Lemmatisation of MWEs in Dutch resources

Carole Tiberius and Lut Colman

Instituut voor de Nederlandse Taal, The Netherlands
{carole.tiberius, lut.colman}@ivdnt.org

Relevant UniDive working group: WG2

1 Introduction

In this paper we present an analysis of the lemmatisation and presentation formats of MWEs in different Dutch lexical resources. The analysis was carried out in the context of the *Woordcombinaties*[1] (Word Combinations) project, a relatively new online lexicographic resource for advanced learners of Dutch as a second or foreign language combining access to collocations, idioms, conversational routines and constructions in one tool.

In *Woordcombinaties* collocations are treated at microstructural level under the respective noun and verb lemmas of their components. They are shown per syntactic relation (e.g. subject or object) following the example of *Sketch Engine for Language Learning* (*SkeLL*)[2]. For verbs, usage patterns are annotated and encoded using a version of Patrick Hanks' Corpus Pattern Analysis (CPA) (Hanks 2004, 2013) which has been tailored to the needs of the target audience. Sentences are sorted before annotation, using a specially developed GDEX[3] configuration to enable the output of short, comprehensible and yet informative sentences. For increased readability, argument and complement slots in the patterns are represented by dummies, such as *iemand* 'someone', *iets* 'something', *ergens* 'somewhere', *zo* 'such' (or combinations thereof) where this is possible instead of by semantic types as in Hanks' Pattern Dictionary of English Verbs (PDEV). This practice was inspired by the German valency dictionary E-VALBU[4]. The dummy slots in *Woordcombinaties* are further enriched with collocations offering a kind of advanced word sketch in the patterns.

Idioms and conversational routines are also included in *Woordcombinaties*. Currently, they are encoded at microstructural level as special instances among the collocations and the patterns. However, separate access with advanced search options for idioms and conversational routines is planned and currently being designed. For instance, it will be possible to search for idioms based on image categories, such as 'body parts' and 'food' for *een vinger in de pap hebben* 'have a finger in the pie' and less specific sense categories, such as 'have a property'. Conversational routines will be linked to speech acts, such as 'greeting', 'apologising' or 'draw attention' (e.g. *luister eens* 'listen', *kijk eens* 'look').

2 Lemmatisation of MWEs

Making the more fixed MWE types accessible also at macrostructural level raises the question of their lemma form and whether to align the lemmatisation practices of MWEs with those of words.

The flexible nature of MWEs makes lemmatisation challenging and as Svensén (2009:199) notes there are no ready-made solutions in lexicography for representing the different types of variation of idioms. The number of variants shown depends on the type of dictionary. Svensén also notes that idioms must be presented in their full form and in their usual constructions, i.e. the syntactic valency of the idiom must be shown (e.g. 'look at/see sth through the rose-tinted glass').[5] However, it is also important not to include too much context, as the idiom should not appear to be more restricted contextually than it actually is. Svensén further writes that the grammatical

[1] We use the term *woordcombinaties* (word combinations) for any meaningful type of combination of words with spaces. This includes free combinations and multiword expressions, like collocations, fixed expressions, idioms and conversational routines, but also more abstract semantically motivated valency patterns.

[2] https://skell.sketchengine.eu/#home?lang=en

[3] https://www.sketchengine.eu/guide/gdex/

[4] https://grammis.ids-mannheim.de/verbvalenz

[5] Note that adding this information to the lemma form of MWEs is not in line with lemmatisation practices for words, where syntactic valency is not normally part of the lemma form.

form in which the idiom is to be presented in the dictionary depends on how frozen in form it is and that idioms are usually presented in a kind of base form.

Often, but not always, there is a dominant form that we can consider canonical, e.g. the form with the highest-frequency in the corpus. In *het paard achter de wagen/kar spannen* 'put the cart before the horse', *het paard achter de wagen spannen* is the canonical form and *het paard achter de kar spannen* is a lexical variant.

In so-called constructional idioms (Booij 2002: 302)[6], it is more difficult and often even impossible to detect a canonical form. The idiomaticity is mainly in the syntactic pattern that allows a wider but not unlimited lexicalisation than fully lexicalised idioms. An example is the construction consisting of a reflexive verb with a resultative complement (zich + RESULT + V): *zich ziek lachen* 'lit. laugh oneself sick', etc.

In the recently released DUCAME (DUtch CAnonicalised Multiword Expressions) resource (Odijk, To Appear), the canonical form of verbal MWEs is a finite sentence with a form of the future tense auxiliary verb *zullen* 'will' as its main verb (e.g. ***de laatste loodjes*** *zullen* ***het zwaarst wegen*** 'the tail end is the most difficult').[7] Special annotations are used to encode restrictions and variations (e.g. *dd:[die] vlieger zal 0niet opgaan* 'that's (simply) not on'). While this approach is suitable for more NLP oriented work, this is not a canonical form to be presented to the end user of a dictionary.

3 Lemmatisation of MWEs in Dutch resources

With this in mind, we carried out a comparative analysis of lemmatisation practices for MWEs in a number of Dutch general dictionaries and idiom dictionaries[8]. It comes as no surprise that MWEs are not treated consistently at all in the Dutch resources. Dictionaries tend to show the syntactic valency of verbal MWEs (as recommended by Svensén (2009)), but there are inconsistencies in the encoding across different dictionaries. For instance, the Dikke Van Dale Online has *iem. naar zijn hand zetten* 'force someone to one's will', whereas the Van Dale Online woordenboek hedendaags Nederlands has *iem. of iets naar zijn hand zetten* 'force someone or something to one's will'. This information is undoubtedly useful for the dictionary user, but should ideally be consistent across dictionaries.

The dictionaries include lexical variation and usually show this by giving a limited paradigm in one slot of the MWE separating the variants by means of e.g. slashes or commas. This practice can, however, lead to very complex lemma forms, e.g. *die Hand darauf/dadrauf/aus das/ein Versprechen… geben ('*to give one's hand on it/sth./on a promise…*'* (Ermakova et al. 2022:854). Furthermore, inconsistencies can be observed in the variants that are included even for one and the same expression within one dictionary as well as in the encoding of variants as separate entries or not. Differences in the positioning of the arguments in the lemma form can also be observed, e.g. *geen kaas gegeten hebben van iets* (Van Dale) and *er geen kaas van gegeten hebben* (Met zoveel woorden) 'not have a clue about something'. Syntactic variation is rarely given in the resources we consulted. Both Van Dale dictionaries, for example, mention *iem. zand in de ogen strooien* 'throw someone dust in the eyes', but not *zand in iemands ogen strooien '*throw dust in someone's eyes' or *zand in de ogen strooien van iemand* 'throw dust in the eyes of someone'.

Taking all this into account, we will define a lemmatisation strategy for MWEs in *Woordcombinaties* that is user-friendly but also compatible with more NLP oriented work. For instance, a fixed order of components will be followed as much as possible such that place and direction complements in verbal MWEs will usually occur before the verb and prepositions after it (e.g. *in de bres springen voor iemand of iets* 'throw oneself into the

[6] "Syntactic constructions with a (partially or fully) non-compositional meaning contributed by the construction, in which—unlike idioms in the traditional sense—only a subset (possibly empty) of the terminal elements is fixed."

[7] This resembles the prototypical form of verbal MWEs in PARSEME: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=home

[8] Dikke Van Dale Online (Den Boon & Hendrickx 2015), Van Dale Onlinewoordenboek hedendaags Nederlands (De Boer 2015), Algemeen Nederlands Woordenboek (ANW), Van Dale Idioomwoordenboek (de Groot 1999) en Met zoveel woorden. Gids voor trefzeker taalgebruik (Schutz & Permentier 2016).

breach for someone'). We believe that the in-depth study of Dutch phraseology from a lexicographic perspective can contribute to a cross-lingually unified lexicography of idiosyncratic constructions and harmonising lemmatisation rules (for words and MWEs).

References

Booij, G. (2002). Constructional idioms, morphology, and the Dutch lexicon. In *Journal of Germanic Linguistics*, 14(4), pp. 301-329.

Ermakova, M., A. Geyken, L. Lemnitzer, B. Roll (2022). Integration of multi-word expressions into the Digital Dictionary of German Language (DWDS). Towards a lexicographic representation of phraseological variation. In *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, pp. 851-860. Mannheim: IDS-Verlag

Hanks, P. (2004). Corpus pattern analysis. In *Proceedings of the 11th EURALEX International Congress*, pp. 87-98.

Hanks, P. (2013*). Lexical analysis: Norms and exploitations*. Cambridge, MA: The MIT Press.

Odijk, J. (To Appear) MWE-Finder: Querying for multiword expressions in large Dutch text corpora. Language Science Press.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge University Press.