Finding Dutch multiword expressions

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

Martin Kroon
Utrecht University, the Netherlands
m.s.kroon@uu.nl

Tijmen Baarda
Utrecht University, the Netherlands
t.c.baarda@uu.nl

Ben Bonfil
Utrecht University, the Netherlands
b.bonfil@uu.nl

Sheean Spoel
Utrecht University, the Netherlands
s.j.j.spoel@uu.nl

Abstract

We present MWE-Finder, which enables a user to search for occurrences of multiword expressions (MWEs) in large Dutch text corpora. Components of many MWEs in Dutch can occur in multiple forms, need not be adjacent, and can occur in multiple orders (such MWEs are called *flexible*). Searching for occurrences of such flexible MWEs is difficult and cannot be done reliably with most search applications. What is needed is a search engine that takes into account the grammatical configuration of the MWE. MWE-Finder is therefore embedded in GrETEL, a treebank search application for Dutch. A user can enter an example of a MWE in a specific canonical form, after which the system searches for sentences in which the MWE occurs, using queries generated automatically from the canonical form. The MWE can also be selected from a list of more than 11k canonical forms for Dutch MWEs that MWE-Finder offers. We will show that MWE-Finder also offers facilities to find examples with unexpected modifiers or determiners on components of the MWE

1 Introduction

A multiword expression (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). 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 unpredictabe form (e.g. *ter plaatse* 'on location', with idiosyncratic use of *ter* and *e*-suffix on the noun), 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.

Many Dutch multiword expressions (MWEs) are flexible in the sense that their components can have different forms, can occur in different orders, or can have words that do not belong to the MWE between them. This flexible nature of such MWEs makes it difficult to reliably search for occurrences of 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 developed in the context of CLARIN 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 can find all instances but one will at the same time find many instances where all these words occur but not the MWE. One should be able to search for a flexible MWE in such a way that its grammatical structure is taken into account.

We present MWE-Finder, which is intended as a research tool for any linguist or lexicographer interested in research into multiword expressions, in particular *flexible* multiword expressions. MWE-Finder can take the grammatical structure of a MWE into account because it is embedded in a new version (version 5) of GrETEL,[1] an existing web application for searching Dutch treebanks developed in the context

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

of CLARIN (Augustinus et al., 2012; Augustinus et al., 2017; Odijk et al., 2018). The distinguishing feature of GrETEL is its query-by-example feature. In its regular search mode, it leads the user through a number of steps to get from an example sentence to search results and analysis of the search results. MWE-Finder mimics this approach specifically for MWEs. We describe the relevant steps in section 2.

A second important feature of GrETEL is that one can upload one's own text corpus, which is then automatically parsed and made available as a treebank to search in. This feature is therefore automatically also available for MWE-Finder.

GrETEL is open source and its code is available on GitHub.[2] The part of the application that generates queries for MWEs and that performs the tree manipulation is available as a separate Python package, so that it may also be used to create scripts that search treebanks without using GrETEL.[3] We are in contact with researchers of the Institute for the Dutch Language (INT) to host GrETEL5 when it is completely finished at the recognized CLARIN Type B-center INT.[4]

2 MWE-Finder

The user goes through a number of steps to obtain the desired results: (1) example MWE; (2) treebank selection; (3) query results ; (4) analysis of the query results. We describe each of these steps here.

2.1 Example MWE

MWE-Finder enables a user to search for occurrences of a MWE in a treebank based on an example MWE. The example MWE must be in a specific canonical form. For single words the canonical form is its lemma. However, for reasons that will be described in the full paper, in many verbal MWEs one cannot use just the lemma for the head of the MWE. Instead, we require that a verbal MWE is an infinitival complement to the future auxiliary verb *zullen* 'will'. A concrete example is (1) in which the indefinite pronoun *iemand* in the canonical form means that any phrase can occur here:

(1) iemand zal de dans ontspringen
someone will the dance spring
'someone will have a lucky escape'

It is assumed that the head of the MWE can be inflected, modified and determined, but that other parts of the MWE cannot. Of course, there are many exceptions to this, and these are indicated by means of annotations. There are annotations to mark (un)modifiability and (un)inflectability of MWE components, not being a component of the MWE, specific limited types of variation, and for variables parts of the MWE, etc., as will be explained in detail in the full paper. With the canonical form of the MWE the user implicitly formulates a hypothesis about the properties of this MWE.

MWE-Finder offers the user a large list of MWEs in canonical form to select from. This list was derived from the DUtch CAnonicalised Multiword Expressions (DUCAME) resource.[5]

2.2 Treebank selection

The user selects the treebank or treebanks that the MWE should be searched in. As a concrete example, one could choose the treebank MEDIARGUS, which contains texts from Belgian newspapers (more than 103 million sentences).

2.3 Query results

MWE-Finder automatically generates three queries from the MWE example in canonical form to search for occurrences of this MWE in a treebank. These are the *major lemma query*, the *near-miss query*, and the *MWE query*.[6] The query generation process has been described in detail in (Odijk et al., to appear).

[2] https://github.com/UUDigitalHumanitieslab/gretel

[3] https://github.com/UUDigitalHumanitieslab/mwe-query

[4] https://www.ivdnt.org/.

[5] See (Odijk et al., to appear) and https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP.

[6] Note that MWE-Finder can identify potential occurrences of a MWE in a treebank. It cannot determine for an expression that is ambiguous between a literal and an idiomatic reading which of these alternative readings is applicable in a specific sentence.

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.

The major lemma query searches for sentences in which at least the lemmas of the major words of the MWE occur (in any grammatical configuration). Major words are the content words if there are at least two in the MWE, and content and function words if there is at most one content word in the MWE. The query yields a superset of the results of both other queries. This query is applied to the full treebank, making use of indexes on the treebank to speed up the process. The major lemma query yields a list of syntactic structures, and can be used to identify the MWE in a grammatical configuration that was not expected at all, to retrieve occurrences of the MWE in sentences that the Alpino parser used in GrETEL parsed incorrectly, or to retrieve occurrences of the MWE for which MWE-Finder did not generate the correct other two queries on the basis of the canonical form. The syntactic structures in the output of the major lemma query are modified in ways described in (Odijk et al., to appear). The near-miss query and the MWE query are applied to the modified output of the major lemma query.

The near-miss query searches for sentences in which the lemmas of the major words of the MWE occur in the grammatical configuration derived from the canonical form. It can find potential examples of the MWE that deviate from the canonical form provided by showing differences in forms, arguments, modification and determination. It yields a superset of the MWE query results and can be used to fine-tune the hypothesis on the MWE as encoded in the canonical form supplied by the user.

The MWE query finds sentences in which the MWE occurs. This query takes into account the hypothesis on the MWE implied by the canonical form and its annotations supplied by the user.

For the canonical form (1), applied to the MEDIARGUS treebank, the results are as follows:[7] The major lemma query yields 1309 results. The near-miss query yields 1271 results. The MWE query yields 1158 results

2.4 Analysis

Finally, there is the analysis step, which is identical to the one in GrETEL.

For a MWE, one would additionally like to analyse the result set in ways that cannot be done by GrETEL's standard analysis component. We are working on adding a special analysis step for MWEs, in which the system gathers statistics on the components of the MWE, the arguments of the MWE (their grammatical relations and syntactic categories, and their heads), the argument frames[8] that occur with the MWE, and about modifiers and determiners for the MWE as a whole and for each of its components. It does this for the results of the MWE query, for the results of the near-miss query, and for the difference between the near-miss query and the MWE query. We have an initial version available but at the time of writing it has not been integrated yet in the actual application.

But even without this dedicated analysis component MWE-Finder enables the user to quickly analyse the search results. In the results of the near-miss query one can exclude the results of the MWE query, leaving only 113 results for manual inspection. Even a cursory look shows that different determiners than *de* occur with *dans* (in particular *die*), that the determiner can be absent (but apparently only in headlines), and that the word *dans* can be modified by adjectives (e.g. *gerechtelijke* 'judicial', *fiscale* 'fiscal', *fatale* 'fatal') and by PPs (e.g. *van het faillissement* 'of the bankruptcy'). Several other results are due to wrong parses.

In the results of the major lemma query one can exclude the results of the near-miss query, which leaves 38 utterances for manual inspection. Most of these involve wrong parses, some examples involve the variant *aan de dans ontspringen*.

These result convince us to revise our hypothesis on the expression *de dans ontspringen* as implicit in the canonical form. Better canonical forms for this expression are probably *iemand zal dd:[de] *dans*

[7] The query names are links to the actual queries. All queries last retrieved 2023-03-20.

[8] With *argument frame* we mean a list of (extended relation, syntactic category) pairs for the arguments that the MWE occurs with, where an extended relation is a sequence of grammatical relations. For example, in *Marie brak Piets hart.* lit. 'Marie broke Piet's heart.', the argument frame is [(su, NP), (obj1/det, NP)], i.e., it combines with two arguments, a subject NP and a NP functioning as the determiner of the direct object.

ontspringen or *iemand zal 0de *dans ontspringen*, where *dd:[de]* means that *de* can be replaced by any definite determiner, **dans* means that the word *dans* is modifiable, and *0de* means that the word *de* is not part of the MWE. Furthermore, we found a variant of this MWE, with canonical form *iemand zal aan 0de *dans ontspringen*.

3 Limitations

MWE-Finder is fully dependent on the syntactic structures generated by the Alpino parser. If Alpino cannot parse a sentence correctly, MWE-Finder will not be able to identify any MWE in it. This is one of the reasons why MWE-Finder includes the major lemma query: this query will find sentences in which the MWE occurs even if Alpino cannot parse it correctly, so a researcher still has data to work with.[9] However, this query will also find many sentences in which the MWE does not occur, so it will require more manual work by the researcher. We aim to reduce the amount of manual work required by providing statistics on the results and the results minus the results of the other two queries in the dedicated MWE Analysis step. In particular, it will provide statistics on the grammatical relation between the lemmas of the major words. However, at the time of writing this has not been integrated in the online version yet.

4 Conclusions

MWE-Finder makes it possible to reliably and quickly search for occurrences of a MWE despite its flexible nature. We submit that MWE-Finder is a useful research instrument for linguistic and lexicological research into MWEs, and can form an exemplary research instrument in the CLARIN research infrastructure.

References

Augustinus, L., Vandeghinste, V., & Eynde, F. V. (2012). 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)*. European Language Resources Association (ELRA).

Augustinus, L., Vandeghinste, V., Schuurman, I., & Van Eynde, F. (2017). GrETEL: A tool for example-based treebank mining [DOI: http://dx.doi.org/10.5334/bbi.22. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 269–280). Ubiquity.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Sang, E. T. K., & van den Bosch, A. (2016). 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)*. European Language Resources Association (ELRA).

de Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab [DOI: http://dx.doi.org/10.5334/bbi.20. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 245–257). Ubiquity.

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., Kroon, M., Baarda, T., Bonfil, B., & Spoel, S. (to appear). 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.

Odijk, J., van der Klis, M., & Spoel, S. (2018). Extensions to the GrETEL treebank query application [http://aclweb.org/anthology/W/W17/W17-7608.pdf]. In *Proceedings of the 16th international workshop on treebanks and linguistic theories (tlt16)* (pp. 46–55).

[9] Assuming Alpino can at least lemmatize all major words correctly.

van de Camp, M., Reynaert, M., & Oostdijk, N. (2017). WhiteLab 2.0: A web interface for corpus exploitation [DOI: http://dx.doi.org/10.5334/bbi.19. License: CC-BY 4.0]. In J. Odijk & A. van Hessen (Eds.), *Clarin in the low countries* (pp. 231–243). Ubiquity.