Lemmatisation of MWEs in Dutch resources

Carole Tiberius, carole.tiberius@ivdnt.org | Lut Colman, lut.colman@ivdnt.org

/instituut voor de Nederlandse taal/

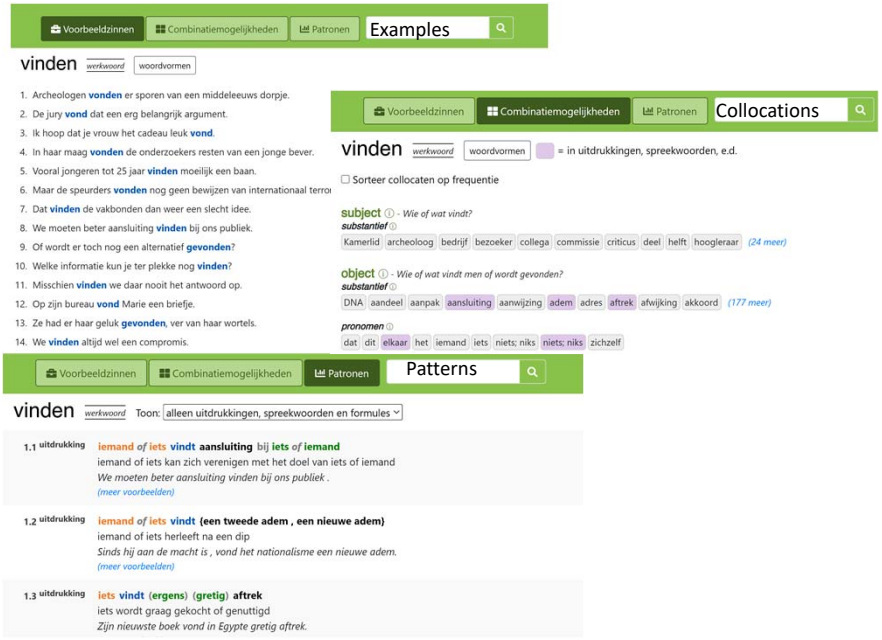
UniDive WG2

Woordcombinaties

MWEs in Woordcombinaties

Woordcombinaties is 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.

Woordcombinaties distinguishes a number of mwe types, including collocations (*een misdaad begaan* 'commit a crime'), multiword nouns, adjectives, adverbs and adpositions (*blinde muur* 'blank wall', *om en nabij* 'approximately', *baseren op* 'base on'), idioms (*de kat de bel aanbinden* 'bell the cat'), proverbs (*morgenstond heeft goud in de mond* 'the early bird catches the worm'), quotes (*ik denk, dus ik ben* 'I think, therefore I am'), slogans (*twee halen, één betalen* '~get two, pay one'), aphorisms (*gedenk te sterven* 'remember to die') and conversational and text formulas (*goede morgen* 'good morning').



Lemma forms of MWEs in Dutch resources

Arguments and open slots

iem. *naar* ***zijn*** *hand zetten* 'force **someone** to **one's** will' [1]
iem. *of* ***iets*** *naar zijn hand zetten* 'force **someone** or **something** to **one's** will' [2]
iets *naar* ***je*** *hand zetten* 'force **something** to **your** will' [5]

in weerwil van … [1,5]
in weerwil van [2]
in weerwil van ***iets*** [6]
'in spite of'

Position of arguments

geen kaas gegeten hebben ***van iets*** [1,2]
er *geen kaas van gegeten hebben* [5]
ergens *geen kaas van gegeten hebben* [6]
'not have a clue **about something**'

Lexical variation

een schat van een baby, kind, man, vrouw (entry for **schat**) [1]
een schat van een kind (entry for **een**) [1]
een schat van een [kind] [2]
'a gem of a baby, child, man, woman'
een boom van een /kerel/vent/ 'a great big fellow' [2]
als de dood van ***of*** *voor iets zijn* 'be scared to death of **or** for sth.' [1]
oude (verdronken, dode) koeien uit de sloot halen 'bring up old (drowned, dead) matters' [1]
zich naar/rot lachen & zich ziek lachen 'split one's sides laughing' & 'be in stiches' [2]
botertje ***tot*** *de boom & botertje* ***aan*** *de boom* 'situation with prosperity and welfare or excellent mutual understanding' [3]

NOTE: Syntactic variation is rarely included

Resources
[1] Dikke Van Dale Online (Den Boon & Hendrickx 2015)
[2] Van Dale Onlinewoordenboek hedendaags Nederlands (De Boer 2015),
[3] Algemeen Nederlands Woordenboek (ANW),
[4] Van Dale Idioomwoordenboek (de Groot 1999)
[5] Met zoveel woorden. Gids voor trefzeker taalgebruik (Schutz & Permentier 2016)
[6] Open Dutch WordNet (Postma, M. et al. 2016).

Canonical forms of MWEs in the lexicographic literature

There are no ready-made solutions in lexicography for representing the different types of variation of idioms. (Svensén 2009:199)

Harras and Proost's (2002:289) Citation Form Maxim:
Idioms should basically be entered in their basic or canonical form. This means that the citation form should contain only general pronouns like *someone* or *somebody* and *something*. VP-idioms should basically be entered in the infinitive form of the head verb. Where deviations from the canonical citation form are required, these should be in accordance with the following submaxims:
(1) The citation form must indicate as many restrictions as possible. […]
(2) Morphological restrictions should also be indicated by the citation form. […]
(3) The citation form should not be too restrictive. […]

Referenties
Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge, MA: The MIT Press.
Harras, G., & Proost, K. (2002). The Lemmatisation of idioms. In *Symposium on Lexicography XI, Proceedings of the Eleventh International Symposium on Lexicography*. Copenhagen. 277-291.
Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Lemmatising MWEs in Woordcombinaties

A lemmatisation strategy for MWEs that is user-friendly but also compatible with more NLP oriented work. In *Woordcombinaties* a human-friendly form is complemented with a pattern form

Preliminary guidelines:

- MWEs are entered in their canonical form, e.g. infinitive form for verbal MWEs.
- Variable, but obligatory arguments and variable parts of arguments and complements are indicated by means of dummies (e.g. *iemand de ogen openen* 'open someone's eyes') or other generic forms such as **zijn** (e.g. *zijn gezicht laten zien* 'show one's face') and **zich** (e.g. *zich op de vlakte houden* 'not commit oneself').
- A fixed order of components is followed as much as possible: e.g. place and direction complements in verbal MWEs will usually occur before the verb and fixed prepositions after it (e.g. *in de bres springen voor iemand of iets* 'throw oneself into the breach for someone or something').
- Canonical forms, variants and lexical realisations of constructional MWEs will be lemmatised separately and linked.