A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined

Lut Colman, Carole Tiberius
Dutch Language Institute, Leiden
E-mail: lut.colman@ivdnt.org, carole.tiberius@ivdnt.org

Abstract

Woordcombinaties (*Word Combinations*) is to be a new online lexicographic resource in which a Dutch collocation and idiom dictionary will be combined with a pattern dictionary. We believe that the combination of these dictionary types will be of great value to language learners and teachers. In this paper we present the three-year pilot in which we design the project and start with the description of the combinatorics of a selection of verbs for advanced learners of Dutch as a second language. We will merge a pattern dictionary of Dutch verbs, following the example of the *Pattern Dictionary of English Verbs* (*PDEV*),[1] with a collocation application, following the example of *Sketch Engine for Language Learning* (*SkeLL*).[2] In a follow-up to this pilot, more verbs and the combinatorics of nouns and adjectives will be dealt with. The long-term purpose of *Woordcombinaties* is a fully-fledged phraseological resource for Dutch.

Keywords: word combinations, collocations, idioms, proverbs, conversational routines, patterns, e-dictionary for learners of Dutch as a second language, Corpus Pattern Analysis (CPA)

1 Introduction

The importance of phraseology in second language learning and teaching is generally acknowledged (Howarth 1998; Cowie 1981, 2008; Bahns & Eldaw 1993; Wray 2000; Jesen 2006; Granger & Meunier 2008; Peters 2013). Granger and Meunier (2008) rightly propose that phraseological information should be available to learners and teachers and it should be rapidly and easily accessible. Online dictionaries are an obvious means to this end. However, up to now, Dutch dictionaries, in print and online, have many shortcomings as phraseological repositories. On the one hand, they deal with phraseology fragmentedly and rather unsystematically (Fenoulhet 1991; de Kleijn 1999, 2003; Hiligsmann 2005). On the other hand, a fully-fledged phraseological lexicographic resource for Dutch with quick and easy access to the information is still lacking. The focus in general language dictionaries, translation dictionaries and pedagogical dictionaries is mainly on idioms and proverbs, whereas the description of collocations has been given little attention (de Kleijn 1999).

We can illustrate the fragmentation with a few examples. There is a print idiom dictionary *Idioomwoordenboek* (Van Dale 1999) and a print and online collocation dictionary *Combinatiewoordenboek* (de Kleijn 2003).[3] The first one by definition focusses on idioms like *de strijdbijl begraven* (*bury the hatchet*), the latter is a useful collocation dictionary, but idioms are explicitly excluded and it is restricted to collocations of nouns with verbs. The learners' dictionary *Van Dale pocketwoordenboek: Nederlands als tweede taal (NT2)* (Verburg et al. 2017) includes verb valency patterns, like [*iemand schetst iets*] (*someone sketches something*) and includes frequently used idioms, but it does not

[1] http://pdev.org.uk

[2] https://skell.sketchengine.co.uk/

[3] https://combinatiewoordenboek.nl

include lists of collocates to fill the slots in the patterns. A systematized pattern description of about 500 verbs is provided in the Dutch – English – French *Contrastive Verb Valency Dictionary (CVVD)*,[4] but this dictionary by definition is restricted to valency and does not include lists of collocates either. The *Algemeen Nederlands Woordenboek* (*ANW*),[5] the online general language dictionary of modern Dutch, covers collocations and idioms, but there is no systematized pattern description for verbs yet, and easy and quick access to the phraseological information is still an issue.[6]

As a result of this fragmented phraseological landscape, learners of Dutch are dependent on many different resources to meet various phraseological needs, which is an impractical and undesirable learning environment. A project like *Woordcombinaties* can answer various types of phraseological queries by combining access to collocations, idioms and valency patterns in one tool. This tool can be used by second language learners in computer-assisted language learning (CALL) and data-driven language learning (DDL) for comprehension and production and by teachers as a resource to find data for combinatorics in vocabulary and grammar lessons and sentence-building exercises.

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 and idioms, but also more abstract semantically motivated valency patterns. Compounds and phrasal verbs in Dutch are written as one word. The *ANW* encodes these quite comprehensively. As both projects will be linked in the future, multiword expressions of words without spaces are accounted for. We will refer to compounds in *Woordcombinaties* only when they are synonymous with combinations with spaces, for example *slaapwel* (*good night*, *sweet dreams*) as synonym of the combination *slaap lekker*.

2 Background and Related Work

Our own and others' lexicographic experience and usage-based linguistic approaches, like lexicogrammar and construction grammar, made us acknowledge that there is no clear-cut division between lexicon and grammar, and that words get their meaning when used in context, thus in combination with other words (Firth 1957; Fillmore et al. 1988; Halliday & Matthiessen 2014; Sinclair 1991; Goldberg 1995, Gries 2013; Hanks 2013). Hence, in *Woordcombinaties* meanings will be associated with combinations of words rather than with words in isolation.

During the planning stage of the pilot we performed research on several phraseological projects to get a clear picture of the features we wanted to include. Special attention was paid to collocation applications, online valency dictionaries and online pattern dictionaries from a more semantic point of view. An exhaustive account of this research is beyond the scope of this paper, but we will briefly go into the ones that influenced our project most: *Sketch Engine for Language Learning* (*SkELL*), the *Pattern Dictionary of English Verbs* (*PDEV*), the German valency dictionary *E-VALBU* and *StringNet Navigator*.

2.1 Sketch Engine for Language Learning, SkELL[7]

Sketch Engine for Language Learning (*SkELL*) is a fully automated web interface for learners and teachers of English to search for words and phrases in corpora and find example sentences,

[4] http://www.cvvd.ugent.be

[5] http://anw.inl.nl

[6] In the future, *Woordcombinaties* will be accessible both as a stand-alone dictionary and as a plug-in resource for other applications, for example *ANW*.

[7] https://skell.sketchengine.co.uk

frequent collocates and words with similar behavior.[8] Example sentences are retrieved using GDEX. GDEX stands for "**G**ood **D**ictionary **EX**amples" and is a technology that evaluates sentences with respect to their suitability to serve as good dictionary examples. GDEX is the abbreviation for *good dictionary examples*: short and intelligible, but informative sentences elucidating the definition and exhibiting typical patterns of usage (Kilgarriff et al. 2008). Access to multiple examples of usage is useful to language learners. Frankenberg (2012 and 2014), for example, found a beneficial effect on language comprehension and production of data-driven learning through exposure to multiple good examples in experiments with Portuguese learners of English as a second language.

Word sketches in *SkELL* list collocates, which are defined as words which frequently co-occur with the searched word. The collocates are grouped according to syntactic function or another grammatical relation, for example *verbs with x as subject*, *verbs with x as object*, *modifiers of x*, *etc*., and can be part of a collocation or another type of multiword expression, like an idiom. The word sketch of the verb *bury*, for example, contains among others *dead*, *body* and *hatchet* as direct objects. The first two being part of a collocation, the latter being part of the idiom *bury the hatchet*.

The 'similar words' function shows words used in similar contexts. They are listed and visualized with a word cloud.

The main difference between *SkELL* and the *Sketch Engine* as a tool is that the first is intended as a reference and learning tool for language learners, while the latter is a language corpus management and query system for research and lexicography. *SkELL* draws data from a specially built and cleaned corpus and offers only the collocations for a selection of important grammatical relations in language learning, such as subjects, objects and modifiers.

In *Woordcombinaties* we will offer a *SkELL*-like function for Dutch with GDEX examples and word sketches. We expect it to provide a good first and overall impression of the different senses and usage patterns of the searched word and quick access to target collocates for language production. In contrast with the original *SkELL*, we will post-edit the automatically retrieved word sketches to eliminate noise. We will also add more complement types, for example, prepositional objects and clausal collocates.

2.2 Pattern Dictionary of English Verbs, PDEV[9]

The *Pattern Dictionary of English Verbs* (*PDEV*) is a corpus-driven inventory of verb patterns and their implicatures.[10] The implicature is a definition anchored to the arguments in the pattern. Each pattern-implicature pair is illustrated with an example from the British National Corpus (BNC) (Figure 1) and access to more data is provided by links to annotated concordances. Patterns are also linked to FrameNet[11] semantic frames.

PDEV is the first dictionary in which meanings are associated with usage patterns of words instead of with words in isolation (Hanks 2008). Patterns include valency structures, but they are more than that. Other syntagmatic features can also determine a pattern and its implicature (Hanks 2004).[12]

[8] *SkELL* is also available for other languages, for example, *ruSkELL* for Russian (http://ruskell.sketchengine.co.uk) and *csSkELL* for Czech (https://cskell.sketchengine.co.uk).

[9] http://pdev.org.uk

[10] Hanks borrowed the term *implicature* from H.P. Grice (1968), "to denote the act of intentionally implying a meaning that can be inferred from an utterance in context, but it is neither explicitly expressed nor logically entailed by the statement itself." (Hanks 2013: 74).

[11] https://framenet.icsi.berkeley.edu/fndrupal

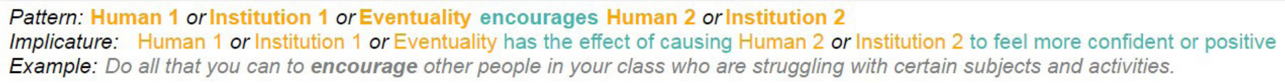
[12] Hanks (2004) illustrates this with the different meanings of *take place* vs. *take his place*, due to the absence or presence of the determiner.

Patterns are semantically motivated: lexical sets in argument slots are linked to semantic types from a shallow ontology.[13]For example, [[Human]], [[Institution]] and [[Beverage]] are the semantic types that label the lexical sets {he, woman, ...}, {school, firm, ...} and {coffee, tea, beer, ...}. The lexical sets, however, are not made explicit in the patterns. The linking of patterns with their typical lexical sets is a new feature which will be encoded in *Woordcombinaties*.

PDEV uses a specific lexicographic technique to identify the usage patterns which is called Corpus Pattern Analysis (CPA) (Hanks 2004). For each verb a sample of 250 (or more) concordance lines is analyzed and annotated with pattern numbers. The annotated concordances are grouped automatically and the lexicographer associates them with a meaning.

Both the dictionary and CPA technique are developed by Patrick Hanks within the scope of his Theory of Norms and Exploitations (TNE) (Hanks 2008, 2013). This theory distinguishes between normal or prototypical uses of words and exploitations of these, like patterns with anomalous collocates or unconventional metaphors. Corpus pattern analysis will reveal the most normal usage patterns of the verbs, which can top the pattern list in the dictionary, whereas exploitations of prototypical patterns can either be left out or put at the bottom of the list.

In *Woordcombinaties* we will describe verb patterns in a similar way, but we will tailor it to the requirements of our user group, advanced learners of Dutch. The German project *E-VALBU* inspired us for this purpose.



Pattern: **Human 1** *or* **Institution 1** *or* **Eventuality encourages Human 2** *or* **Institution 2**
Implicature: Human 1 *or* Institution 1 *or* Eventuality has the effect of causing Human 2 *or* Institution 2 to feel more confident or positive
Example: *Do all that you can to **encourage** other people in your class who are struggling with certain subjects and activities.*

Figure 1: *PDEV* pattern with semantic types.

2.3 E-VALBU[14]

E-VALBU is a semantically motivated valency dictionary for German which is developed at the Mannheim Institute for German Language. It is the electronic implementation of *VALBU*, the largest printed dictionary on German verb valency which was completed in 2004. The entry list of 638 verbs in the printed edition meets the requirements for the certificate “German as a Foreign Language” at the federal Goethe-Institut (Schneider 2008) and is integrated in the electronic version.

Verbs in *E-VALBU* are described in a maximum valency frame, which means that all elements essential to explain the meaning of the verb are encoded as complements. The verb *mieten* (rent), for example, needs five complements[15] to distinguish the meaning of the verb from that of *kaufen* (buy) (Kubczak 2014). For each verb-meaning pair *E-VALBU* codes obligatory and optional complements within the sentence structure (called *Satzbauplan*). Optional complements are put between round brackets. In *PDEV* complements in the patterns are embedded as semantic types, whereas in *E-VALBU* they are embedded as dummies, like *jemand*, *etwas*, *irgendwieviel* in *jemand mietet etwas/ jemanden für irgendwieviel von jemandem irgendwielange*, so that semantic roles are more or less implicitly recognizable. At mouseover on the dummies, semantic types become visible. In the examples section (*Beispiele*) a few corpus examples are given and in the section *Belegungsregeln* the formal grammatical categories of the complement types are listed. Each dummy in a pattern is assigned a color, which re-occurs in the complement type names in the *Satzbauplan*, the formal categories in

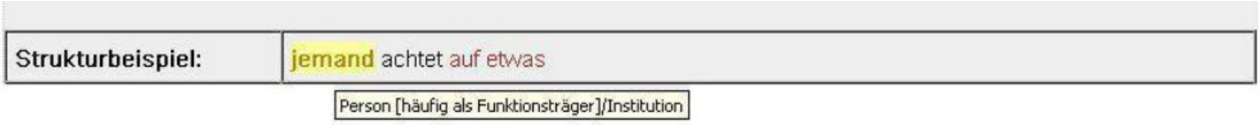
[13] http://pdev.org.uk/#onto

[14] http://hypermedia.ids-mannheim.de/evalbu

[15] With roles such as renter, tenant, rented object, rent and term of lease.

the *Belegungsregeln* and in the lexical items in the examples. This way, language learners can easily recognize the complements, even if the word order in the example differs from the pattern.

In *Woordcombinaties* we will adopt the dummy practice of *E-VALBU*. A direct embedding of semantic types in patterns may hamper readability, especially when patterns contain many semantic types. We will not use the mouseover, however, because this is impractical in phone apps. Instead, we will show semantic types and the sets of collocates on mouse click. Sentence structures which code obligatory and optional complements will be adopted from *E-VALBU* as well.



Strukturbeispiel:	jemand achtet auf etwas

Person [häufig als Funktionsträger]/Institution

Figure 2: *Achten auf* in *E-VALBU* with dummies and semantic types on mouseover.

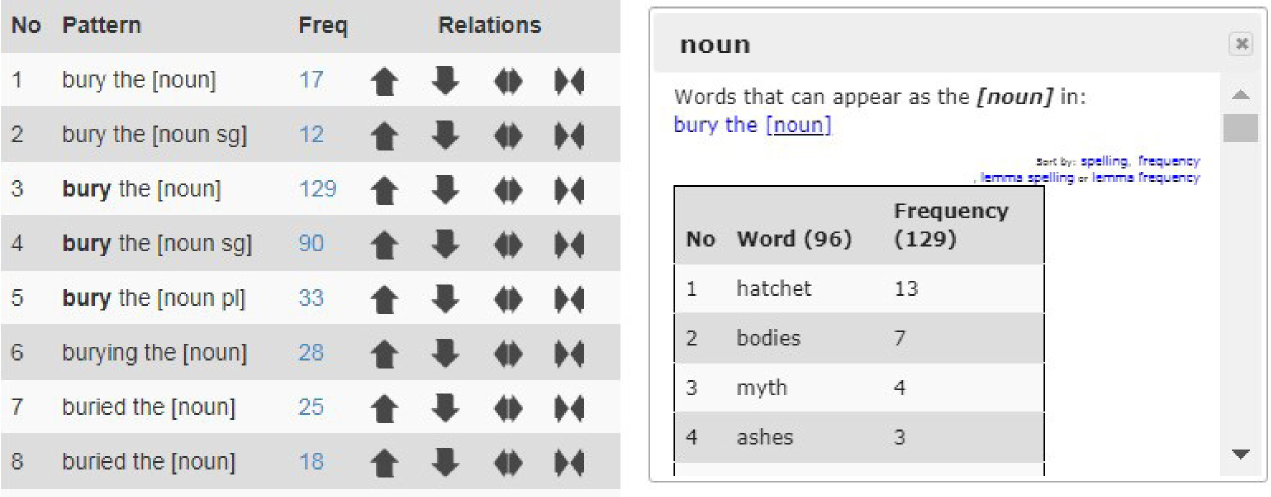
2.4 StringNet Navigator[16]

StringNet 4.0 is a corpus-derived online inventory of 1.6 billion English hybrid n-grams (Wible & Tsao 2011). *StringNet Navigator* is the user interface to navigate the resource. Hybrid n-grams, unlike traditional linear n-grams, can be any co-occurrence of POS tags, lexemes, and word forms. For example, *leave [pers pn] in no doubt [conj]*or a construction like *leave me in no doubt that*, which has word forms in the POS tag slots. This way, hybrid n-grams are not just linear strings, but they form a structured net in which one can also search for parent-child relations of constructions. An interesting application of such a structured net is that one can investigate degrees of frozenness and variation in multiword expressions. Wible and Tsao (2011) illustrate this with the string *keep a close eye on*, which, when navigating upward, reveals that *eye* in this string can be replaced by *watch*, and that *close* can be replaced by *careful*, but also that, in the construction *keep a [Adj][N] on*, the verb *keep* is the unsubstitutable lexical anchor to the expression. Tools to search hybrid n-grams will no doubt be of great relevance to constructionists and lexicographers dealing with phraseology in the future.

Another innovative aspect of the *StringNet Navigator* is that it has a collocation search in which collocations are linked to the patterns that contain that collocation. The idea to link patterns and collocations is interesting. However, navigating the tool is still a daunting task, especially for language learners. Suppose one wants to search for possible objects of the verb *bury*. If the learner would search in the *Sketch Engine for Language Learning*, he would get results in a list of possible objects, like *treasure*, *body*, *dead*, *hatchet*. In the collocation tool of *StringNet*, however, one can only search for formal grammatical categories, i.e. nouns, verbs, adjectives, etc., before or after the searched word. These categories are not assigned any complement type or semantic type, which means that the search for nouns before or after *bury* yields results like *churchyard*, *cemetery*, *grave*, *sand*, *face*, etc. These are indeed nominal collocates that frequently co-occur with *bury*, but they are not the target object collocates the language learner was looking for. To find the target collocates, he/she could start from the patterns, but then, he/she must already have an idea of all the possible phrase structures and/or word forms the object can have, like *bury the [noun]*, *bury their [noun]* or *bury the [noun pl]*, etc. In other words, the restriction to POS tags, lexemes and word forms in the hybrid n-grams has the disadvantage that one cannot search for phrasal collocates with a syntactic function linked to a semantic type or role and lexical sets, for example *bury NP* in which NP = direct object / [[physical_object]] or [[Body]] / *treasure*, *hatchet*, *dead*, *body*,....

[16] http://nav4.stringnet.org

In *Woordcombinaties* we will implement a functionality to link collocations and patterns, but the collocate sets will be assigned semantic types and will be linked to a syntactic function in the pattern.



No	Pattern	Freq	Relations			
1	bury the [noun]	17	⬆	⬇	◆	⧓
2	bury the [noun sg]	12	⬆	⬇	◆	⧓
3	**bury** the [noun]	129	⬆	⬇	◆	⧓
4	**bury** the [noun sg]	90	⬆	⬇	◆	⧓
5	**bury** the [noun pl]	33	⬆	⬇	◆	⧓
6	burying the [noun]	28	⬆	⬇	◆	⧓
7	buried the [noun]	25	⬆	⬇	◆	⧓
8	buried the [noun]	18	⬆	⬇	◆	⧓

noun

Words that can appear as the ***[noun]*** in:
bury the [noun]

Sort by: spelling, frequency, lemma spelling or lemma frequency

No	Word (96)	Frequency (129)
1	hatchet	13
2	bodies	7
3	myth	4
4	ashes	3

Figure 3: Patterns and collocates in StringNet 4.0 Navigator.

3 *Woordcombinaties* (Word Combinations)

We will now turn to the description of our own project in which we want to adopt the strengths of the projects described in the previous paragraphs while also including some innovative features which could address their weaknesses. The macrostructure of the pilot will consist of a selection of mid-frequency lexical verbs. These are selected from a vocabulary list of a remedial teaching application for academic Dutch[17] which contains a module for advanced learners of Dutch as a second language. Advanced learners of Dutch as a second language are the user group aimed at in the pilot, but in the long term we can tailor the application to any level of learners, provided that level-stratified corpora are available. As we aim at quick and easy access to phraseological information, it is necessary to deviate from a traditional layered microstructure in which this information is dispersed over a number of senses and subsenses. We offer immediate access to usage patterns in a toolbar instead (Figure 4). Demo screenshots of the combinatorics of the test verb *aanmoedigen* (encourage) will offer a preview of the web application.

3.1 Application Features and Search Options

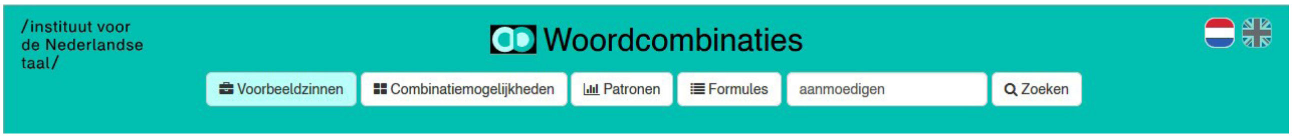


Figure 4: Toolbar of *Woordcombinaties*.

All toolbar buttons can be accessed directly in random order, but the search options are also arranged according to their increasing degree of complexity, ranging from a simple search of the verb in example sentences (*voorbeeldzinnen*), over word sketches with collocates (*combinatiemogelijkheden*), to

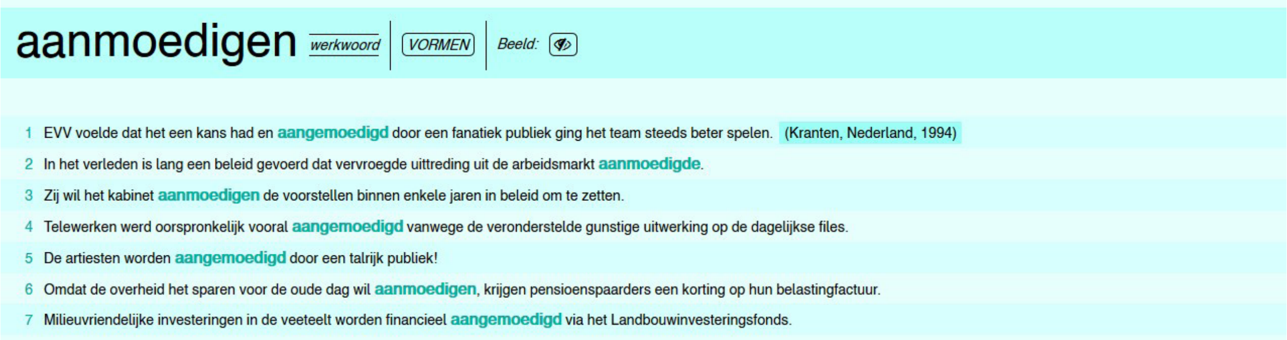
[17] http://www.hogeschooltaal.nl

pattern-meaning pairs (*patronen*) and conversational routines (*formules*), which are prepatterned situation-bound or speech act-bound utterances with a pragmatic rather than a referential meaning, e.g. apologies, greetings, asking for information. In addition to word searches, we will also offer thematic searches for idioms, proverbs and conversational routines. This option has not been implemented in the first demo screenshots, but we will include the feature in the course of the pilot.

3.1.1 Example Sentences (Voorbeeldzinnen)

Multiple example sentences provide a bird's eye view of the usage patterns and meaning potentials of the verb. As we already mentioned in discussing the *SkELL* (2.1), these can be used in data-driven learning by learners to work out the different verb senses and usage patterns. Teachers can select good examples for grammar and vocabulary lessons.

The best examples will be retrieved automatically using the GDEX functionality in the Sketch Engine. The GDEX configuration for Dutch will be set such that short and intelligible, but informative sentences elucidating the definition and exhibiting typical patterns of usage are selected. We will post-edit the examples to eliminate mistakes. Post-editing will be restricted to the correction of spelling mistakes and obvious grammar mistakes in order to maintain authenticity. The original *SkELL* provides 40 examples per verb. We will perform tests on which number of examples is sufficient and necessary for an optimal overview of a verb's behavior in Dutch. Above the examples the lemma form of the verb is given, as well as a link to a pop-up window with its conjugation forms.



1 EVV voelde dat het een kans had en aangemoedigd door een fanatiek publiek ging het team steeds beter spelen. (Kranten, Nederland, 1994)
2 In het verleden is lang een beleid gevoerd dat vervroegde uittreding uit de arbeidsmarkt aanmoedigde.
3 Zij wil het kabinet aanmoedigen de voorstellen binnen enkele jaren in beleid om te zetten.
4 Telewerken werd oorspronkelijk vooral aangemoedigd vanwege de veronderstelde gunstige uitwerking op de dagelijkse files.
5 De artiesten worden aangemoedigd door een talrijk publiek!
6 Omdat de overheid het sparen voor de oude dag wil aanmoedigen, krijgen pensioenspaarders een korting op hun belastingfactuur.
7 Milieuvriendelijke investeringen in de veeteelt worden financieel aangemoedigd via het Landbouwinvesteringsfonds.

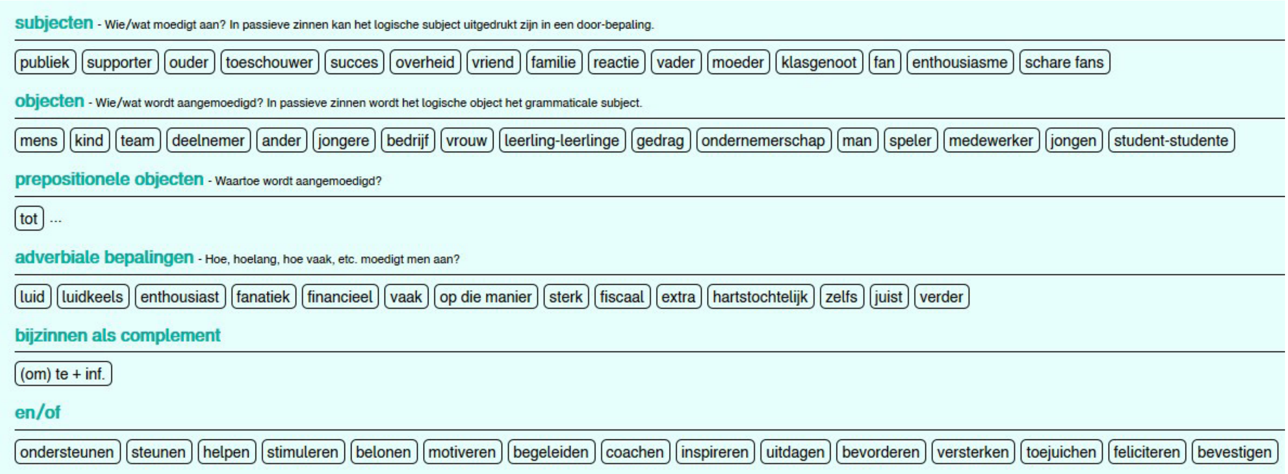
Figure 5: GDEX examples.

3.1.2 Word Sketches (Combinatiemogelijkheden)

Word sketches with collocates can be accessed through the second option on the toolbar: *combinatiemogelijkheden* (combination possibilities). We prefer this term to *word sketch*, because it is self-explanatory: the option shows which words and phrases combine with the searched verb. The functionality can support language learners in finding the right collocations to build sentences. Teachers can use the functionality to decide on which collocations to teach. We will include the following grammatical relations: subjects, direct objects, indirect objects, prepositional objects, subject complements, object complements, adverbials, co-ordination and clausal complementation. To help learners who may not be familiar with syntax terminology, the complement type names are paraphrased by means of questions, such as *who or what encourages?* or *who or what is encouraged?* A notification also mentions that the logical subject is expressed in a *by*-adverbial (*door*-bepaling) in passive sentences.

SkELL lists up to 15 collocates for each grammatical relation. We will perform tests on the number of collocates sufficient and necessary for production tasks. We will also examine various rankings

– by score, frequency or alphabetically – and the possibility for users to switch between them. The collocate lists will be post-edited. Mouse clicks on collocates result in example sentences of the collocation. It will not be possible, however, to post-edit all examples of all collocations as well. To give learners something to hold on to, we could check every first example of the collocations.



subjecten - Wie/wat moedigt aan? In passieve zinnen kan het logische subject uitgedrukt zijn in een door-bepaling.

publiek | supporter | ouder | toeschouwer | succes | overheid | vriend | familie | reactie | vader | moeder | klasgenoot | fan | enthousiasme | schare fans

objecten - Wie/wat wordt aangemoedigd? In passieve zinnen wordt het logische object het grammaticale subject.

mens | kind | team | deelnemer | ander | jongere | bedrijf | vrouw | leerling-leerlinge | gedrag | ondernemerschap | man | speler | medewerker | jongen | student-studente

prepositionele objecten - Waartoe wordt aangemoedigd?

tot ...

adverbiale bepalingen - Hoe, hoelang, hoe vaak, etc. moedigt men aan?

luid | luidkeels | enthousiast | fanatiek | financieel | vaak | op die manier | sterk | fiscaal | extra | hartstochtelijk | zelfs | juist | verder

bijzinnen als complement

(om) te + inf.

en/of

ondersteunen | steunen | helpen | stimuleren | belonen | motiveren | begeleiden | coachen | inspireren | uitdagen | bevorderen | versterken | toejuichen | feliciteren | bevestigen

Figure 6: *Combinatiemogelijkheden* (word sketch) of *aanmoedigen* (encourage).

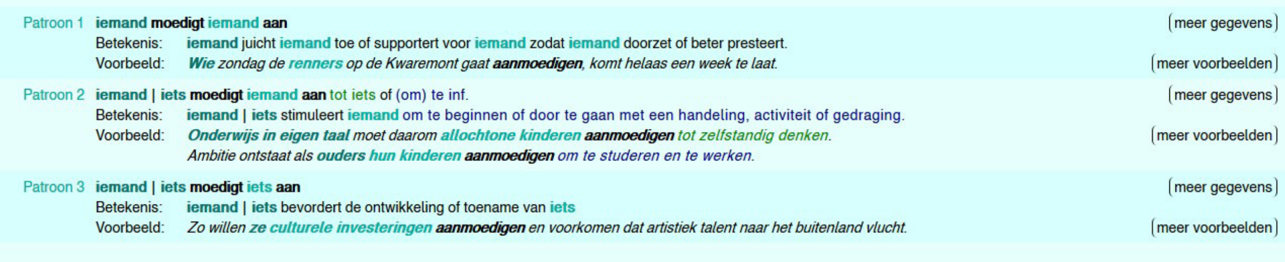
3.1.3 Patterns with Definitions (Patronen met Definities)

Examples and word sketches provide a good first impression of usage patterns and meaning. This can be very helpful for advanced learners trying to find target collocates or seeking confirmation of their intuitions regarding a collocation. However, both options have one major disadvantage: the examples and collocations are not explicitly linked to their meaning in context. As patterns and meanings often have a preference for particular sets of collocates, this information is essential and has to be encoded. The pattern functionality will enable learners to build constructions longer and more complex than the binary combinations provided by the word sketches. Patterns with their associated meanings and collocate sets will be accessible through the option *patronen* (patterns)[18]. The slots in the patterns are filled with dummies for the sake of readability. The first pattern in Figure 7, for example, is *iemand moedigt iemand aan* (*someone encourages someone*). The dummies form a limited set which will be established in the course of the pilot. Dummies are assigned colors to distinguish them from similar ones in other syntactic functions in the same pattern. The colored dummies re-occur as anchors in the definition (*Betekenis* in Figure 7) and the colors also re-occur in the corresponding lexical items in the example sentence (*Voorbeeld* in Figure 7). Clicking on the dummies will reveal collocate sets. For example, clicking on *iemand* in the subject position of pattern 1 would reveal the lexical set {*publiek* (*public*), *supporter*, *toeschouwer* (*spectator/audience*)} (Figure 8). The lexical sets will be grouped in semantic types in the database. We will use the same ontology as *PDEV* along with a Dutch version linked to it. In the example of pattern 1 the semantic types are [[Human]] / [[Mens]] and [[Human_Group]] / [[Mens_Groep]]. Dependent on the preferences of the users, we can show or hide the semantic types in the application. Links on the right of the screen give access to more examples and to more information categories, for example, the sentence structure (*Bouw* in Figure 9) with obligatory or optional complements and with information on passivisation (*Passief* in Figure 9).

Idioms and proverbs are special types of patterns with very limited lexical preferences and/or specific phrase structures in some or all of the slots. Hence, an idiom like *de strijdbijl begraven* (*bury the*

[18] From the first screenshots it is not clear yet that patterns are assigned meaning. In the final version of the application the toolbar icon '*patronen'* will be replaced by the icon '*patronen met definities*' (patterns with definitions).

hatchet) can be encoded as *iemand begraaft de strijdbijl* (*someone buries the hatchet*), but only the subject position is lexically variable whereas the object must include *strijdbijl* (*hatchet*) and the definite article *de* (*the*). Idioms and proverbs will be listed in the 'patterns with definitions' section and will be labelled. They will also be accessible through a thematic search option, for example, idioms or proverbs with animal names or body parts, idioms and proverbs for weather conditions, emotions, etc.

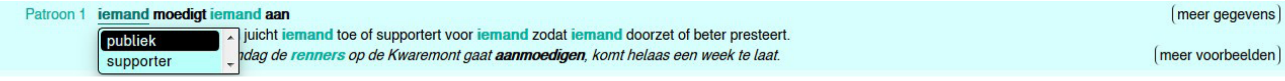


Patroon 1 **iemand moedigt iemand aan**
Betekenis: **iemand** juicht **iemand** toe of supportert voor **iemand** zodat **iemand** doorzet of beter presteert.
Voorbeeld: ***Wie*** *zondag de* ***renners*** *op de Kwaremont gaat* ***aanmoedigen****, komt helaas een week te laat.*

Patroon 2 **iemand | iets moedigt iemand aan** tot iets of (om) te inf.
Betekenis: **iemand | iets** stimuleert **iemand** om te beginnen of door te gaan met een handeling, activiteit of gedraging.
Voorbeeld: ***Onderwijs in eigen taal*** *moet daarom* ***allochtone kinderen aanmoedigen*** *tot zelfstandig denken.*
Ambitie ontstaat als ***ouders hun kinderen aanmoedigen*** *om te studeren en te werken.*

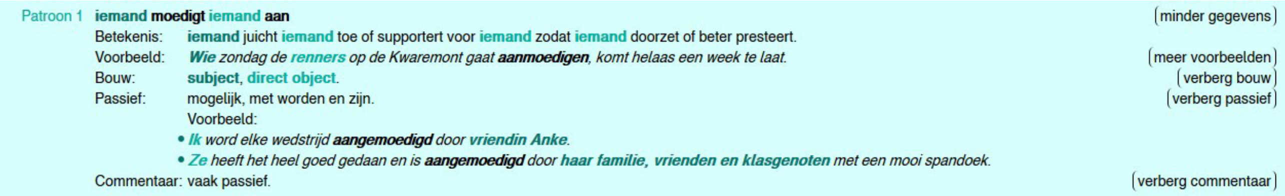
Patroon 3 **iemand | iets moedigt iets aan**
Betekenis: **iemand | iets** bevordert de ontwikkeling of toename van **iets**
Voorbeeld: *Zo willen* ***ze culturele investeringen aanmoedigen*** *en voorkomen dat artistiek talent naar het buitenland vlucht.*

Figure 7: Triples pattern – definition – example.



Patroon 1 **iemand moedigt iemand aan**
juicht **iemand** toe of supportert voor **iemand** zodat **iemand** doorzet of beter presteert.
ndag de ***renners*** *op de Kwaremont gaat* ***aanmoedigen****, komt helaas een week te laat.*

Figure 8: Lexical set on mouse click.



Patroon 1 **iemand moedigt iemand aan**
Betekenis: **iemand** juicht **iemand** toe of supportert voor **iemand** zodat **iemand** doorzet of beter presteert.
Voorbeeld: ***Wie*** *zondag de* ***renners*** *op de Kwaremont gaat* ***aanmoedigen****, komt helaas een week te laat.*
Bouw: **subject**, **direct object**.
Passief: mogelijk, met worden en zijn.
Voorbeeld:
- ***Ik*** *word elke wedstrijd* ***aangemoedigd*** *door* ***vriendin Anke****.*
- ***Ze*** *heeft het heel goed gedaan en is* ***aangemoedigd*** *door* ***haar familie, vrienden en klasgenoten*** *met een mooi spandoek.*

Commentaar: vaak passief.

Figure 9: More pattern information.

3.1.4 Conversational routines (formules)

Conversational routines, the fourth option *formules* in the toolbar, deserve special attention in a learners' application. Like idioms and proverbs, conversational routines are a special type of pattern. Coulmas defines them as "highly conventionalized prepatterned expressions whose occurrence is tied to more or less standardized communication situations" (1981, 1-3). Aijmer (2014: 2) distinguishes three major classes: formulaic speech acts, such as apologizing, thanking and greeting, for example *I'm sorry, but* ... or *Thank you!*, discourse markers, such as *As I say* ... and attitudinal routines expressing the speaker's attitude or emotion, such as *Go to hell!* Routines come natural to native speakers, but they are difficult to master for non-native language learners.

As the pragmatic function of the routines predominates over the referential meaning (Aijmer 2014: 11), it is not practicable to list them with the semantically motivated pattern-meaning pairs. Aijmer acknowledges that "the referential meaning does not completely disappear, however, but it is 'overlaid' with a pragmatic function which may be more or less dominant" (2014: 11). Moreover, verb patterns can serve as templates in "free" sentence building, but conversational routines are conventionalized and fixed to the extent that they cannot be produced by language learners by simply using the patterns as templates. One just has to know the conversational routine. Therefore, it is better to provide access to these formulae in a different way. A separate icon on the toolbar can be used for word searches, but as formulae are bound to specific communication functions and situations, it is only logical to offer a thematic search option as well. Hence, the conversational routines will also be accessible through predefined lists of speech acts and/or communication situations. Common speech acts and situations will be collected from textbooks and applications for second language learners. For example, ‘asking information’ (speech act/function) in the theme or situation ‘public transport’ would yield a formula like *Hoe laat vertrekt/gaat de trein naar x?* (*What time does the train for x depart?*). The functionality is still being developed at the time of writing this paper, so there is no screenshot to illustrate it. Possibly, thematic search options will be provided in a vertical toolbar on the left.

3.2 Corpus, Tools and Methodology

Parts of the project will be automated and parts will be manually produced. For the pilot we compiled a test corpus of about 300 million tokens which consists of newspaper material, spoken material, domain specific texts and fiction. The corpus contains material from the Netherlands and Belgium to reflect language variety in Dutch. The corpus is loaded in the Sketch Engine.[19] It is lemmatized and part-of-speech tagged and has a word sketch grammar to retrieve word sketches. We will also run tests with a parsed corpus. A parsed corpus may (or may not) open better possibilities of automatically preprocessing word sketches and patterns.

Example sentences will be automatically generated with the GDEX functionality, but they will be checked manually in the example search option (option 1 in the toolbar). Word sketches (option 2) will also be automatically retrieved in the Sketch Engine, but noise in the collocate lists will be eliminated manually. Collocates are clustered in complement types, such as subject, direct object, prepositional object, etc., which will make them easily and quickly accessible in the application for sentence building tasks.

Patterns (option 3) will be manually annotated using the CPA-technique in corpus samples of 250 concordances, or 500 or more for more polysemous verbs. The Sketch Engine supports CPA-annotation with pattern numbers in both concordance lines and word sketches. The possibility to annotate in the word sketches as well and the technique of TickBox Lexicography (TBL)[20], makes it practicable to assign relevant collocates to slots in patterns of the pattern-meaning pairs. This task is best performed manually, as collocates may occur in more than one pattern. The CPA-tool can cluster the annotated patterns automatically, which makes it possible to rank them according to frequency in the dictionary application.

For the lexicographic description of the patterns a tailored pattern editor will be developed, which will be connected with the Sketch Engine. Collocate sets will be grouped in semantic types in the database, using the *PDEV*- ontology and a linked Dutch version of it. Semantic type annotation may be useful for semantic parsing in NLP (El Maarouf et al. 2014). However, as some semantic types may be too abstract for language learners with little or no linguistic or semantic background, it is advisable to display them in the language learners’ application only on demand. In order to visualize semantic differences between syntactically identical patterns, one or two superscript lexical items can represent semantic types and serve as sense markers instead. For example, patterns like *someone checks something*$^{\text{e-mail}}$ and *someone checks something*$^{\text{oil level, tyre}}$ will immediately draw the learner’s attention to the target pattern-meaning pair in a straightforward and simple way. The default pattern view option will display elementary information: pattern, definition and example. More information categories, such as information on sentence structure, optionality of complements and passivization, will be made accessible through fold-out links. More examples will be made accessible through links to the corpus.

CPA is a corpus-driven approach which is a workable method to discover socially salient (frequent) patterns of use. However, many idioms and proverbs are cognitively salient, which means they are

[19] http://sketchengine.co.uk

[20] https://www.sketchengine.co.uk/user-guide/user-manual/tickbox-lexicography

salient, not so much in terms of frequency, but in terms of ease of recall (memorability) (Hanks 2013: 5, 344). In a strictly corpus-driven approach in which only samples of the corpus are analyzed, we may miss out on cognitively salient idioms and proverbs which are not at all that infrequent or rare. Dictionaries and lexical databases have already encoded many of them, so we advocate combining corpus-driven CPA with a corpus-based approach in which we check the currency of already encoded idioms and proverbs in our up-to-date corpus.

Conversational routines (option 4 in the toolbar) pose yet another challenge. They are mainly used in spoken language and in specific social contexts, but so far spoken language is underrepresented in Dutch corpora. Acquisition of a larger corpus of spoken Dutch or language coming close to spoken Dutch, such as subtitles of television programs, will be aimed at, but is beyond the scope of this pilot. Textbooks for language learning and online teaching materials will also be used to make an inventory of the routines used in specific situations.

4 Discussion

In the previous paragraphs we expounded upon projects which inspired us to develop *Woordcombinaties* and we offered a preview of our project and the methodology applied. One has to be aware of the fact that the pilot is an experimental lexicographic resource which leaves room for improvement.

We tried to make the application as self-explanatory as possible: on the home page all search options are clearly defined in short instruction sentences. Still, any application for computer-assisted and/or data-driven language learning requires some instruction or brief training (Boulton 2012). Supported by a user-friendly tool, learners and teachers can develop the motivation to learn and teach phraseology, and they can develop better and faster search strategies. More motivation and better search strategies can generate more fluency and proficiency in the target language.

However promising and useful we may find our project, we are well aware of a few issues which will have to be dealt with as it develops. In the first place, there is the issue of the Dutch corpora and their suitability as bases for language learning and teaching tools. A substantial portion of the present corpora is written language from newspapers, domain corpora and the web. It will be suitable enough for the target user group of advanced learners in the pilot, especially if GDEX configurations are fine-tuned to retrieve the best possible examples. However, if we want to develop more level-stratified applications for more user groups in the follow-ups of the pilot, learner corpora for Dutch will have to be structurally developed. More spoken standard language in conversations is required to disclose conversational routines and other situation-bound utterances. To disclose school language, which younger language learners have to master in education, and the combinatorics of words used in education in various topics, subcorpora of textbooks, simplified literature and easy non-fiction texts are required. A textbook corpus should contain books for language learning and textbooks for other topics in education, like history, geography, physics, and so on. Level-stratified corpora are needed to differentiate our application between levels in language learning. Many teachers and institutions provide free access to online teaching materials. In the course of the project we will try to acquire some of these, but a more structural acquisition policy is required in the long term.

Secondly, we are aware of the fact that our project is an eclectic synergy of the projects which inspired us, but we do not see this as a disadvantage. On the contrary, we took the best of all worlds by adopting the positive aspects of these projects and by providing solutions for the missing links in them. Post-editing automatically retrieved information is only one improvement. The most innovative but also the most challenging aspect of the Dutch project is the insertion of the lexical preferences and semantic types in the patterns, and the possibility to switch from collocation dictionary to pattern dictionary and vice versa. Language learners will not only have access to patterns or collocate lists separately, but will be able to see which collocates fill the slots in a pattern. The collocate lists in the word sketches provide a quick overview of target collocations, but often one also wants to know which collocates are preferred in or restricted to specific pattern-meaning pairs. Therefore, easy switching between the search options will be a useful functionality. Another important innovation is that we will pay more attention to conversational routines and thematic search options to make them easily accessible.

As for the work-flow of the project, we expect the production of the examples-section and the word sketches to be relatively easy and fast, as they will be semi-automatically produced. Pattern description will be a much more complicated and time-consuming task. Another more daunting task in due time could be a clustering of collocations according to some common Mel'čukian lexical functions as is suggested by Atkins and Rundell (2008: 151) and/or a clustering of collocations by generic semantic categories. Promising experiments in distributional semantics have been conducted with regard to this (Wanner, Ferraro, & Moreno, 2017), so we will follow developments in this field attentively.

Last but not least, with a view to reusability of existing resources, we have the intention to assimilate other phraseological resources, such as the phraseological information from *ANW* and multiword expressions in computational lexicons, such as *DuELME*[21], *RBN*[22] and *Cornetto*[23]. Information in the latter resources is not presented in a learner-friendly environment, but the contents are useful. They can be checked for currency in our corpus and adapted to the format of our application.

5 Conclusion

The pilot of *Woordcombinaties* is only the first move towards a fully-fledged phraseological resource for Dutch. We realize that such a resource is an ambitious long-term goal, but at the same time we have to acknowledge that the good match of collocations, idioms and patterns is necessary in order to support language learners and users of Dutch better in the future. Follow-ups of the pilot will be essential to prevent the project from becoming another fragment in the fragmented Dutch phraseological landscape. A first sequel to this pilot will consist of the description of more verbs and the combinatorics of nouns and adjectives. Special attention will be paid to core vocabulary, level-stratified spin-offs of the application and the optimization of Dutch corpora for this purpose.

References

Aijmer, K. (2014). *Conversational Routines in English: Convention and Creativity*. Routledge.

Algemeen Nederlands Woordenboek (*ANW*). Accessed on: http://anw.ivdnt.org [22/03/2018].

Atkins, S.B.T., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Bahns, J., Eldaw, M. (1993). Should we teach EFL students collocations? In *System*, 21 (3), pp. 101-114.

Boulton, A. (2012). What Data for Data-Driven Learning? European Association for Computer-Assisted Language Learning (EUROCALL).

Contrastive Verb Valency Dictionary (CVVD). Accessed on: http://www.cvvd.ugent.be [21/03/2018].

Cornetto Demo. *Combinatorial and Relational Network as Toolkit for Dutch Language Technology*. Accessed at: http://cornetto.clarin.inl.nl/index.html [23/03/2018].

[21] http://duelme.inl.nl

[22] http://tst.inl.nl/producten/rbn

[23] http://cornetto.clarin.inl.nl/index.html

Coulmas, F. (Ed.). (1981). *Conversational Routine: Explorations in Standarized Communication Situations and Prepatterned Speech*. Mouton.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. In *Applied Linguistics*, 2 (3), pp. 223–235.

Cowie, A. P. (2008). Phraseology. In *Practical Lexicography*, pp. 163–167.

DuELME. Dutch Electronic Lexicon of Multiword Expressions. Accessed at: http://duelme.inl.nl [23/03/2018].

El Maarouf, I., Baisa, V., Bradbury, J., & Hanks, P. (2014). Disambiguating Verbs by Collocation: Corpus Lexicography meets Natural Language Processing. *In Proceedings of LREC*, pp. 1001–1006.

E-VALBU. Das Elektronische Valenzwörterbuch Deutscher Verben. Accessed at: http://hypermedia.ids-mannheim.de/evalbu/index.html [22/03/2018].

Kleijn, P. de (1999). Nederlandse woordenboeken als basis voor een woordenboek van vaste verbindingen? In *Neerlandica Extra Muros*, 37, pp. 14-22.

Kleijn, P. de (2003). *Combinatiewoordenboek. Nederlandse substantieven met hun vaste verba*. Rozenberg Publishers, Amsterdam. Online version accessed at: https://combinatiewoordenboek.nl [21/03/2018].

Fenoulhet, J. (1991). Fraseologie en lexicografie. In *Handelingen Elfde Colloquium Neerlandicum Utrecht 1991,* Woubrugge, IVN, pp. 107-120.

Fillmore, C.J., Kay, P. & O'Connor, M.C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. In *Language* 64 (3), pp. 501-538.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis*.

Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. In *International Journal of Lexicography*, 25(3), pp. 273-296.

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. In *ReCALL*, 26(2), pp. 128-146.

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press.

Granger, S., and Meunier, F. (2008). Phraseology in language learning and teaching: Where to from here? In *Phraseology in Foreign Language Learning and Teaching*, (Amsterdam: John Benjamins Publishing Company), pp. 247–252.

Gries, S. T. (2013). 50-something years of work on collocations. In *International Journal of Corpus Linguistics*, 18 (1), pp. 137-166.

Groot, H. (1999). *Van Dale Idioomwoordenboek*. Utrecht: Van Dale Lexicografie.

Halliday, M., Matthiessen, C. M. & Matthiessen, C. (2014). *An Introduction to Functional Grammar*. Routledge.

Hanks, P. (2004). Corpus pattern analysis. In *Euralex Proceedings*, 1, pp. 87-98.

Hanks, P. (2008). Mapping meaning onto use: a Pattern Dictionary of English Verbs. *Proceedings of the AACL*.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hiligsmann, Ph. (2005). Enkele recente woordenboeken Nederlands onder de NVT-loep. In *Neerlandica extra Muros*, 43, pp. 27-38.

Hogeschooltaal. Accessed at: https://www.hogeschooltaal.nl [23/03/2018].

Howarth, P. (1998). Phraseology and second language proficiency. In *Applied Linguistics*, 19 (1), pp. 24-44.

Jesen, V. (2006). Phraseologie und Fremdsprachenlernen. Zur Problematik einer angemessenen phraseodidaktischen Umsetzung. In *Linguistik Online*, 27 (2), pp. 137- 147.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, July). GDEX: Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

Kubczak, J. (2014). Das Versteckspiel der Komplemente - wie obligatorisch sind obligatorische Komplemente und wie geht man damit in den VALBUS um. Accessed at: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-29323 [22/03/2018].

Pattern Dictionary of English Verbs (PDEV). Accessed at: http://pdev.org.uk [21/03/2018].

Peters, E. (2013). Collocaties leren in een vreemde taal. In *Handelingen der Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis*, 56, pp. 177-192.

Referentiebestand Nederlands Online (*RBN*). Accessed at: http://tst.inl.nl/producten/rbn [22/03/2018].

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford Unversity Press.

Sketch Engine for Language Learning (SkELL). Accessed at: https://skell.sketchengine.co.uk/run.cgi/skell [21/03/2018].

StringNet Navigator 4.0. Accessed at: http://nav4.stringnet.org [22/03/2018].

Verburg, M. E., Stumpel, R. J. T., & de Groot, H. (Eds.). (2017). *Van Dale pocketwoordenboek Nederlands als tweede taal (NT2)*. Utrecht: Van Dale Lexicografie.

Wanner, L., Ferraro, G., & Moreno, P. (2017). Towards distributional semantics-based classification of collocations for collocation dictionaries. In *International Journal of Lexicography*, 30(2), pp. 167-186.

Wible, D., Tsao, N. L. (2011). The StringNet lexico-grammatical knowledgebase and its applications. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 128-130. Association for Computational Linguistics.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. In *Applied Linguistics*, 21 (4), pp. 463-489.