retrieved a well-formed variant of the MWE with the possessive pronoun *zijn* 'his' instead of *het* before *lijf*. It may be expected that the performance of MWE-Finder's MEQ will be even higher when using the updated canonical forms.

The MLQ proved useful in identifying points of improvement for MWE-Finder's system. While the wrong parses produced by Alpino may not be mitigated in the near future, especially the case study on *iemand zal 0dat *+varken wassen* laid bare that in quite a few (though infrequent) constructions it fails to recognise the object relation between *varken* and *wassen*, a problem likely to arise in other MWEs as well. MWE-Finder can thus be improved by, e.g., enriching the Alpino parses by making the object relation in such constructions explicit. Also the issue concerning the labelling of relative pronouns (`vg` vs. `vnw`) is an obvious point of improvement, as well as the reformulation of the way the NMQ is generated and how it deals with adpositional complements. Another valuable addition can be found in a way to check for the presence of NPLs.

There is, though, one caveat with regards to the performance as presented in Table 3: it may be – and probably is – the case that the MLQ has missed cases of the target MWEs, for instance where a homonym received a wrong POS tag or was lemmatised wrongly, as was already pointed out in Section 5. The MLQ specifically searches for the lemmas of the content words in the MWE, along with their POS tags as derived from the canonical form, and will miss any such case that was tagged otherwise. The results presented in Table 3 may therefore be somewhat skewed, but the argument that MWE-Finder works well, we believe, still holds, also considering that the potential cases missed by the MLQ would be the result of Alpino's parse, not MWE-Finder's algorithm directly. As said before, however, we are working on the formulation of a fourth query (the Related Word Query, RWQ) that is even more relaxed than the MLQ in order to retrieve candidates with wrongly tagged content words.

Using MWE-Finder does sometimes entail closer, manual inspection of the data – also in the case studies presented in this paper. This closer inspection can be streamlined with a special analysis component built into MWE-Finder. A preliminary version of such a component already exists, but will require more testing before being definitively implemented into the application.

In the future MWE-Finder's performance may be evaluated more formally using the Dutch MWE-corpus currently being produced by Bouma et al. (2024). For now, we believe that the current case studies presented in this paper provide a simple evaluation, showcasing and roadmap for further development of MWE-Finder.

References

Augustinus, L., Vandeghinste, V., & Van Eynde, F. (2012, May). Example-based treebank querying. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC 2012)* (pp. 3161–3167). European Language Resources Association (ELRA).

Bouma, G., Odijk, J., & Tiberius, C. (2024, February). Towards a Dutch Parseme corpus. In *Proceedings of the second general UniDive meeting, Naples, Italy*. UniDive Project. https://unidive.lisn.upsaclay.fr/doku.php?id=meetings:general_meetings:2nd_unidive_general_meeting

Broekhuis, H., Corver, N., & Vos, R. (2016, January). 3.2.2.5. The reflexive middle construction [Retrieved January 29, 2024 from https://taalportaal.org/taalportaal/topic/pid/topic-14406719669673231]. https://taalportaal.org/taalportaal/topic/pid/topic-14406719669673231

Broekhuis, H., & Keizer, E. (2015, December). 3.2.III. Modal infinitives [Retrieved January 29, 2024 from https://taalportaal.org/taalportaal/topic/pid/topic-14286557605977117]. https://taalportaal.org/taalportaal/topic/pid/topic-14286557605977117

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., & van den Bosch, A. (2016, May). Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard,

J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)* (pp. 1277–1281). European Language Resources Association (ELRA).

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892. https://doi.org/10.1162/COLI_a_00302

Constant, M., & Nivre, J. (2016, August). A transition-based system for joint lexical and syntactic analysis. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: Long papers)* (pp. 161–171). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1016

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 245–257). Ubiquity. https://doi.org/http://dx.doi.org/10.5334/bbi.20

Kruyskamp, C. (1974). *Groot woordenboek der Nederlandse taal* (10th ed.). Martinus Nijhoff.

Nagy T., I., & Vincze, V. (2014, April). VPCTagger: Detecting verb-particle constructions with syntax-based methods. In V. Kordoni, M. Egg, A. Savary, E. Wehrli, & S. Evert (Eds.), *Proceedings of the 10th workshop on multiword expressions (MWE)* (pp. 17–25). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0803

Odijk, J. (2013). Identification and lexical representation of multiword expressions. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch. Results by the STEVIN-programme* (pp. 201–217). Springer. http://link.springer.com/content/pdf/10.1007

Odijk, J. (2023). A canonical form for Dutch multiword expressions (version 1.0) [Part of the DUCAME documentation]. https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

Odijk, J., & Kroon, M. (2024, May). A canonical form for flexible multiword expressions. In *Proceedings of LREC-COLING 2024*. European Language Resources Association (ELRA).

Odijk, J., Kroon, M., Baarda, T., Bonfil, B., & Spoel, S. (2024). MWE-Finder: Querying for multiword expressions in large Dutch text corpora. In V. Giouli & V. B. Mititelu (Eds.), *Multiword expressions in lexical resources. Linguistic, lexicographic and computational perspectives*. Language Science Press.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500 million word reference corpus of contemporary written Dutch [http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1]. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch: Results by the STEVIN-programme* (pp. 219–247). Springer.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.-P., Moortgat, M., & Baayen, H. (2002). Experiences from the Spoken Dutch Corpus project. In *Proceedings of the third international conference on language resources and evaluation (LREC-2002)* (pp. 340–347). ELRA.

Ramisch, C., Cordeiro, S. R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., … Walsh, A. (2018, August). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In A. Savary, C. Ramisch, J. D. Hwang, N. Schneider, M. Andresen, S. Pradhan, & M. R. L. Petruck (Eds.), *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (LAW-MWE-CxG-2018)* (pp. 222–240). Association for Computational Linguistics. https://aclanthology.org/W18-4925

Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., & Xu, H. (2020, December). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In S. Markantonatou, J. McCrae, J. Mitrović, C. Tiberius, C. Ramisch, A. Vaidya, P. Osenova, & A. Savary (Eds.), *Proceedings of the joint workshop on multiword expressions and electronic lexicons* (pp. 107–118). Association for Computational Linguistics. https://aclanthology.org/2020.mwe-1.14

Ramisch, C., Walsh, A., Blanchard, T., & Taslimipoor, S. (2023, May). A survey of MWE identification experiments: The devil is in the details. In A. Bhatia, K. Evang, M. Garcia, V. Giouli, L. Han, & S. Taslimipoor (Eds.), *Proceedings of the 19th workshop on multiword expressions (MWE 2023)* (pp. 106–120). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.mwe-1.15

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2001). Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper*, *2001-03*. http://lingo.stanford.edu/csli/pubs/WP-2001-03.ps.gz

Savary, A., Cordeiro, S., Lichte, T., Ramisch, C., Iñurrieta, U., & Giouli, V. (2019). Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, *112*(1), 5–54. https://doi.org/doi:10.2478/pralin-2019-0001

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., & Doucet, A. (2017, April). The PARSEME shared task on automatic identification of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary, & V. Vincze (Eds.), *Proceedings of the 13th workshop on multiword expressions (MWE 2017)* (pp. 31–47). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1704

Schneider, N., Hovy, D., Johannsen, A., & Carpuat, M. (2016, June). SemEval-2016 task 10: Detecting minimal semantic units and their meanings (DiMSUM). In S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, & T. Zesch (Eds.), *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)* (pp. 546–559). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1084

van de Camp, M., Reynaert, M., & Oostdijk, N. (2017). WhiteLab 2.0: A web interface for corpus exploitation. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 231–243). Ubiquity. https://doi.org/http://dx.doi.org/10.5334/bbi.19

van Noord, G. (2006). At last parsing is now operational. In P. Mertens, C. Fairon, A. Dister, & P. Watrin (Eds.), *TALN06 verbum ex machina. Actes de la 13e conférence sur le traitement automatique des langues naturelles* (pp. 20–42).

van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Tjong Kim Sang, E., & Vandeghinste, V. (2013). Large scale syntactic annotation of written Dutch: Lassy. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch* (pp. 147–164). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30910-6_9

Appendix: Queries

In this appendix we list the queries for the MWE *iemand zal 0dat *varken wassen*. For this particular example, the MEQ and NMQ queries are identical.

MEQ
```
//node[
       node[@rel="obj1" and @cat="np" and
           node[@lemma="varken" and @rel="hd" and
                @pt="n" and @ntype="soort" and
                (@genus="onz" or @getal="mv")]] and
       node[@lemma="wassen" and @rel="hd" and @pt="ww"]]
```

NMQ
```
//node[
       node[@rel="obj1" and @cat="np" and
           node[@lemma="varken" and @rel="hd" and
                @pt="n" and @ntype="soort" and
                (@genus="onz" or @getal="mv")]] and
       node[@lemma="wassen" and @rel="hd" and @pt="ww"]]
```

MLQ
```
//node[@lemma="varken" and @pt="n"]/ancestor::alpino_ds/
       node[@cat="top" and
            descendant::node[@lemma="wassen" and @pt="ww"]]
```