Untangling Multiword Expressions

A study on the representation and variation of Dutch multiword expressions

Published by
LOT
Janskerkhof 13
3512 BL Utrecht
The Netherlands

Phone: +31 30 253 6006
Fax: +31 30 253 6000
e-mail: lot@uu.nl
http://www.lotschool.nl/

Cover illustration: Wordle creation by www.wordle.net

ISBN 978-94-6093-010-2
NUR 616

Untangling Multiword Expressions
A study on the representation and variation of Dutch multiword expressions

Ontwarren van meerwoordexpressies
Een studie naar de representatie en variatie van Nederlandse meerwoordexpressies

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
dinsdag 10 november 2009
des middags te 4.15 uur

door

Nicole Hendrika Wilhelmina Grégoire

geboren op 5 maart 1979 te Delft

Promotor: Prof.dr. J.E.J.M Odijk

You don’t need eyes to see, you need vision.

(Faithless, Reverence)

CONTENTS

Acknowledgements

I believe that one of the best feelings one can have is to know exactly what you want: no doubts, but motivated and disciplined to accomplish your goal. One of my I-know-what-I-want experiences came during my master's study. Small groups and motivated staff and students made it a pleasure to study Linguistics. I really enjoyed writing my master's thesis and there was no doubt about it: I wanted to do a PhD.

It soon became clear that knowing what you want is not the same as getting what you want. I applied for several positions, but there was always another candidate who had taken precisely those courses required for the position, which were never courses on computational linguistics. A year after my graduation I knew again why I worked so hard for my master's thesis; my supervisor offered me a research position. Initially for two years, but with his help and effort it turned into a full PhD position, which was exactly what I wanted.

I am very grateful to Jan Odijk for his confidence and for giving me the opportunity to explore the scientific world. I highly appreciate the many hours he put into guiding me and into reading, rereading and rerereading my work. Thank you for being my promotor, co-promotor, daily supervisor, English teacher, etc.

During the first two years I was working within the STEVIN project *Identification and Representation of Multiword Expressions* in collaboration with Begoña Villada Moirón of the University of Groningen. I would like to thank Begoña for introducing me to many people in the field and for taking me in tow at the conferences we attended. Many thanks to Gertjan van Noord of the University of Groningen for answering all my questions about Alpino and for providing me with the relevant data for

the second part of my research.

I really enjoyed working at UiL-OTS and being part of Utrecht University. I am grateful to Paola Monachesi, Eline Westerhout, Rick Nouwen, Marjo van Koppen and Martin Everaert with whom I could talk about my research or who have read and commented on parts of my dissertation. A special thanks goes to Huib Kranendonk for being a pleasant office mate and sparring partner.

As a member and chair of the PhD network of Utrecht University (PrOUt) I have met many fellow PhD candidates. I enjoyed the monthly meetings with my PrOUt colleagues and the fruitful discussions on PhD policy. A special thanks goes to PrOUt colleagues Martijn Duvoort, Floor Kroese and Richard van der Put for helping me to revive PrOUt.

Yes, besides all the hard work, there was also a private life. I would like to thank my friends for their support, the good talks and of course the "witte wijn in een koeler". Especially Amy, my best friend and personal coach: thank you for being there.

Thanks to my in-laws and a big hug for my niece and nephew, Aimee and Julian.

I am grateful to my parents Jan and Aga for their constant love and support. Thanks to my brother Paul for beating me at golf and tennis. You just set me a new goal.

I would like to end with a special thanks to the most important person in my life, my treasure. Vincent, thank you for your endless love, support and encouragement. I know that you are holding your breath for my next I-know-what-I-want experience, but I promise you: no more hermit behaviour.

CHAPTER 1

GENERAL INTRODUCTION

It is the regularity of natural language that makes a language understandable and learnable. It is its (apparent) irregularity and unpredictability that keeps many linguists (and language learners) busy. Regularities in language can be expressed by grammatical rules, e.g. the past tense of the majority of English verbs can be formed by a rule that suffixes *-ed* to the stem. There are, however, a number of English verbs that do not obey this rule, e.g. the past tense of *eat* is *ate* and not *eated*, and are therefore called irregular.

Regularity and irregularity can be found at both the word and the phrase level. There are rules that alter word forms to create new ones and, as shown in the past tense example, there are exceptions to these rules. There are syntactic rules that combine words into grammatically correct phrases and sentences, and although there are no rules that assign meaning to words, there is a rule (or principle) that says that the meaning of a phrase or sentence can be derived from the meaning of its parts and the way they are combined. But also at the phrase level there are exceptions to these rules, yielding linguistically idiosyncratic combinations. It is this type of irregularity, i.e. combinations of words with idiosyncratic and unpredictable properties, that I am interested in and that is the subject of this study.

Multiword expressions, *multiword units*, *extended lexical units*, *complex units*, *idiomatic expressions*, *idiomatic phrases*, *idiomatic combinations*, *idioms*, *collocations*, *fixed phrases*, *phrasemes*, *phrasal lexical items*, *frozen sentences*: just a selection of terms found in the literature that are used to refer to combinations of words that show some sort of idiosyncrasy.

Some terms cover a wider range of types of combinations than others, although this often depends on how the term is defined. Many terms lack a common definition. Take for instance the term *idiom*, of which various definitions can be found in the literature:

Weinreich (1967) "[...] any expression in which at least one constituent is polysemous, and in which a selection of a subsense is determined by the verbal context, [is called] a *phraseological unit*. A phraseological unit that involves at least two polysemous constituents, and in which there is a reciprocal contextual selection of subsenses, will be called an *idiom*." (p. 42)

Fraser (1970) "I shall regard an idiom as a constituent or series of constituents for which the semantic interpretation is not a compositional function of the formatives of which it is composed." (p. 22)

Schenk (1994) "Idioms are expressions for which a literal interpretation does not yield the correct meaning of the idiomatic expression." (p. 2)

Mel'čuk (1995) "An idiom is a multilexemic expression **E** whose meaning cannot be deduced by the general rules of the language in question from the meaning of the constituent lexemes of **E**, their semantically loaded morphological characteristics (if any) and their syntactic configuration." (p. 167)

O'Grady (1998) "I assume that idioms have a meaning that is not a simple function of the literal (i.e., non-figurative) meaning of their parts and that they manifest a high degree of conventionality in the choice of component lexical items." (p. 279)

Riehemann (2001) "I use the term 'idiom' to refer to an expression made up out of two or more words, at least one of which does not have any of the meanings it can have outside of the expression. As will become clear from the discussion below, this is not intended as an exact definition." (p. 2)

It is beyond the scope of this dissertation and beyond its purpose to come up with appropriate terms and definitions for all types of multiword combinations explored in this study. Instead I will focus on mul-

tiword expressions in general and, moreover, classify them not on the basis of predefined labels but on the basis of their individual properties.

In this study, the term *multiword expression* is used as defined in (1).

(1) A multiword expression (MWE) is a combination of words that has linguistic properties not predictable from the individual components or the normal way they are combined.

MWEs show linguistic idiosycrasies at different levels of analysis: specific lexical items must be used and cannot be replaced by synonyms or near-synonyms without changing the meaning or the well-formedness of the expression; individual components may be restricted in their phonological, morphological and syntactic flexibility; and their meaning is often not straightforwardly compositional, i.e. the meaning of an expression cannot always be deduced from the meaning of the individual components and the way they are combined. Examples of MWEs are given in (2) - (5).[1]

(2) a. blunder maken/begaan/*doen/*slaan
mistake make/commit/*do/*hit
'make a mistake'

b. flater *maken/*begaan/*doen/slaan
mistake *make/*commit/*do/hit
'make a mistake'

(3) het loodje/*lood leggen
the piece-of-lead-DIM/*piece-of-lead lay
id. 'to die'[2]

(4) in opdracht van
in order of
'by order of'

[1]In this dissertation, I use the following typographical conventions: An asterisk (*) is used to indicate ill-formedness in both the literal interpretation and the non-literal interpretation, and a dash (#) is used to indicate ill-formedness in only the non-literal interpretation.

[2]The abbreviation *id.* stands for idiomatic meaning, i.e. the non-literal meaning of the MWE, whereas *lit.* stands for literal meaning.

(5) met de handen in het haar zitten
with the hands in the hair sit
id. 'to be at loss what to do'

In all four examples specific lexical items must be used. The choice of the specific items is unpredictable and cannot be derived from the semantic properties of the individual items; although *blunder* and *flater* in (2) are synonyms and share the same semantic properties, the verbs they can co-occur with are different and the choice of the verb is unpredictable, which makes the combinations MWEs. An example of morphological inflexibility is (3), where the noun *lood* must be in the diminutive form (DIM) to form an MWE with *leggen*. The noun *opdracht* in (4) is a singular count noun that should in general be preceded by a determiner according to standard Dutch grammar. Because of the lack of a determiner preceding the noun, the combination is syntactically idiosyncratic and hence an MWE.

Because of their idiosyncratic properties and the unpredictable relation between form and meaning, but also because MWEs make up a large part of natural language, there is a long tradition in the study of MWEs. Linguists from various disciplines are interested in the concept of MWE, see e.g. Everaert et al. (1995) and Granger and Meunier (2008). In this dissertation, MWEs will be studied from multiple perspectives. As a lexicographer, I will analyze over 9,000 word combinations automatically extracted from corpora and proposed as being (part of) an MWE by the identification methods applied. As described in the next chapter, the annotation process includes selecting true MWEs according to the definition given in (1) and deciding on their precise form. Although the concept of MWE has been clearly defined, it is not always easy to determine whether a combination is an MWE. Moreover, even if a combination is classified as an MWE, it can be a problem to determine its correct form: What are the obligatory and what are the optional components? Is the determiner fixed or can it vary? Are there any restrictions on the morpho-syntactic flexibility of the individual items? etc.

The result of the selection procedure is a list of over 5,000 MWEs that can be used for different purposes. Although the list of MWEs would be suited for inclusion in a dictionary focusing on humans users, the goal of this work is to create a dictionary for natural language pro-

cessing (NLP) technology. Dictionary makers who focus on human users usually rely on the user's knowledge of the language and general knowledge and represent information that is not common knowledge, such as meaning and usage information, in an informal and implicit way. An NLP system, however, does not know anything unless it is explicitly encoded. Therefore all information, such as the assumptions about the general rules of the system, must be represented in a formal and explicit way. For instance, to form the past tense of English regular verbs, the grammar of an NLP system should contain a rule that suffixes *-ed* to the stem of the verb to form its past tense, but to avoid ungrammatical past tense forms, the past tense of irregular verbs, such as *eat*, should be explicitly listed. The same holds for NLP systems to adequately deal with large numbers of MWEs; in order to do so they should contain (1) an adequate method for handling various types of MWEs in the grammar, and (2) a large number of lexical entries for MWEs compatible with the grammar.

Various methods have been developed to adapt existing grammatical frameworks and implementations so that they can handle (types of) MWEs. Examples include the incorporation of various types of expressions in the compositional grammars of the Rosetta approach, see (Schenk, 1994), (Rosetta, 1994); in the Tree Adjoining Grammar (TAG) framework, see (Abeillé and Schabes, 1989), (Abeillé, 1995); in X-bar theory, e.g. (van Gestel, 1995); and in Head-Driven Phrase Structure Grammar (HPSG), e.g. (Riehemann, 1997), (Riehemann, 2001), (Richter and Sailer, 2002), (Sailer, 2003). Research on this topic is still going on. The current study, however, does not concentrate on one specific framework or implementation. Instead it focuses on making lexical descriptions of over 5,000 Dutch MWEs available in such a way that they can be used in a wide variety of different grammatical frameworks, approaches to MWEs, and their implementations.

This study builds on previous work done by Odijk (2003), see also (Odijk, 2004b), (Odijk, 2004a), (Odijk, 2005). Odijk (2003) argues that the lexical description of an MWE must have the same properties as a simple lexical item and furthermore the following elements: (1) a syntactic structure of the MWE, (2) a unique identification of the MWE components and (3) a listing of the MWE components in an order that is compatible with the syntactic structure. These conclusions are based on the treatments of MWEs in the Rosetta machine translation (MT)

system (Rosetta, 1994), but other approaches, such as the TAG framework and HPSG, require these elements as well.

No de facto standard for representing these aspects currently exists. The main criticism against proposed standards for the representation of the syntactic structure of (types of) MWEs is that the structures assigned are highly theory- and implementation-dependent and often too complex to create and maintain. A standard lexical representation of MWEs should be as theory- and implementation-neutral as possible, technically simple and moreover, it should be convertable to any system specific representation with a minimal amount of manual effort.

Avoiding the problem of theory-neutrality and too complex structures, while keeping the focus on reusability, Odijk (2003) proposes a standard that does not prescribe the structure of an MWE, but that requires that it is specified which MWEs have the same structure. The idea is that MWEs that have the same structure require the same treatment in an NLP system. MWEs with the same structure form so-called equivalence classes, hence the name of the proposed standard: the Equivalence Class Method (ECM).

The ECM, including the accompanying standardized lexical representation for MWEs, constitutes the starting point of the creation of a Dutch Electronic Lexicon of Multiword Expressions (DuELME). Describing and discussing the development of DuELME is one of the objectives of this dissertation. The development process can be divided into four parts: (1) selecting MWEs to include in DuELME by analysing over 9,000 word combinations automatically extracted from corpora; (2) refining the original ECM and (3) extending the standard lexical representation for MWEs, both to increase the successfulness of ECM; and (4) designing and implementing the DuELME database structure.

The enhanced lexical representation does not include the encoding of restrictions on the use of MWEs; MWEs often seem to be restricted in their use, i.e. they do not always allow all the combinations and modifications that a literal use would. It is unclear whether restrictions on their use can be predicted on the basis of their properties. To improve successful treatment of MWEs in NLP, a better understanding of potential MWE variation is necessary. A corpus-based study on the variation potential of MWEs will be carried out in Part II of this dissertation. The focus will be in particular on OBJ1-V idioms, i.e. direct-object verb (OBJ1-V) combinations of which at least the noun is not used literally.

Central in this study is the Idiom Variation Potential Hypothesis, which postulates that the presence of certain properties of the idiom parts are responsible for their in principle unlimited variation potential. The hypothesis will be tested using the corpus data of 25 OBJ1-V idioms taken from DuELME as the primary empirical material.

Outline of this dissertation

This dissertation is structured in the following way. Besides the general introduction and a general conclusion, there are two parts. Part I describes the design, implementation and population of DuELME, and Part II studies the variation potential of idioms. Both parts start with an introduction addressing related topics, such as related literature, and end with a conclusion in which the main points will be summarized and the approach taken discussed.

Part I

A Dutch Electronic Lexicon of Multiword Expressions

CHAPTER 2

INTRODUCTION

Successful treatment of a large number of MWEs is still a major problem for NLP technologies (Sag et al., 2001). This problem can be overcome by having (1) an adequate method of handling each type of MWE in the grammar of an NLP system, and (2) the availability of a large number of lexical entries for MWEs compatible with the requirements of the grammar (Odijk, 2004b). As stated in the general introduction, various methods have been developed to adapt existing grammatical frameworks and implementations so that they can handle MWEs, and although research in this area is still going on, the current study focuses on the second requirement.

The creation of lexical descriptions for MWEs is very time consuming and they can hardly be reused when designed for one specific framework or implementation. It is more efficient to make a large number of lexical descriptions of MWEs available that both are as theory- and implementation-neutral as possible and at the same time rich enough to be compatible with a wide variety of different grammars. With this in mind we have developed a Dutch Electronic Lexicon of Multiword Expressions (DuELME), and the first part of this dissertation describes its development, i.e. its design, implementation and population.

DuELME[1] is a resource that contains a set of data, in this case Dutch

[1]DuELME is one of the results of the project Identification and Representation of Multiword Expressions (IRME), which has been carried out within the STEVIN programme funded by the Dutch and Flemish Governments. The current version of DuELME is v1.1 and is available through the *TST-centrale* (HLT Agency, http://www.tst.inl.nl/).

MWEs, that are represented electronically, hence the name electronic lexicon. DuELME has been created out of a need for a large number of lexical descriptions of Dutch MWEs organized in such a way that they can be used in a wide variety of different grammatical frameworks, approaches to MWEs and their implementations. The result is a resource that contains lexical descriptions of over 5,000 MWEs, and that can be integrated into Dutch NLP systems with a minimal amount of manual effort.

Data for DuELME have been extracted from corpora, which has resulted in a list of over 9,000 so-called candidate expressions, i.e. combinations of two or more words that may form an MWE or may be part of an MWE. Given this list, the development process of DuELME includes:

1. manually selecting true MWEs for inclusion in DuELME and deciding on their precise form;

2. designing a standard lexical representation for MWEs which is as theory- and implementation-independent as possible, and that comprises at least the core properties needed to convert the standard representation into system specific representations as efficiently as possible; and

3. implementing the lexical database so that views on the data can be provided in a flexible way.

The design of DuELME is based on the Equivalence Class Method (ECM) introduced by Odijk (2003). The ECM is an innovative method that groups MWEs according to their syntactic pattern to form so-called Equivalence Classes (ECs). Given the idea that MWEs that have the same structure require the same treatment in NLP, it takes just one instance of an EC that needs to be converted manually to a system specific representation, while all other instances of the same EC can be converted fully automatically.

The proposed ECM includes a suggestion for a standardized lexical representation for MWEs. The design part of DuELME's development involves refining the original ECM and extending the proposed lexical representation for MWEs to increase the successfulness of converting the standard representation into representations required by a specific system.

In the remainder of this chapter I will first give an overview of related literature on the representation of MWEs and MWE resources. Section 2.2 introduces the research approach and discusses its innovativeness with respect to related approaches. Section 2.3 elaborates on the proposed ECM. This chapter ends with an outline of the remainder of Part I.

2.1 RELATED RESEARCH

A central question is what the structural design of a lexical entry must be. In the past, various suggestions have been made, for instance to treat MWEs as strings, i.e. fixed sequences of words (see e.g. Chomsky (1981)). Such an approach might work for fixed expressions such as *ad hoc*, but encounters a lot of problems when dealing with more flexible expressions. Even the smallest inflectional variation, e.g. tense of verbal expressions, cannot be accounted for easily by such an approach.

Dictionaries designed for human users usually represent MWEs as simple strings as it is assumed that lexico-grammatical information is generally known by the user. Where such a representation is usually sufficient for the purpose of such dictionaries, a more sophisticated representation is needed for an adequate treatment of MWEs in NLP.

Besides general dictionaries that include a selection of MWEs (which may not always be actually marked as such), there are two Dutch MWE resources, viz. *Idioomwoordenboek* ('Idiom dictionary'), which is a commercial printed dictionary (de Groot et al., 1999), and an electronic resource called the *Referentiebestand Nederlands* (RBN, 'Reference Database of The Dutch Language') (Martin and Maks, 2005). Both resources are primarily meant for human users, although the entries in the RBN have been represented in a more formal way than usually done in dictionaries that are intended for human users.

No exhaustive research on creating a Dutch MWE resource that can be used to generate system specific MWE lexicons has been done before. Work on the representation of MWEs and the creation of MWE resources has been carried out for other languages. This section provides an overview of related research. The overview is certainly not exhaustive, but does reflect the variety of approaches to the development of MWE resources.

Gross (1986, 1996) discusses the classification of various types of MWEs in the lexicon-grammar framework. In this framework MWEs are classified according to their structure, which is described as the part-of-speech of the head of the MWE and of each of its essential components. For example, the expression *s.o. took the bull by the horns* is represented as N_0 V C_1 Prep C_2 (where *C* stands for fixed noun phrase). Each class, containing instances with the same structure, is represented as a so-called matrix, where each row is an entry and each column defines a particular transformational property, e.g. passivization, or distributional property, e.g human or non-human. Whether or not this property applies to the entry is notated with a plus or minus sign respectively. Subclasses can be defined on the basis of these properties.

The lexicon-grammar covers 25,000 frozen verb phrases, 50,000 constructions containing a support verb and a predicative noun, and 7,000 frozen adverbial phrases (see also Laporte and Voyatzi (2008)).[2]

The lexicon-grammar taxonomic approach has been adopted by Baptista et al. (2004) and Català and Baptista (2007). The former present research on building an electronic dictionary of over 4,000 so-called frozen sentences of European Portugese, whereas the latter discuss the construction of an electronic resource describing over 4,000 Spanish adverbial frozen expressions.

Dormeyer et al. (1998) report on work on an electronic dictionary for German verbal idioms, called *Phraseo-Lex*. *Phraseo-Lex* has been designed both as a dictionary for humans and as a source to generate lexicons for NLP systems. The representation of idioms in the database is very advanced, i.e. it not only includes the idiom's lemma and base lexemes, but it also contains a syntactic, a semantic and a pragmatic level of description. The implementation of a mapping between a part of the *Phraseo-Lex* database and an NLP system has been described in Dormeyer and Fischer (1998).

Krenn (2000a,b) has implemented a relational database for lexical collocations, called CDB (Collocation Database). She uses the term *collocation* for "word combinations that are lexically determined and constitute particular syntactic dependencies such as verb-object, verb-subject,

[2]http://infolingu.univ-mlv.fr/

adjective-noun relations, etc." (Krenn, 2000a, p. 1). The database contains more than 1,000 German PP-verb collocations.

In her approach, the representation of a collocation consists of a *competence base* and an *example base*. In the competence base, collocation instances (CI) and CI-analyses are stored. CIs are representations of the major lexical elements of a collocation, in which nouns are represented as full forms and verbs as infinitives. Each CI is related to a collocation type, either support verb construction, figurative expression, or pseudo-collocation. The CI-analysis lists the values of eight attributes for each collocation, among which causativity, Aktionart, the syntactic arguments required by the collocation, modification of the noun, and modification of the whole expression. The example base contains example sentences extracted from corpora. Each corpus example is associated with exactly one CI.

The development of CDB came to a halt before it had been made available. Some demo data are still accessible.

Kuiper et al. (2003) have created a Syntactically Annotated Idiom Database (SAID). SAID contains 13,467 phrasal lexical items (PLIs). The data come from four dictionaries of English Idioms. The syntactic analysis has been conducted manually and is based on a pre-barriers generative framework. SAID is available via the Linguistic Data Consortium.[3]

Villavicencio et al. (2004) propose a possible architecture for the lexical encoding of MWEs built on ideas presented by Copestake et al. (2002). Villavicencio et al. (2004) analyse two different types of expressions, viz. idioms and verb-particle constructions. The central idea behind the design of the encodings is to minimize the amount of information that needs to be specified for each entry by maximising the information that can be inherited from simplex verbs. For each simplex verb, its orthography and its syntactic and semantic type are stored in the database.

Villavicencio et al. subdivide the class of idioms into fixed idioms and flexible idioms, based on the semantic decomposability of an idiom, i.e. an idiom is semantically decomposable, and said to be flexi-

[3]http://www.ldc.upenn.edu/

ble, if "a meaning can be assigned to individual words (even if some of them are non-standard meanings) from where the meaning of the idiom can be compositionally constructed." (Villavicencio et al., 2004, p. 2)

Fixed idioms are treated as strings ('words with spaces') and encoded as simplex entries. Elements that can inflect, e.g. *kick* in *kick the bucket*, are marked as such.

The encoding of flexible idioms is dealt with in three stages. First the idiomatic components of the idiom are defined in the same way as simplex verbs. In addition, each component is linked to a non-idiomatic simplex entry from which they obtain by default many of the characteristics. Furthermore a non-idiomatic paraphrase for the idiomatic element is defined. In the second stage, all the components that are part of the idiom are listed. This is done to ensure that the idiomatic reading of the individual components is only used when it occurs in the presence of the other components. Meta-types, in the sense of predefined syntactic and semantic relations, are specified in the last stage. Examples of meta-types specified are *verb-object-idiom* and *verb-particle-np*. The generality of the meta-types has been tested using a sample of 25 randomly selected idioms (out of a larger sample of 100 idioms that was used to determine the requirements of the standard encoding). The majority of the 25 idioms could be described by the meta-types defined, yielding a classified list of idiom entries.

Fellbaum et al. (2006) discuss the motivation as well as the design and development of a large lexical resource focusing on German verb phrase idioms and light verbs. Lexical annotation of the idioms is based on corpus-based investigation. Both the annotations and the corpus data are accessible.[4] The properties of an idiom are recorded in eight so-called data sheets. The idiom description contains, inter alia, the citation form of the idiom, some corpus examples, a paraphrase of the idiom, a dependency structure, the syntactic transformations found in the example corpus, and semantic properties. A total of 1,000 expressions have been investigated in this project.

[4] http://kollokationen.bbaw.de/

2.2 THE RESEARCH APPROACH

The goal of this work is to make available a resource that is organized and describes MWEs in such a way that it can be integrated into a wide variety of NLP systems as efficiently as possible. To achieve this goal, the approach taken concentrates on representing the core properties needed and on organizing the MWEs according to their syntactic pattern. The result is a resource that comprises both pattern descriptions describing the characteristics of a group of MWEs and descriptions of individual expressions. Although the approach is in line with some of the projects addressed in the previous section, it is also distinctive and innovative.

First of all, the approach taken builds on the Equivalence Class Method (ECM) (Odijk, 2003). The ECM is based on the idea that MWEs that have the same structure require the same treatment in NLP, and therefore classifies MWEs according to their syntactic pattern. A classification of expressions has also been suggested by Villavicencio and Copestake (2002). They propose two types of classifications for idioms, viz. a classification based on syntactic patterns and a classification based on semantic decomposability (cf. Sag et al. (2001)). A random selection of 43 VP idioms shows 20 different syntactic patterns. From the 43 VP idioms, 33 idioms are decomposable and the remaining idioms are non-decomposable. Since one of the requirements of DuELME is that it should be organized in such a way that its integration into Dutch NLP systems can be done with a minimal amount of manual effort, a classification that comprises just two classes will hardly decrease the amount of manual work as compared to no classification. A classification based on the syntactic pattern of MWEs suits the requirement better, provided that the annotation of the pattern is as theory- and implementation-neutral as possible.

In most approaches addressed, some kind of syntactic analysis is assigned to individual expressions. The most sophisticated syntactic analysis is done in the SAID-database Kuiper et al. (2003), but just like Fellbaum et al. (2006) who provide a dependency structure for each expression, the structures have been assigned to individual MWEs and not with the intention of grouping the expressions accordingly. In the lexicon-grammar as proposed by Gross (1996), classes are formed on the basis of, inter alia, the parts-of-speech of the individual compo-

nents that make up an MWE. Although, the lexicon-grammar could be taken as the starting point of DuELME, the approach is very specific for the lexicon-grammar taxonomy lacking theory-independence. We have chosen to follow the ECM, mainly because the ECM offers the possibility to describe any group of MWEs theory-independently, i.e. it allows one to abstract away from specific syntactic structures. Moreover, the ECM comes with a standard representation for MWEs and with an incorporation method which has been illustrated by converting the standard representation to the representation as required in the Rosetta MT system (Odijk, 2004b).

The approach taken does not solely focus on one type of MWEs, but on MWEs in general. Moreover, the selection of the lexical entries and their properties is corpus-based. Both Fellbaum et al. (2006) and Krenn (2000a,b) support their annotation of lexical entries using corpora as empirical material, with the difference that the PP-verb collocations in Krenn (2000a,b) have been selected from a set of automatically extracted data.

Except for the SAID-database and the various lexicon-grammars, the resources described contain no more than 1,000 highly frequent expressions. DuELME contains over 5,000 unique expressions that have been selected manually from data automatically extracted from corpora.

The approach taken is furthermore distinctive, because the resource created is intended for use in NLP systems. To that end a conversion to the Dutch NLP system Alpino[5] has been tested in theory and in practice (see Section 5.2). Moreover, a conversion to the Rosetta MT system (Rosetta, 1994) has been theoretically examined. One of the goals of Copestake et al. (2002) and Villavicencio et al. (2004) was also to build a resource that can be used in multiple frameworks. They have proposed a design which has been applied to a small set of expressions but which has not been evaluated.

To conclude, DuELME is a unique resource optimized for reusability. Its design is described in detail in Chapter 4. The next section introduces the principles of the original ECM, including the proposed standardized lexical representation for MWEs.

[5]http://www.let.rug.nl/vannoord/alp/

2.3 THE EQUIVALENCE CLASS METHOD

2.3.1 THE PROPOSED STANDARD

The approach taken builds on the Equivalence Class Method developed by Odijk (2003). The ECM is based on the idea that MWEs that have the same syntactic pattern require the same treatment in NLP. So instead of assigning a syntactic structure to individual MWEs, the method specifies which MWEs have the same structure.

The original ECM contains a proposal for a standard lexical representation of MWEs and an incorporation method. The proposal states that an MWE description should consist of:

1. an MWE pattern name: an identifier that uniquely identifies the structure of the MWE;
2. a list of MWE components (Component List: CL); and
3. an example sentence that contains the MWE.

Equivalence classes are defined with the help of the MWE patterns, i.e. MWEs with the same MWE pattern name belong to the same equivalence class (EC). The CL takes the form of a sequence of strings, where each string represents the lexicon citation form of an MWE component. The order of the sequence is free, but the standard requires that the same order is used for each MWE in the same equivalence class. As for the example sentence, the standard requires that the structure should be identical for each example sentence within the same equivalence class.

Besides the MWE description, there must be a description of the MWE patterns. An MWE pattern description consists of two parts:

1. an MWE pattern name; and
2. comments, i.e. free text in which the uniqueness of the MWE pattern is described. The information in this field is meant for human users and not to be interpreted automatically.

The proposed standard is illustrated in (6) and (7).

(6)

MWE pattern	**comments**
MWEp1	expressions headed by a verb taking a subject and a direct object NP that consists of a determiner and a singular noun

(7)

MWE pat.	cl	example sentence
MWEp1	de boot missen 'to miss the boat'	hij heeft de boot gemist 'he has missed de boot'
MWEp1	de geest geven id. 'to die'	hij heeft de geest gegeven id. 'he has died'
MWEp1	...	...

An example of an MWE pattern description is shown in (6), whereas (7) shows some instances of the EC *MWEp1*. Having such an EC, representations for a specific theory and implementation can be derived. The procedure is that one instance of an EC, e.g. *de boot missen* of the EC *MWEp1*, must be converted in part manually. By defining and formalizing the conversion procedure, the other instances of the same EC can be converted fully automatically. In other words, having the ECs consisting of MWEs with the same pattern, it requires some manual work to convert one instance of each EC into a system specific representation, but all other members of the same EC can be done fully automatically.

Besides a standard representation for MWEs, the ECM proposes a conversion procedure which focuses on systems that deal with MWE surface syntactic structures in the same way as they deal with normal syntactic structures, such as the Rosetta MT system. The proposed procedure consists of two parts, viz. a manual part and an automatic part, and has been illustrated for the Rosetta system in Odijk (2004b, 2003).

The manual part has to be carried out once for each MWE pattern *P* and consists of five steps:

Step 1. Select an example sentence for MWE pattern *P*, and have it parsed by the system. This will usually yield multiple parses, from which the correct one for this example sentence must be selected manually,[6] yielding the *Reference Parse* (RP).[7]

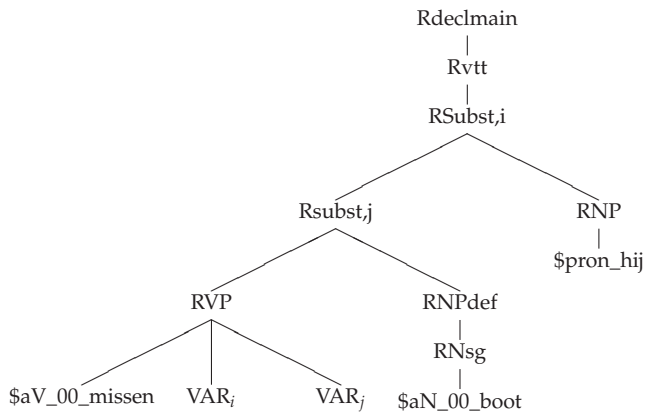
For Rosetta: For MWE pattern *MWEp1* as described in (6), we select the first example sentence from (7), which is *hij heeft de boot*

[6]The free text comments describing the MWE pattern should be of help in selecting the right parse.

[7]It should be noted that before one can start with the conversion procedure it should be checked that the grammar is correct and that the lexical components that are needed for parsing the MWE example sentence are present in the system (see Odijk (2004b)).

gemist. Parsing this sentence by the Rosetta system yields the following RP:

(8)



Step 2. Define a transformation to turn the RP into the MWE structure.

For Rosetta: Delete everything above the node containing the rule *Rsubst,j*.

Step 3. Determine the list of unique identifiers of the lexical components used in the MWE, using the derived MWE structure, yielding the *Component ID List* (CIDL).

For Rosetta: The CIDL of this MWE is:
⟨$V_00_missen,$aN_00_boot⟩, in this order.

Step 4. Define a transformation to relate CL and CIDL.

For Rosetta: The components listed in the CL in (7), viz. *de boot missen*, can be brought in correspondence with the CIDL by applying the transformation 1 2 3 → 3 2, i.e. delete the first element and reverse the remaining list.

Step 5. Apply this transformation to the CL, yielding the *Transformed CL* (TCL) and check that the citation form of each lexical item equals the corresponding element in the CIDL.

For Rosetta: Applying this transformation turns the CL *de boot missen* into the TCL *missen boot*. The citation forms of the CIDL correspond to the elements in the TCL:

- citation form ($aV_00_missen) = *missen*

- citation form ($aN_00_boot) = *boot*

After the manual work is done for one instance of an equivalence class, the transformation of all other members of the EC with MWE pattern *P* can be done automatically. The automatic part of the conversion procedure also consists of five steps, and is applied to the MWE *de geest geven* as illustration.

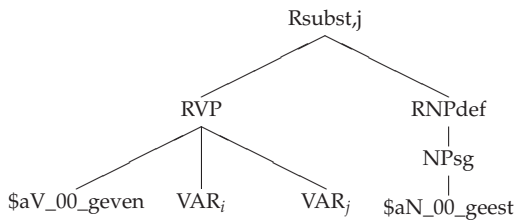
Step 1. Parse the example sentence of the MWE and check that it is identical to the RP for the example sentence used in the manual step, except for the lexical items.

For Rosetta: Parsing the example sentence *hij heeft de geest gegeven* indeed leads to an RP that is identical to the one in (8), except for the lexical items.

Step 2. Use the transformation defined in the manual step to turn the RP into the structure of the MWE.

For Rosetta:

(9)



Step 3. Select the component identifiers from the parse tree, in order to obtain the CIDL.

For Rosetta: The CIDL is ⟨$aV_00_geven,$aN_00_geest⟩, in this order.

Step 4. Apply the MWE component transformation to the CL, in order to obtain the TCL.

For Rosetta: The MWE component list transformation applied to the CL *de geest geven* yields *geven geest*.

Step 5. Check that the citation form of each item in the CIDL equals the corresponding element in the TCL.

For Rosetta: The citation form of each item in the CIDL equals the corresponding element in the TCL:

- citation form ($aV_00_geven) = *geven*
- citation form ($aN_00_geest) = *geest*

This concludes the description of the conversion procedure as proposed by the ECM. Applying this procedure results in (1) a set of syntactic structures and (2) identifiers to the lexical components. Both are specific to the NLP system, in this case Rosetta, and form the basic ingredients for the treatment of MWEs in the target system.

A potential problem of the ECM as proposed is the risk that the number of ECs will run into thousands the majority of which contains only a small number of MWEs.[8] Since the ECM concentrates on minimizing the manual work when incorporating a large number of MWEs in a specific system, the method will be less successful if there are many ECs with only a few instances. In order to reduce the number of ECs and to increase the number of members within each EC, Odijk (2004b) introduces parameterized equivalence classes.

2.3.2 THE PARAMETERIZED ECM

The central idea behind the parameterized ECM is that many MWE patterns describe structures that are for a large part identical and differ only locally. Take for example pattern description *MWEp1* in (6), which requires a singular noun. However, another pattern is required that is identical except that it requires a plural noun. Moreover, the description of another pattern is needed for a diminutive singular noun, and yet another one that requires a diminutive plural noun. In most theories and NLP systems such local differences are treated locally, e.g. locally different rule names (Rosetta, 1994) or features. Odijk (2005) makes use of this fact by introducing parameters to represent local variation, in this case the variation of the noun, which makes it possible to reduce the number of MWE pattern descriptions from four different MWE patterns to one single MWE pattern that takes two parameters: one to specify the number of the noun and one to specify whether the diminutive form must be used or not. In other words, instead of having a pattern description *MWEp1*, described in (6), for MWEs such as *de boot missen*

[8] This problem has also been raised by Copestake et al. (2002), though not in relation to the ECM.

and another pattern description *MWEp2*, described in (10), for MWEs such as *de benen nemen* (lit. 'to take the legs', id. 'to run off'), there is one pattern description *MWEp3*, described in (11), for both types of MWEs.

(10)

MWE pattern	comments
MWEp2	expressions headed by a verb taking a subject and a direct object NP that consists of a determiner and a plural noun

(11)

MWE pattern	comments
MWEp3	expressions headed by a verb taking a subject and a direct object NP that consists of a determiner and a noun

Though extending the ECM with parameters introduces more theory-dependent assumptions, the approach as a whole is still as theory-neutral as possible: NLP systems that can make use of these parameters will profit from it, while systems that cannot make use of these parameters are not harmed since the original equivalence classes can still be identified by grouping instances with the same parameter combinations.

In Section 4.1, I will elaborate on the organization of DuELME using parameterized equivalence classes and give an overview of the linguistic phenomena that are dealt with by parameters.

2.4 OUTLINE

In the following chapters of this part of the dissertation, I will discuss the development of DuELME in detail. Chapter 3 describes the extraction of candidate expressions from corpora and elaborates on the selection of the lexical entries. The design and implementation of DuELME will be described in Chapter 4. Chapter 5 concludes the first part of this dissertation with a discussion and an evaluation of DuELME.

CHAPTER 3

EXTRACTION AND SELECTION OF MWES

No resource without data. The data for DuELME have been automatically extracted from corpora.[1] We make use of corpora, because we want our lexicon to reflect actual language usage and because we do not want to restrict ourselves to a linguist's imagination of which uses are possible or actually occur. However, we are aware of the fact that some MWEs are relatively rare and the corpora used may not be large enough to fully show their uses. For this reason and because of the unreliable output produced by the extraction methods, the data extracted have been carefully analyzed manually before creating entries for MWEs.

One of the reasons why we cannot fully rely on the output from the extraction methods is because the extraction techniques sometimes erroneously identify groups of words as an MWE or they group different expressions that share some but not all words together. Furthermore, the extraction is in part based on an automatic syntactic parse of the corpus sentences, and these parses may be incorrect. Since there is no straightforward way to interpret the data fully automatically, the selection of MWEs has been done manually.

The automatic extraction of the data from corpora is addressed in Section 3.1, and Section 3.2 elaborates on the manual selection of MWEs for DuELME.

[1] The identification of MWEs has been carried out by Begoña Villada Moirón (University of Groningen, The Netherlands) within the STEVIN IRME project.

3.1 EXTRACTION

The MWE extraction process consists of two steps. The first step constitutes the automatic identification of candidate expressions, which is followed by collecting their possible morpho-syntactic variation.

3.1.1 AUTOMATIC MWE IDENTIFICATION

Before the extraction techniques can be applied, it needs to be decided what we want to extract. The intended result is to obtain an exhaustive list of real MWEs. However, the output of the extraction method applied is a list of so-called candidate expressions, i.e. combinations of two or more lemmas that may form an MWE or may be part of an MWE. The number of lemmas in the candidate expressions depends on the input; the automatic extraction techniques require predefined syntactic patterns as input. Very specific patterns can be used, but the amount of extracted data is higher with a simple pattern, such as a pattern the head of which is a verb that takes a direct object headed by a noun.

A disadvantage of the input requirement is that it determines for a large part the variety of different MWE patterns in DuELME and that MWEs that do not match the predefined patterns are not identified and hence not included in DuELME.

We chose the six most frequently occurring patterns from a list of patterns created by parsing a random selection of MWEs taken from the Van Dale Lexical Information System (VLIS) database. The selected patterns are shown in (12).

(12) **NP_V** NP(DIRECT OBJECT) - verb
(NP)_PP_V variable NP(DIRECT OBJECT) - PP - verb
NP_NP_V NP(INDIRECT OBJECT) - NP(DIRECT OBJECT) - verb
A_N adjective - noun
N_PP noun - PP
P_N_P preposition - noun - preposition

It should be noted that the patterns have been used as defined, i.e. the patterns do not include any other complements than the ones stated. Moreover, only the head of the patterns and the head of the complement(s) are taken into account with the automatic extraction,

i.e. no explicit search is done for e.g. adjectives modifying the head of the direct object in the NP_V pattern. Since combinations such as determiner-adjective-noun have been formed during the manual selection process, DuELME contains a lot more MWE patterns than the five input patterns.

The first step in the automatic MWE identification is the extraction of tuples, i.e. sequences of lemmas formed by the head of the pattern and the head(s) of the complement(s), from the Dutch CLEF corpus, a collection of newspaper articles from 1994–1995 taken from the Dutch daily newspapers Algemeen Dagblad and NRC Handelsblad. The corpus contains 80 million words and 4 million sentences, which have been annotated automatically with the Alpino parser. The tuples form the input for the identification models. Based on experiments with various machine learning techniques, Villada Moirón (2006) chose to apply a decision tree classifier. The classifier learns a notion of MWE-hood on the basis of training data that consist of a collection of tuples and a number of features that encode linguistic information. Such features measure the lexical affinity between the component words, the syntactic flexibility, the strength of the dependence between the words, passivizability, etc. Each tuple in the training data has been labeled as either *MWE* or *non-MWE*. Two existing lexical databases, VLIS and the RBN (Martin and Maks, 2005), have been used to annotate the training data.

The identification model also makes use of an absolute frequency threshold, i.e. tuples that occur infrequently are not taken into account since they would introduce noise and degrade the performance of the classifier. A desirable threshold has been established empirically per syntactic pattern, see Table 3.1. The chosen threshold was the one yielding the best performance of the classifier.

The decision tree classifier proposes a class (MWE|noMWE) for each input tuple. Although the classification includes a probability that suggests how confident the classifier is in assigning a given class to a tuple, no use has been made of this probability. The identification provides a list of candidate expressions, i.e. tuples that are assigned the class *MWE*, yielding a total of 9,451 expressions, see table 3.2.[2] No manual

[2]It should be noted that the extraction techniques used fail to come up with all the MWEs that occur in the corpus: Villada Moirón (2007) reports a recall of 0.53 in the identification of Dutch MWEs using the decision tree classifier.

pattern	threshold used
NP_V	f>=10
(NP)_PP_V	f>=10
NP_NP_V	f>=10
A_N	f>=50
N_PP	f>=30
P_N_P	f>=50

Table 3.1 *Absolute frequency threshold (f) used for each pattern.*

pattern	# of candidate expressions
NP_V	3,894
(NP)_PP_V	2,405
NP_NP_V	202
A_N	1,001
N_PP	1,342
P_N_P	607
Total	**9,451**

Table 3.2 *Distribution of candidate expressions over the extracted patterns.*

filtering or correction has been applied to this list at this stage.

3.1.2 COLLECTING MORPHO-SYNTACTIC INFORMATION

MWEs allow morpho-syntactic variation, e.g. verbs may show different forms depending on tense, person, etc.; nouns may allow number alternation, etc. Evidence of morpho-syntactic variation for the candidate expressions has been collected from two different corpora, viz. the CLEF corpus, which has been used for the candidate expressions with the patterns A_N and N_PP, and the Twente Nieuws Corpus (TwNC) (Ordelman, 2002), which has been used for the other patterns. The TwNC comprises 500 million words of newspaper text and television news reports and has also been syntactically annotated with the Alpino parser. For each candidate expression the following set of properties

and their frequencies have been extracted:[3]

1. the subcategorization frame assigned by the Alpino parser;
2. the absolute frequency of the tuple;
3. a list of heads of co-occurring subjects;
4. number information of the noun;
5. diminutive information of the noun;
6. a list of determiners co-occurring with the noun;
7. a list of heads of modifiers pre-modifying the noun; and
8. a list of heads of modifiers post-modifying the noun.

The candidate expressions, their properties and for each candidate six examples are stored in so-called data records. Appendix A describes the data record format for each pattern extracted illustrated with an example.

3.2 SELECTION

The 9,451 data records representing the candidate expressions, their properties and corpus examples form the input for the manual data analysis. The analysis includes selecting true MWEs and deciding on their precise form.

The MWEs for DuELME have been manually selected according to the definition given in Chapter 1, and for convenience repeated in (13).

(13) A multiword expression is a combination of words that has linguistic properties not predictable from the individual components or the normal way they are combined.

MWEs show idiosyncrasies at different levels of analysis, see the examples given in the general introduction. Not all types of linguistic properties have been detected during the analysis, mainly because of the extraction methodology and the text type, viz. news papers, of the corpora used.

[3]It is important to note that the extraction techniques do not distinguish between different interpretations of the corpus examples, such as literal and non-literal. This means that the frequencies extracted for each property represent the total number of occurrences found for the tuple.

Although we clearly defined the concept of MWE, it is not always easy to determine whether a combination is a true MWE. In order to classify a combination as idiosyncratic, one not only needs to know the properties of the individual components and the standard grammar rules, but there must also be a general agreement about these rules and properties. MWEs with morpho-syntactic idiosyncratic properties are relatively easy to identify, because this type of idiosyncrasy exhibits in the form of the expression. Also combinations the meaning of which is not straightforwardly compositional often do not need much argumention to be selected as true MWE. More problematic are combinations for which it is unclear whether the choice of the specific items can be derived from the semantic properties of these items.

To illustrate the problem, an example of a clear MWE is *een gesprek voeren* ('have a conversation'): although one meaning of *voeren* is 'being actively occupied with', and although one can be actively occupied with a conversation, the combination is unpredictable since *gesprek* cannot be substituted by its synonym *praatje* ('chat'), i.e. **een praatje voeren* ('have a chat') is out. For this reason, *een gesprek voeren* is classified as a true MWE and thus entered in the lexicon.

An illustration of a not-so-clear-cut example is the expression *een getuigenis afleggen* ('give a testimony'). The extracted data contain five other nouns that occur with *afleggen*, three of which requiring the same meaning of *afleggen* as required by the noun *getuigenis*: *verklaring* ('statement'), *eed* ('oath'), and *bekentenis* ('confession'). The question is whether the lexical selection of the noun is predictable from its semantic properties. In this case we are not sure, since we do not know which semantic proporties a noun that the verb *afleggen* selects requires. Although the expression seems semantically regular, a detailed study for each of such cases is required to determine the precise properties, and hence to establish whether the set of nouns that co-occurs with the given verb can be described by means of semantic selection restrictions or that the noun-verb combination is lexically determined. Since we are not aware of the exact rules for these types of combinations, I assume that NLP systems have no explicit encoding of these rules, and therefore these types of expressions are regarded as MWEs and included in DuELME.

A single data record may contain a lemma tuple that is part of more than one MWE. An extreme example of such a data record is:[4,5]

```
heb#hand
frame transitive_ndev 1280,np_ld_pp 181,aci_simple 22,
freq 1497
hd heb
subject hij 149,die 96,ik 70,ze 67,je 46,zij 28,we 27,
compl1 hand
hd1 hand
hdcomp1
dep1 obj1 1497,
mor1 sg 908,pl 589,
dim1 nodim 1497,
det1 de 696,een 235,NO 208,geen 90,zijn 74,hun 62,
premod1 NO 875,gelukkig 123,vrij 118,schoon 75,
postmod1 NO 1186,in 115,van 99,op 24,bij 14,vol 12,
328.xml|hij had zijn handen vol om een boterham te verdienen
234.xml|en heeft de handen vol aan drugssmokkelaars
469.xml|Hij is een pianist die vier handen leek te hebben
958.xml|Het Iraakse regime heeft de hand gehad in de dood van
452.xml|Ook daar had God de hand in
796.xml|De meisjes hadden hun handen op de gebogen knieen
```

The tuple, in this example *hand hebben* ('hand have'), is formed by the head of the predefined pattern, here NP_V, and the head of the complement. Combinations that include for example a determiner-adjective-noun constituent have to be created and checked manually by analyzing the extracted properties and the corpus examples, and by using knowledge of the language and in some cases a dictionary. Given the information in the data record, at least four different expressions can be identified:

(14) a. de vrije hand hebben
the free hand have
id. 'to have a free hand'

b. een gelukkige hand hebben
a lucky hand have

[4]The numbers represent the absolute frequency of the number of occurrences of the value.

[5]As stated, the extracted pattern does not include any other complements than the ones defined. In this case the extracted pattern is NP_V. Given the example sentences we can conclude that the Alpino parser analyzes PPs as modifiers instead of complements, because the subcategorization pattern of *hebben* ('have') individually differs from the subcategorization pattern of *hebben* in the expressions *the hand hebben in iets* and *de handen vol hebben aan iets*, cf. the examples 234.xml and 452.xml.

id. 'to be lucky'

c. de hand hebben in iets
the hand have in s.th.
id. 'to have a hand in s.th.'

d. de handen vol hebben aan iets
the hands full have on s.th.
id. 'to have one's hands full with s.th.'

Because the data records have been analyzed manually, more types of MWEs have been detected than would have been identified by using only automatic techniques. For this reason, DuELME contains a total of 141 MWE patterns, while only five patterns have been used as input for the automated extraction. The manual analysis resulted in DuELME containing a total of 5,607 MWEs.[6]

To summarize, MWEs for the lexicon have been selected from lists of candidate expressions, their properties and example sentences according to the definition given in (13). The selection needs to be done manually, since there is no straightforward way to interpret the data fully automatically. The information given in the data record needs to be analyzed carefully to identify one or more MWEs and to determine the correct form of an MWE.

[6]This total includes a number of 375 MWEs that have not been assigned an MWE pattern, see Chapter 5.

CHAPTER 4

DESIGN AND IMPLEMENTATION

An MWE resource that is meant to be used in a wide variety of NLP systems should be designed and implemented in such a way that its integration into an NLP system can be done with a minimal amount of manual effort. To achieve this goal I take an innovative approach based on the Equivalence Class Method (Odijk, 2003). In Chapter 2 I have presented the original ECM including the standardized lexical representation for MWEs as proposed by Odijk (2003). Several extensions and improvements of the proposed method are possible. Odijk (2004b) discusses the extension of the ECM with parameters, with the goal of reducing the number of ECs and increasing the number of members within each EC, and hence reducing the number of MWEs that need to be dealt with manually. Odijk illustrates the introduction of parameters using two properties of the noun, viz. number and whether the diminutive form must be used. There are, however, a lot more linguistic phenomena that can be parameterized. One of the goals of this chapter is to optimize the use of parameters and hence the use of the ECM.

Besides refining the parameterized ECM, the focus is on extending the standardized lexical representation with core properties to improve the compatibility with existing frameworks and implementations. A uniform representation has been designed for the description of the MWE patterns and the description of the individual expressions for in DuELME.

This chapter starts with refining the parameterized ECM in Section 4.1. This is followed by elaborating the representation of the MWEs and their patterns in Section 4.2. Finally, the implementation of the database

structure is described in Section 4.3.

4.1 REFINING THE PARAMETERIZED ECM

Parameters have been introduced by Odijk (2004b) in order to reduce the number of ECs in the original ECM, and hence to reduce the amount of manual effort when integrating the standard into NLP systems. Odijk illustrates that by parameterizing two properties of the noun, viz. the number and whether the diminutive form must be used, the number of ECs can be reduced from four different ECs to one single EC.

To optimize the use of parameters it needs to be studied (1) which linguistic phenomena qualify for parameterization, and (2) how to represent parameters in the lexical descriptions.

Given the structure of an MWE, it is local variation, i.e. the morpho-syntactic properties of individual elements, that can be specified independently of the description of the structure. In Dutch, for example, an adjective premodifying a noun always inflects (*-e* inflection) except when the adjective forms a constituent with an indefinite determiner (or no determiner) and a neuter noun, cf. (15) and (16).

(15) het mooie huis
the beautiful house

(16) een mooi huis
a beautiful house

Adjectives in MWEs, however, may behave idiosyncratically with respect to inflection. For example, the adjective *geheim* in the MWE *geheim agent* ('secret agent') does not allow inflection despite of the noun being masculine. Inflection of the adjective is typically a local phenomenon that can be parameterized.

A total of 26 parameters have been defined for Dutch. Before I give an overview and discuss how to represent parameters, I will first elaborate on the term *parameter*. In this study, the term *parameter* is a feature and can be defined as an occurrence of the pair <parameter category,parameter value>, where *parameter category* refers to the aspect that is parameterized, and *parameter value* to the value a parameter category takes. Examples of parameters are <nnum,sg> for singular nouns, <afrm,sup> for superlative adjectives, <vfrm,part> for

particle verbs. Table 4.1 gives an overview of the parameters used in DuELME including a description of each parameter category and the parameter values.[1]

The next question is how to represent the parameters. Recall that the ECM distinguishes between MWE pattern descriptions and descriptions of individual MWEs. In the original ECM, an MWE description consists of: (1) an MWE pattern name, (2) a list of MWE components (CL), and (3) an example sentence. In DuELME, the CL contains the obligatory lexically fixed components of an MWE in the lemma form.[2] Since parameters specify the variation of the individual components of individual MWEs, the CL is the best place to represent them. As can be seen in Table 4.1, each parameter value is unique and therefore we only need to represent the parameter value of each parameter. The parameter values are notated between square brackets directly to the right of the item they parameterize. Default values, here the first values given in the table for each PC, are not represented. An example of the CL representation of the MWE *de benen nemen* is *de been[het][pl] nemen*, where the noun is represented in the non-inflected form and where *[het]* specifies the gender of *been*, viz. neuter, and *[pl]* specifies that the form of *been* in the MWE must be plural.

Extending the ECM with parameters introduces more theory-dependent assumptions, though by parameterizing only very local phenomena, the theory-dependence is minimized as much as possible. Some local variation, such as the number of the noun, is part of standard Dutch grammar and is treated in grammars in a local way, e.g. by using local features or rules. This means that in general the parameters defined can be linked to individual rules or features of specific NLP systems.

The extension with parameters contributes to reducing the number of ECs and increasing the number of members within each EC. As a result the number of MWEs that have to be dealt with manually decreases, whereas the number of MWEs that can be automatically incorporated into an NLP system increases. This immediately contributes to the successfulness of the method, which depends, inter alia, on (1)

[1] The gender and countability of the noun are basically not a type of variation, but are properties of the noun that need to be specified in order to determine the correct form of other components, e.g. the standard inflection of the adjective depends on the gender of the noun.

[2] See Section 4.2.2.

parameter <PC,PV>	PC description	PV description
<*dbin,sb*>	binding type	subject bound
<*dbin,dob*>		direct object bound
<*dbin,iob*>		indirect object bound
<*ngen,de*>	the gender of the noun	masculine and feminine nouns
<*ngen,het*>		neuter nouns
<*ncount,count*>	the countability of the noun	count noun
<*ncount,mass*>		mass noun
<*nnum,infl*>	the number of the noun	inflectable
<*nnum,sg*>		singular
<*nnum,pl*>		plural
<*nfrm,pos*>	the form of the noun	positive
<*nfrm,dim*>		diminutive
<*afrm,norm*>	the form of the adjective	normal
<*afrm,noe*>		never *-e* inflection
<*afrm,optepl*>		No *-e* inflection when noun is singular, optional when noun is plural.
<*afrm,noesg*>		No *-e* inflection when noun is singular.
<*afrm,opte*>		*-e* inflection is optional
<*afrm,comp*>		comparative
<*afrm,sup*>		superlative
<*vfrm,fin*>	the form of the verb	finite
<*vfrm,inf*>		infinitive
<*vfrm,part*>		particle verb
<*vfrm,presp*>		present participle
<*vfrm,passp*>		passive participle
<*ppos,prep*>	the way the adposition	preposition
<*ppos,post*>	must be realized	postposition

Table 4.1 *Overview of parameters, with descriptions of the parameter category (PC) and the parameter value (PV).*

how many different ECs are distinguished (the fewer the better), and (2) how many instances each ECs contains (the more the better).

To determine the effectiveness of the method, measurements have been carried out on DuELME. A total of 5,232 unique expressions have

Cov.	# MWEs	# ECs	# parameterized ECs
50%	2,616	101	10
60%	3,139	166	16
70%	3,662	272	25
80%	4,186	441	38
85%	4,447	572	48
90%	4,709	785	63
95%	4,970	1,046	87
100%	5,232	1,308	140

Table 4.2 *Coverage of ECs.*

been included in the evaluation.[3]

To measure the number of ECs without parameters, I counted the number of unique parameter combinations from the CL-fields of each parameterized EC. For example, in the parameterized ECM the CLs *de boot[sg] missen* and *de been[het][pl] nemen* occur in the same EC. In the original ECM, these CLs would appear in different ECs, due to the variation of the number of the noun.

Table 4.2 shows the major findings of the measurements. The first row, for example, means that 50% (or 2,616) of the expressions can be dealt by 101 ECs in the original ECM and just 10 classes in the parameterized ECM. The main conclusion that can be drawn from the results is that introducing parameters in the ECM reduces the number of ECs by almost 90%, and multiplies the average cardinality of the ECs with a factor of over 9.3 for the whole set of MWEs.

To conclude, once a mapping of the parameters to rules of a specific NLP system has been established, it takes only 140 manual conversions of the standard lexical representation to system specific representations, instead of manually converting 1,308 entries in the original ECM. This means that NLP systems that can make use of these parameters will profit from the extended ECM. Systems that cannot make use of these parameters are not harmed since the original equivalence classes can still be identified by grouping instances with the same parameter combinations.

[3]The 375 MWE that have not been assigned an MWE pattern have been excluded from this test.

4.2 REPRESENTATION

To optimize the compatibility of DuELME with existing frameworks and implementations, the standardized lexical representation for MWEs as proposed within the ECM has been extended. The result is a uniform representation for MWE descriptions as well as for MWE pattern descriptions. The MWE pattern description is discussed in Section 4.2.1 and the MWE description is presented in Section 4.2.2. Detailed information about the ingredients that are part of the descriptions can be found in Grégoire (2007a). The implementation of the lexical entries in DuELME is described in Section 4.3.

4.2.1 MWE PATTERN DESCRIPTION

The MWEs in DuELME are classified according to their pattern, yielding besides individual MWE descriptions also MWE pattern descriptions. In the original ECM, an MWE pattern description consists of (1) a pattern name; and (2) comments, i.e. free text in which the uniqueness of the pattern is described. This description specifies the syntactic category of the head of the MWE, the complements it takes and their internal structure. The pattern name is not more than an identifier that labels a group of MWEs with the same pattern. The comment field is not required for the purpose of the ECM, but can be of help to humans when creating new MWE descriptions or during the conversion procedure when comparing the structure of the standard representation with a system specific structure. Of course, a linguist must understand what is stated in the comments. Since comparing structures during the conversion procedure is in part a manual task, it is easy to make mistakes. In the current approach, a formal representation of the patterns has been added to the pattern descriptions and it is indicated whether individual components can be modified, see (17).

(17) Expressions headed by a verb, taking a direct object consisting of a fixed determiner and an unmodifiable noun.
[.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]

The notation used to describe the patterns is a formalization of dependency trees, in particular CGN (*Corpus Gesproken Nederlands* 'Corpus of Spoken Dutch') dependency trees (Hoekstra et al., 2003), which

have become a de facto standard for Dutch (van Noord et al., 2006). Most recent projects make use of this format, see e.g. the D-Coi project,[4] LASSY,[5] and SoNaR.[6] CGN dependency structures are based on traditional syntactic analysis described in the Algemene Nederlandse Spraakkunst (Haeseryn et al., 1997) and are aimed to be as theory-neutral as possible. Because the formal representation is in agreement with a de facto standard for Dutch, most Dutch NLP systems are able to use it for the conversion procedure, yielding an optimal reduction of manual labor. For instance, the formal representation can be used for automatic verification of the reference parse, which might be suitable for a conversion to the Rosetta system (see Section 2.3), or it can be used for direct conversion, which has been done in the conversion to the Alpino system (see Section 5.2).

The patterns are encoded using a formal language, which is short and which allows easy visualization of dependency trees. The dependency labels (in lower case) and category labels (in upper case) are divided by a colon (:), e.g. *obj1:NP*. For leaf nodes, the part-of-speech is represented instead of the category label. To cover the modifiability of the noun and adjective,[7] additional labels have been created, see Table 4.3.

Label	Description
A	not modifiable adjective
A1	modifiable adjective
N	not freely modifiable noun
N1	modifiable noun
N2	limitedly modifiable noun

Table 4.3 *Additional labels to cover modifiability of nouns and adjectives.*

Leaf nodes are followed by an index that refers to the MWE component as represented in the CL-field (see Section 4.2.2), e.g. (1) refers to the first component of the CL, (2) to the second, etc. Obligatory open

[4] http://lands.let.ru.nl/projects/d-coi/

[5] http://www.let.rug.nl/vannoord/Lassy/

[6] http://lands.let.ru.nl/projects/SoNaR/

[7] Modifiability of the adjective includes variation of the form, e.g. comparative and superlative.

slots are represented with the label *var* instead of an index, e.g. [obj1:NP (var)], [obj2:NP (var)], etc.:

(18) *iemand de loef afsteken* (id. 'steal a march on s.o.')
[.VP [.obj2:NP (var)] [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]

I have not only extended the standard MWE pattern description with a formal notation of the structure, but also created five more fields mainly meant to ease the assignment of MWE patterns to individual expressions and to help the linguist with the conversion to an NLP system. This means that besides a pattern name, a pattern and a textual description, the MWE pattern description contains the following fields:

POS encodes the part-of-speech tag for each leaf node in the PATTERN-field. The POS-field is mainly used for maintenance reasons, i.e. with the help of this field it is possible to limit the number of candidate pattern descriptions for an expression. For example, if we want to assign a pattern description to the MWE *de boot missen*, we can use its part-of-speech sequence *d n v* to narrow down the number of possible pattern descriptions.

MAPPING indicates the relation between the position of a component in the Component List (CL) and its position in the EXAMPLE-field, i.e. the relation between non-inflected forms and full forms, see Section 4.2.2 for an illustration;

EXAMPLE_MWE contains an example of how to represent the MWE in the EXPRESSION-field of the MWE description.

EXAMPLE_SENTENCE illustrates by means of an example sentence how the example sentences should be constructed in the EXAMPLE-field of the MWE description.

COMMENTS is used to specify notes.

An example of an MWE pattern description stored in DuELME is given in Table 4.4.

4.2.2 MWE DESCRIPTION

In addition to the MWE pattern description a standard has been created for the representation of individual MWEs. We started from the

PATTERN_NAME	ec1
POS	d n v
PATTERN	[.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]]
MAPPING	3 4 5
EXAMPLE_MWE	de boot missen
EXAMPLE_SENT.	hij heeft de boot gemist
DESCRIPTION	Expressions headed by a verb, taking a direct object consisting of a fixed determiner and an unmodifiable noun.
COMMENTS	

Table 4.4 *Example of an MWE pattern description.*

MWE description as proposed by Odijk (2003), which contains three components: (1) an MWE pattern name, (2) a list of MWE components, and (3) an example sentence. To enrich the description of MWEs and to improve the compatibility of the standard representation with different frameworks and implementations, the standard has been extended, yielding an MWE description that contains two parts, viz. a basic MWE description and an extended MWE description.

For the development of the representation, the parsers of two Dutch NLP systems have been consulted, viz. the Alpino parser and the Rosetta MT system (Rosetta, 1994). Both parsers contain a method to deal with (certain types of) MWEs and include lexical representations for MWEs. In Alpino the components that form the MWE are either listed in their full (inflected) form, or in their non-inflected form, depending on the type of expression (see Section 5.2), whereas in Rosetta they are listed in their non-inflected form. In Rosetta, syntactic rules are used to derive the full MWE form and variants of the MWE are realized using different rules. In Alpino, variants of an MWE that is listed in the inflected form need to be explicitly specified in the lexicon, e.g. the MWE *de/zijn hielen lichten* (lit. 'to lift the/his heels', id. 'to take to one's heels') has one entry for *de hielen lichten* and one entry for *zijn hielen lichten*, and furthermore entries for the different forms of the possessive, i.e. *mijn hielen lichten*, *je hielen lichten*, *haar hielen lichten*, etc. The difference between the two representations needs to be accounted for in the standard representation for MWEs. Moreover, special attention is required for the representation of various grammatical properties, e.g. how to

EXPRESSION	de boot missen	blunder ('mistake')
CL	de boot[sg] missen	blunder
PATTERN_NAME	ec1	ec2
LISTA	n.a.	maken ('make')
LISTB	n.a.	begaan ('commit')
EXAMPLE	hij heeft de boot gemist	hij heeft een blunder begaan

Table 4.5 *Two examples of basic MWE descriptions.*

represent negative polarity items? How to deal with variation of individual components of MWEs? What to do with MWEs containing a copular verb: is the verb part of the MWE or not? How to represent optional arguments? etc.

The MWE description comprises a basic description and an extended description. Since the main focus is on representing those properties that are needed for a successful implementation of the MWE lexicon in any specific NLP system, the priority is on properly describing the fields that are part of the basic MWE description, and although the additional description fields also form an important part of the MWE description, less attention has been paid to this part of the representation.

Table 4.5 shows two examples of basic MWE descriptions. The following fields are part of the basic description:[8]

EXPRESSION The first description field is the EXPRESSION-field which contains the obligatory lexically fixed components of an MWE. The components are represented in their full form, i.e. the form they take in the MWE. For nouns this means that if the noun can be both singular and plural, the singular form is represented. Determiner alternation, if applicable, is represented using a slash (/) to separate the alternations, e.g. *de/zijn hielen lichten*.[9] The order of the components should match the order of the pattern in the MWE pattern description.

[8]Clear representation guidelines have been defined for each description field in DuELME. A detailed overview of the representation rules can be found in (Grégoire, 2007d).

[9]The representation of determiner alternation is limited to single words, i.e. a determiner group such as *een paar* ('a few') cannot be represented in combination with the alternation symbol (/).

CL The Component List contains the same components as the EXPRESSION-field. The difference is that the components in the CL are represented in the non-inflected form, instead of in the full form. Parameters are used to specify the full form characteristics of each component, see Section 4.1. An overview of the 26 parameters that have been defined for Dutch is given in Table 4.1. There are special representation rules for adpositions and determiners.

PATTERN_NAME contains a reference to an MWE pattern description. Some MWEs can have optional arguments, such as an optional indirect object or *aan*-PP ('to-PP'). To account for these cases, up to three patterns can be specified for each MWE. An example of an entry with multiple patterns represented is *het woord vragen* (lit. 'to ask the word', id. 'to ask to be able to speak'): the assignment of PATTERN_NAME1 yields the MWE *het woord vragen*, and the assignment of PATTERN_NAME2 yields the MWE *iemand het woord vragen* ('to s.o.').

LISTA and LISTB The use of these fields is restricted to three types of expressions:

- Combinations of a verb that seems to have very little semantic content, e.g. copulas and support verbs/light verbs, and a prepositional phrase, a noun phrase or an adjectival phrase. Since the complement of the verb is used in its normal sense, the constructions are subject to standard grammar rules, which include topicalization, internal modification, etc. Examples are:

  (19) aan de bak raken/komen
  on the bin get/come
  'get a job'

  (20) fout maken
  mistake make
  'make a mistake'

- Combinations of a noun and a verb for which the exact properties of the individual components and the existence and character of the rules to combine them are unknown. See the

discussion in Section 3.2 for more information on this type of MWEs.

- Lexically restricted combinations of an adjective and a noun and combinations of an adjective with an irregular meaning and a noun that is used in its literal sense, e.g. (21). Both components are subject to standard grammar rules, and also occur in grammatical constructions other than in the same NP, e.g. the noun as a subject and the adjective as a predicative complement.

  (21) zwaar accent
  heavy accent
  'strong accent'

The lexical selection of the verb and the adjective is highly restricted, but not always limited to one. The alternation of the verb or the adjective should be specified in the LIST-fields. The reason for using two LIST-fields is to separate predefined list values, which are represented in the LISTA-field, from special list values, represented in LISTB-field. The predefined list values are high-frequency verbs that are known to occur often as so-called light verbs, especially with PPs. Two sets of verbs are predefined:

(a) *blijken* ('appear') *blijven* ('remain') *gaan* ('go') *komen* ('come') *lijken* ('appear') *raken* ('get') *vallen* ('fall')[10] *worden* ('become') *zijn* ('be')

(b) *brengen* ('bring') *doen* ('do') *geven* ('give') *hebben* ('have') *houden* ('keep') *krijgen* ('get') *maken* ('make') *zetten* ('put')

A complement co-occurs either with verbs from set (a) or with verbs from set (b). Each verb from the chosen set is checked against the occurrences found in the corpus data. If a verb does not occur in the corpus data and is also not plausible in constructed data, it is deleted from the LISTA-field.

The LISTB-field contains lexemes, either verbs or adjectives, that are not listed in the predefined set but do co-occur with the com-

[10]The literal meaning of *vallen* is 'fall', but it has a variety of different meanings in MWEs of this type, including 'become', 'is experienced as', etc.

ponent(s) in the EXPRESSION-field. The information in the LISTB-field is merely based on corpus data and therefore may not be exhaustive.

EXAMPLE contains an example sentence with the expression. The only requirement of this field is that the structure is identical for all example sentences in the same equivalence class, i.e. with the same PATTERN_NAME.

The example sentence should be used when the conversion procedure requires a parse of the MWE. Furthermore, it can be used to determine the full form of individual components. As stated in the previous section, the MAPPING-field in the MWE pattern description indicates the relation between the position of a component in the Component List (CL) and its position in the EXAMPLE-field. For example, the CL of *de boot missen* is 'de boot[sg] missen', EXAMPLE is 'hij heeft de boot gemist' and MAPPING is '3 4 5': the first component of CL refers to the third component of EXAMPLE, the second component of CL refers to the fourth component of EXAMPLE, and the third component of CL refers to the fifth component of EXAMPLE.

The extended MWE description has mainly been designed to specify additional information of obligatory free arguments and contains the following fields:

SUBJECT encodes restrictions on the possible realizations of variable subjects and can contain both a list of heads of possible subjects extracted from corpora and predefined labels such as [sg] for singular subject.

OBJECT encodes restrictions on the possible realizations of obligatory variable objects and can contain both a list of heads of possible objects extracted from corpora and predefined labels such as [anim] for animate object.

MODIFIER encodes a list possible modifiers both for expressions in which the modifier is obligatory but variable, e.g. *het MOD voorbeeld geven* ('set a MOD example'), and for expressions that contain a (limitedly) modifiable noun. In the current encoding this field is mainly filled with modifiers coming from extracted data.

RPRON In MWEs containing a prepositional phrase with a variable complement, such a *een hekel hebben aan iets* ('to hate s.th.'), the PP can pronominalize, i.e. be realized as *er* ('there') + adposition. In some cases the adposition can be followed by a clause, either a clause starting with a complementizer, see the example in (22), or an infinitive clause, see the example in (23). To encode this possibility, the RPRON-field can take two predefined labels, viz. [ssub] for clauses starting with a complementizer and [vp] for infinitive clauses.

(22) hij heeft er een hekel aan dat zij komt
he has there a hackle to that she comes
'he hates it that she comes'

(23) hij heeft er een hekel aan weg te moeten
he has there a hackle to away to have
'he hates to leave'

CONJUGATION specifies whether the head of the expression conjugates with *zijn* ('to be'), or *hebben* ('to have'), or both.

POLARITY Some MWEs can only occur in positive or negative polarity environments. An example is the expression *een oog dichtdoen* (lit.' close an eye'), which can only be used in negative polarity environments. This field is unspecified by default and takes the value *NPI* (Negative Polarity Item) if the expression can only occur in negative polarity environments, and *PPI* (Positive Polarity Item) if the expression can only occur in positive polarity environments. See e.g. van der Wouden (1997) for more information about polarity.

Furthermore, the MWE description contains a field with a reference to a plain text file in which the information extracted from the corpora is stored. Any comments regarding the MWE description are entered in the optional COMMENT-field.

4.3 IMPLEMENTATION

Not only should the standard lexical representation be designed for optimal compatibility with various grammars, but it must also be imple-

mented in such a way that the lexical entries can simply be created and maintained and that any information needed can easily be provided. The implementation part includes, inter alia, developing a systematic way for avoiding the creation of duplicate MWEs and determining the correct MWE pattern.

The lexical entries are represented as a set of records which are stored in a MySQL database.[11] A Graphical User Interface for DuELME (DuELME-GUI)[12] has been built with PHP, which makes it possible to access the data stored in the MySQL database using a web browser. The DuELME-GUI combined with the data can be used to extend and enhance the resource and moreover as a research tool to study the MWEs analysed. Furthermore, the DuELME-GUI can be used to create a similar resource for another language.

DuELME has been populated by analysing the data records of the candidate expressions one by one using the input screen shown in Figure 4.1. The example shows the data record of the tuple *mis#boot* and the input of the lexical entry *de boot missen*. The representation guidelines can be found in Grégoire (2007d), whereas the DuELME-GUI is documented in Grégoire (2007b).

[11] http://www.mysql.com/

[12] The DuELME-GUI is available through the *TST-centrale* (HLT Agency, http://www.tst.inl.nl/).

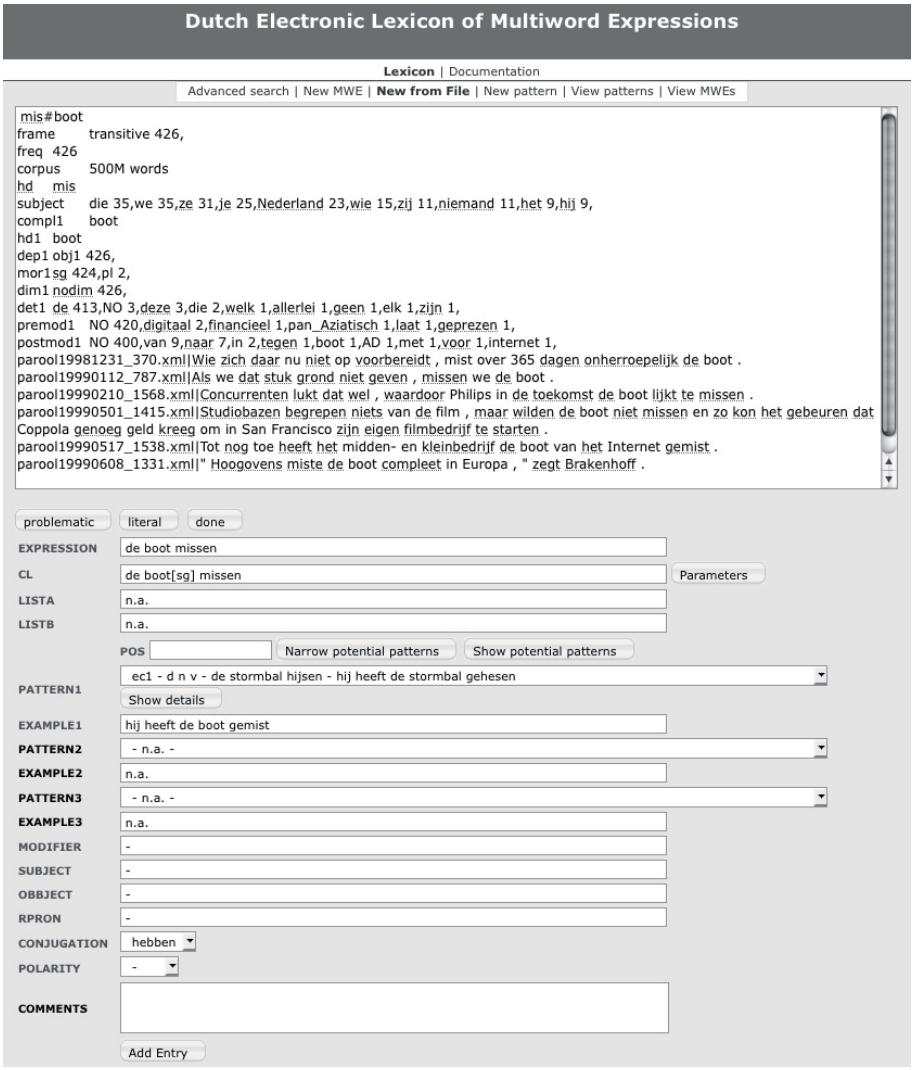


Figure 4.1 *DuELME GUI – Create new MWE screen.*

CHAPTER 5

CONCLUSION

The first part of this dissertation described the design, implementation and population of DuELME, a Dutch Electronic Lexicon of Multiword Expressions. DuELME has been created out of a need for a large number of lexical descriptions of Dutch MWEs organized in such a way that they can be incorporated in a wide variety of different grammatical frameworks and implementations with a minimal amount of manual effort. The result is an electronic resource that contains a total of 141 MWE pattern descriptions[1] and 4,416 MWE entries.[2]

I conclude this part with a discussion of some of the steps of the development in Section 5.1. Section 5.2 first describes the automatic conversion of DuELME into the Alpino, which is followed by discussing the effect of incorporating DuELME into the Alpino system.

[1] It should be noted that pattern descriptions have only been created if the pattern could be assigned to at least two MWE descriptions. A small number of MWEs in the lexicon have a unique pattern, and although these expressions must be analyzed properly, creating a new pattern description for each of these expressions does not contribute the main requirement of the encoding, which is that it can be converted into system specific representations with a minimal amount of manual work. For these expressions a special class has been created, viz EC70. The expressions in this class should be assigned a pattern, represented in the COMMENT-field. It should be noted, however, that the expressions with the PATTERN_NAME EC70 have not been exhaustively analyzed, and that the majority still needs to be assigned a pattern. Moreover, it may turn out that two or more expressions in this class can be assigned the same pattern, yielding the creation of a new MWE pattern description.

[2] The database contains 4,416 separate MWE entries. Taking into account verb alternation (in de LIST-fields) gives a total of 5,607 expressions including 375 MWEs that have been assigned the pattern EC70.

5.1 DISCUSSION

5.1.1 THE RESEARCH APPROACH

The goal of this work is to make a resource available that is organized and describes MWEs in such a way that it can be integrated into a wide variety of NLP systems as efficiently as possible. To achieve this goal, the approach taken concentrates on representing the core properties needed and on organizing the MWEs according to their syntactic pattern. The result is a resource that comprises both pattern descriptions describing the characteristics of a group of MWEs and descriptions of individual expressions.

The approach taken builds on the parameterized Equivalence Class Method (ECM). The ECM is an innovative approach that is based on the idea that MWEs that have the same syntactic pattern require the same treatment in NLP. So instead of assigning a syntactic structure to individual MWEs, the method specifies which MWEs have the same structure. I have enhanced the original parameterized ECM by defining a total of 26 parameters for Dutch. It was shown that introducing parameters to the ECM decreases the number of equivalence classes needed by almost 90% with respect to the number of equivalence classes needed in the original ECM. Concretely, this means that the use of parameters reduces the number of equivalence classes and increases the number of MWEs in each class, supporting the task of converting the standard format into the structure required in the target NLP system. The ability to handle parameters varies from system to system, which means that some systems will profit more from the parameterized ECM than other systems. Applications that cannot deal with certain parameters will not benefit but are also not harmed, since the original equivalence classes can still be identified.

The original ECM is already sufficient in its current form. To help speeding up the process of converting the standard representation into a system specific representation, I introduced a formal notation using dependency structures based on the format used in the CGN, which aims to be as theory-neutral as possible and has become a de facto standard for Dutch. Although no problems have been encountered with describing MWE patterns using the formal notation, it might turn out that the formal notation fails to fully cover the range of different types

of MWEs. The strength of the ECM, however, is that any expression can be included in the lexicon, regardless of whether it fits our notation, because MWEs with identical patterns can still be assigned pattern identifiers. For expressions that cannot be assigned a formal notation, because of its possible limitations, we can fall back on the original ECM.

5.1.2 SELECTION OF MWES

MWEs for the lexicon have been selected from a list of over 9,000 candidate expressions, their properties and example sentences automatically extracted from corpora. The integration of acquired lexical data in DuELME needs to be done manually, since there is no straightforward way to interpret the data automatically. The information given in a data record needs to be analyzed carefully to identify one or more MWEs and to determine the correct form of an MWE. As discussed in Chapter 3, deciding on whether a combination of words is an MWE is not always easy. It appeared to be particularly hard to decide whether certain combinations of a verb and a noun are lexically determined, i.e. unpredictable from their semantic properties, since the properties of the lexical items and the rules to combine them are not always known. More research is needed to study the co-occurence of noun-verb combinations in detail.

5.1.3 REPRESENTATION OF MWES

Manually describing over 5,000 MWEs is very time consuming. We have spent a lot of time selecting true MWEs, but most time has been spent on choosing the correct form of the MWEs selected. Considerable effort has been expended on selecting the correct determiner variation and on choosing the right pattern, especially with respect to modifiability. The decisions made are merely based on information represented in the data records.

The more idiomatic, i.e. semantically opaque, the expression, the more confident we can be about how to represent the MWE. Deciding on the correct form is more complicated with more transparant MWEs, especially with noun-verb combinations in which the noun is used more or less literally. These types of combinations occur frequently in one form, which often includes a fixed determiner, usually the def-

inite article *de* or *het*. These combinations seem unmodifiable or limitedly modifiable and the first intuitions we have about the form of the MWE are often confirmed by the information in the data records. However, what became clear when analysing a selection of expressions in more detail, i.e. examining all examples sentences extracted, is that some of these expressions are in many cases more flexible than assumed during the first analysis.[3] In the next part I will elaborate on this point and examine in detail the corpus data of a number of MWEs to get a better understanding of their variation potential.

5.2 INCORPORATION INTO THE ALPINO SYSTEM

DuELME has been evaluated by testing whether it can be successfully used for the purpose it was developed for, viz. the semi-automatic incorporation of the lexical representations into NLP systems. We extensively studied the way the Rosetta MT system (Rosetta, 1994) deals with MWEs and moreover what is needed for the incorporation of the parameterized ECM in Rosetta. A conversion procedure has been described in detail in Grégoire (2007e), but could unfortunately not be tested in practice. The incorporation of a part of DuELME into Alpino has been tested in theory and in practice.

Alpino is a dependency parser for Dutch, which uses linguistic knowledge and various heuristics to construct appropriate linguistic structures of Dutch sentences. Although DuELME contains both verbal MWEs and non-verbal MWEs, the conversion to the Alpino lexicon includes just verbal MWEs.

Contrary to how MWEs are dealt with in the Rosetta system (see also Section 2.3), MWEs in Alpino are explicitly listed in the lexicon as being fixed complements of the verb with the label *fixed* followed by the obligatory lexically fixed components. A fixed complement is represented in Alpino as either:

1. fully fixed, i.e. the components are listed in their inflected form

[3] It should be noted that if a data record shows any variation, it is not clear whether the frequencies given count as variation of the MWE under consideration, i.e. the frequencies reflect the total number of occurrences of a value found for the tuple, which may also include occurrences found in literal corpus examples. An advantage of examining actual corpus data is that the examples that are not an example of the MWE under consideration can be excluded from further analysis.

and can only be used as such. For example, the MWE *iemand een rad voor ogen draaien* (id. 'throw dust in someone's eyes') is represented in the Alpino lexicon in the lexical entry for *draaien* with the complement *fixed([[voor,ogen],[een,rad],dat],imp_passive)*;[4]

2. fully flexible, i.e. only the head of the complement is represented in the non-inflected form preceded by a predefined label that specifies the dependency of the constituent, e.g. the MWE *iemand een loer draaien* (id. 'play s.o. a trick') is represented in the lexical entry for *draaien* with the complement *fixed([acc(loer),dat],imp_passive)*.[5]

The incorporation of DuELME in Alpino comprises adding new lexical entries to the Alpino lexicon. For the purpose of this test, we left the Alpino grammar untouched. Therefore only types of MWE constructions that are already present in the Alpino lexicon can be integrated.

We have converted the standard representation following the spirit of the ECM, viz. take one instance from an EC, define and formalize the conversion of this instance, and use the information gathered to automate the conversion of all other instances of the same EC. However, we took a slightly different approach than the conversion procedure proposed as part of the original ECM. Recall that in the original ECM, step one of the manual procedure prescribes to select one lexical entry from an EC and to have the examples sentence of this entry parsed by the system. Although, this step is required when converting the standard representation to the representation used in Rosetta, it is not required for a conversion to the Alpino representation. In the standard conversion procedure the parse is used to obtain the syntactic structure of the MWE and the identifiers to the lexical components as required by the target system. In Alpino, MWE structures differ from normal structures in that the subcategorization pattern and the lexical components that make up the MWE are explicitly encoded as fixed complements in the lexicon. Therefore, instead of parsing example sentences, it is more efficient to map the formal representation of the MWE pattern directly on the Alpino representation. This is possible, because the structures

[4]Where *dat* stands for *dative*, which refers in this example to a free argument that is realized as an indirect object, and *imp_passive* indicates the possibility to occur in (impersonal) passive constructions.

[5]Where *acc* stands for *accusative*, which in this example means that the lexical item *loer* is realized as the head of the direct object.

generated by the Alpino parser are dependency structures based on the CGN. In short, for each EC, a mapping has been made between the PATTERN-field taken from the MWE pattern description and the representation of this pattern as required by Alpino. The target pattern contains open slots so that in the automatic part the relevant lexical items can be extracted from the CL-field and the EXAMPLE-field of each MWE description in the EC.

To illustrate, the standard MWE pattern description for *EC1* is [.VP [.obj1:NP [.det:D (1)] [.hd:N (2)]] [.hd:V (3)]], which is described as 'expressions headed by a verb, taking a direct object consisting of a fixed determiner and an unmodifiable noun'. Since the MWE pattern contains a fixed determiner, the pattern must be interpreted as fully fixed, and accordingly be converted to the Alpino representation, which is *fixed([[determiner,noun]],norm passive)*.[6] In the automatic part the determiner and noun slots in the target pattern are filled with the corresponding lexical components extracted from the EXAMPLE-field using the MAPPING-field. Furthermore, the verb is extracted to determine the lexical entry in which the target pattern must be represented in the Alpino lexicon. For example, the EXAMPLE-field of *de boot missen* is *hij heeft de boot gemist*, hence in the target pattern 'determiner' is substituted by *de* and 'noun' is substituted by *boot*,[7] and *gemist* is the verb, i.e. the head of the lexical entry.

The output of the conversion is basically a new lexicon that includes the original Alpino lexicon extended with the verbal MWEs from DuELME. The implementation of DuELME in Alpino has been described exhaustively in Grégoire (2007c).

The assessment of the effect of incorporating the standard into Alpino has been reported in Villada Moirón (2007). The evaluation that has been carried out is rather small but nonetheless promising. A sample of 10 sentences without an MWE selected from the TwNC and a sample of 100 sentences with an MWE extracted from DuELME have been used

[6]Since passive information is not included in the standard representation, *norm_passive* ('normal passive') is used in the target patterns.

[7]The MWE *de boot missen* is represented in DuELME with the paramater [sg], which means that it can only be used in the singular form. For nouns that can both be used in the singular and in the plural form, two target patterns must be created in Alpino, one containing the singular form of the noun and another one containing the plural form.

to test the accuracy of the parser for both the original Alpino lexicon and the Alpino lexicon extended with verbal MWEs from DuELME. The sentences have been assigned a manually created parse to serve as a reference parse for the evaluation.

The sentences have been parsed both with the original Alpino lexicon and with the extended lexicon. Given that the extended lexicon contains more lexical entries for MWEs, it is expected that when Alpino uses the extended lexicon, more sentences with MWEs are correctly analysed than when Alpino uses the original lexicon. We furthermore expect that Alpino will not perform worse with the extended lexicon when analysing the sentences without an MWE.

To measure the accuracy of the analyses returned by the parser, the *concept accuracy per sentence* has been computed as proposed in van Noord (2006) by comparing the parsed sentences with the manually created reference parses. The higher the concept accuracy the better the performance of the parser. Table 5.1 shows the concept accuracy per sentence for both sets of sample sentences using two different lexica. The results show that the concept accuracy of sentences that contain an MWE improves substantially when using the extended lexicon. Moreover, the concept accuracy of sentences without an MWE has not decreased with the extended lexicon. For a detailed description of the method and an overview of quantitative results see Villada Moirón (2007).

Sample	**Lexicon**	**CA**
MWEs	Alpino lexicon	82.849
	Extended lexicon	94.088
Non-MWEs	Alpino lexicon	95.833
	Extended lexicon	96.389

Table 5.1 *Concept accuracy (CA) scores (Villada Moirón, 2007).*

Part II

A CORPUS-BASED STUDY OF IDIOM VARIATION

CHAPTER 6

INTRODUCTION

Although DuELME has been designed in such a way that MWEs with a varying degree of variation potential can be represented, no detailed study on the linguistic behaviour of MWEs has been carried out during its development process. Some MWEs tend to be more restricted in their use than others, i.e. some MWEs do not allow all the combinations and transformations that a literal use would, while others behave more as literal expressions in that they allow a considerable degree of variation. An example of the former type of MWEs is the oft-cited English expression *kick the bucket*, which loses its non-literal interpretation inter alia when it is passivized, when the NP is topicalized, and when the noun is modified, see the examples in (24)-(26) respectively.

(24) # The bucket was kicked by John.

(25) # The bucket, John kicked last night.

(26) # John kicked the rusty bucket.

By contrast, the expression *spill the beans* is less restricted in its use, as illustrated in (27) and (28).

(27) The beans have been spilled by John.

(28) John spilled the well-kept beans.

To improve successful treatment of MWEs in NLP a better understanding of potential variation is necessary, which is precisely the objective of the second part of this dissertation.

To achieve this goal, I take a corpus-based approach, i.e. actual usage data is used as empirical material to test the central claim of this study. Based on observations made in related work (cf. e.g. Nunberg et al. (1994)), I claim that the variation potential of MWEs depends on whether parts of the MWE have identifiable idiomatic referents. Assumptions and hypotheses underlying this claim are outlined in a theoretical account of MWE variation. Although the theory presented is claimed to be applicable to MWEs in general, the focus of this study is in particular on Dutch direct object - verb idioms, i.e. direct object - verb (OBJ1-V) combinations of which at least the noun is not used literally. Examples of such idioms are *de boot missen* ('to miss the boat'), *de benen nemen* (lit. 'to take the legs', id. 'to escape') and *de stormbal hijsen* (lit. 'to hoist the storm cone', id. 'to warn').

Part II is structured as follows. In this chapter, I first present some theoretical background and discuss the terminology employed in this study (Section 6.1). Section 6.2 introduces the approach taken and discusses its novelty compared to related work. Section 6.3 provides a theoretical basis for the variation potential of idioms. The hypothesis formulated in that section will be tested for a number of Dutch idioms in Chapter 7. This part ends with a conclusion in Chapter 8.

6.1 THEORETICAL BACKGROUND AND TERMINOLOGY

In this section I define some notions that are used throughout this study and discuss briefly the main views on the linguistic behaviour of idioms.

6.1.1 IDIOMATIC MEANING VS. LITERAL MEANING

An idiom is a special type of MWE which has an idiosyncratic meaning. We call this meaning the *idiomatic meaning* of the expression.[1] An expression that is an idiom can also have (and usually has) a *literal meaning*, which can be defined as the meaning of the expression that is computed compositionally from the meanings of its parts (outside of the idiomatic expression) and the way they are combined. An idiom

[1] Other terms that are employed in the literature to refer to the non-literal meaning of idioms are, inter alia, *actual meaning* (cf. Dobrovol'skij and Piirainen (2005)) and *figurative meaning*.

does not have a literal meaning if no such meaning can be derived compositionally, viz. if (1) any of the parts does not have a meaning (e.g. *blow the **gaff***); (2) the parts are combined in an ungrammatical way (e.g. *trip the light fantastic*); or (3) the combination is semantically ill-formed, i.e. the combination has an unrealistic or even impossible meaning (e.g. *beat a hasty retreat*).

An idiom's idiomatic meaning deviates from its literal meaning in that it does not automatically follow from the meaning of the individual parts as used outside of the idiom, but simply needs to be stipulated as a property of the combination as a whole.[2]

6.1.2 IDIOM VARIANT AND VARIATION

Recall that one of the steps in the development of DuELME was selecting true MWEs and deciding on their precise form. One of the decisions that had to be made was whether different examples belong to the same MWE or lead to the creation of new entries. For the OBJ1-V idioms examined in this part of the dissertation, *idiom variants* are considered as realisations of the same idiom if they are made up of identical obligatory lexical components that occupy corresponding positions in the syntactic structure. According to these criteria, variants with local differences, such as inflection of the noun (e.g. *miss the boat* and *miss the boats*), are classified as instances of the same idiom. On the other hand, the expressions *hit the hay* and *hit the sack* are not variants of the same idiom, because they do not contain the same obligatory lexical components.

In order to be able to refer to a class of idiom variants, it is useful to assign a label to each class. In principle, this label can be anything as long as one knows which class it denotes. Both Moon (1998) and Langlotz (2006) use the idiom's citation form in dictionaries, which is often assumed to be the form that is recognized by speakers as the most neutral form. Alternatively, Riehemann (2001) and Stathi (2008) choose the most frequent form of the idiom in the corpus, modulo the inflection of the head. In this study, I take the form that is represented in the EXPRESSION-field in DuELME (see Section 4.2.2) to refer to a class

[2]Note that this does not imply that the individual parts cannot be assigned a non-literal meaning or that there is no relation between the literal meaning of the parts and the idiom's idiomatic meaning (cf. Section 6.3.2).

of idiom variants, for example the label for variants that contain the obligatory lexical components *boot* ('boat') and *missen* ('miss') is *de boot missen*.

Idiom variation is defined as the type of change that is reflected by an idiom variant. Different kinds of variation can be distinguished, see Moon (1998) and Langlotz (2006) for an overview. In the present study, I focus on a selection of types of lexico-grammatical variation, which includes morpho-syntactic variation, such as determiner alternation and inflection of the noun; syntactic variation, such as topicalization, pronominal reference and passivization; and adnominal modification.

6.1.3 COMPOSITIONALITY AND DECOMPOSABILITY

Although in recent research it is generally accepted that idiom variation cannot be accounted for without involving semantics, this has not always been the case. In the 1970's, the dominant view was that idioms must be regarded as non-compositional units. With generative grammar predominating linguistic research, syntax was considered as the principal component of linguistic structure, while the meaning of a grammatical construction followed from the principle of compositionality. According to this principle, the meaning of an expression is formed by combining the conventionalized meaning of the expression's parts, which leads to the literal meaning of the expression, but not to the idiomatic one. This implies that the idiomatic meaning of an idiom cannot follow from the principle of compositionality and hence idioms were either neglected in generative accounts, or marked as exceptions and treated as complex phrases with no internal semantics that should be stored in the same way as other lexical items. Accordingly, transformational deficiency, i.e. the inability of idioms to allow all grammatically possible transformations, was not treated as following from general principles, but as an irregularity that needs to be stipulated in the lexicon (cf. among others Weinreich (1967); Fraser (1970); Katz (1973)).

In the same period, Chafe (1968) argues against the purely descriptive proposals coming from syntax-based approaches and argues in favor of an explanatory treatment of syntactic variability based on a semantic approach. In subsequent studies (inter alia Nunberg (1978); Wasow et al. (1983); Nunberg et al. (1994)), the idea that idiom semantics

might be responsible for idiom variation is expounded. In his thesis, Nunberg (1978) asks how native English speakers know that a sentence such as *the beans have been spilled* has an acceptable idiomatic reading while a sentence such as *the bucket was kicked* does not, even if they have not encountered any of these forms of the idioms before. Instead of regarding idioms as non-compositional, Nunberg (1978) bases his answer on the observation that idiom parts can contribute to the interpretation of the idiom as a whole. Nunberg introduces the term *decomposability*, which he defines as follows:

> "Let us say that verb phrases "refer" to states and activities, and that transitive verb phrases normally refer to states and activities that are best identified as "open relations" of the form Rxb, where "R" stands for the relation referred to by the verb, "x" is a variable for the referent of the sentence subject, and "b" stands for the referent of the object NP.[...] Then we will say that an idiomatic transitive VP is DECOMPOSABLE just in case it is used to refer to a state or activity such that it would be normally believed that that activity could be identified as an open relation Rxb, such that the object NP of the idiom refers to *b*, and the verb to R." (Nunberg, 1978, p. 125)

Nunberg distinguishes between *decomposable* and *non-decomposable* idioms. An example of the former type is *spill the beans* which denotes the event of revealing hidden information and which is decomposable because each part of the idiom refers to an element in its denotation: *spill* refers to 'reveal' and *beans* refers to 'hidden information'. On the other hand, the idiom *kick the bucket*, which refers to the event of dying suddenly, is non-decomposable, because there is no element in the denotation to which the individual parts refer.[3] Nunberg argues that

[3]In fact, Nunberg (1978) proposes a third category of idioms, which he refers to as *abnormally decomposable* idioms, and which differ from normally decomposable idioms in that the object NP of abnormally decomposable idioms does not itself refer to some component of the idiomatic denotation, but only to a (metaphorical) relation that is typically used to identify the idiom's idiomatic denotation. Although, the distinction he makes may be useful to account for some of the lexico-grammatical variation of idioms, the primary focus of this study is on the decomposable versus non-decomposable dichotomy. Nevertheless, I will address the phenomenon of 'abnormally' decomposable in the interpretation of the corpus data (see Section 7.4).

solely the individual parts of decomposable idioms can be focused and hence are subject to types of variation that involves focusing.

For some reason, Nunberg et al. (1994) abstract away from the term *decomposability* and instead they introduce a distinction between semantically compositional *idiomatically combining expressions* and semantically non-compositional *idiomatic phrases*, which are both defined in terms of meaning. Idiomatic phrases are defined as expressions the meaning of which cannot be distributed over the individual parts and hence are not subject to syntactic flexibility. On the contrary, idiomatically combining expressions are idioms whose parts carry identifiable parts of their idiomatic meaning and tend to be syntactically flexible to some degree. Compositionality used in this way must be regarded as an *a posteriori* process, i.e. "speakers are capable of recognizing the compositionality of a phrase like *spill the beans* after the fact, having first divined its meaning on the basis of contextual cues." (Nunberg et al., 1994, p. 499).

Besides the term *decomposability* (cf. Sag et al. (2001); Riehemann (2001)) and the terminology introduced by Nunberg et al. (1994), other terms that are employed in the literature are *isomorphism* (e.g. Van der Linden (1993); Geeraerts (1995); Langlotz (2006)) and *analysability* (e.g. Langlotz (2006); Stathi (2008)). There is however not much consensus in the literature about how to use these terms. Especially the term *analysability* is confusing since it is often not clear whether it is used in the sense of decomposability as the degree to which an independent idiomatic meaning can be assigned to the individual parts, or in the sense of whether the idiom's overall meaning is semantically related to the literal meaning of the individual components (see Moreno (2007) for a detailed discussion on this topic).

6.1.4 IDIOMATIC REFERENT

The concept of decomposability is these days regarded as an important determiner for the variation potential of idioms and forms the basis of the theory presented here. However, in this study, I prefer to use a different terminology viz. *idiomatic referentiality*. One reason is that, different from how it is defined by Nunberg (1978), the term *decomposability* is nowadays defined in terms of meaning and relates the linguistic behaviour of idioms to whether the idiom's idiomatic mean-

ing can be decomposed into the meaning of its parts. However, it can be argued that not finding a meaning that can be distributed over the idiom parts does not mean that there is no such meaning. For example, *kick the bucket* is taken to mean 'die (suddenly)', which is a one-place predicate. Although it is difficult to see how *kick the bucket* could be paraphrased as a two-place predicate, this is not because the situation of someone dying can only be described as a one-place predicate (see Nunberg (1978) where *kick the bucket* is compared with the expression *give up the ghost*).

Another reason for using a different terminology is that, especially in psycholinguistic research, the term *decomposability* is often confused (or used interchangeable) with the term *analysability*, which is mainly used as the extent to which the literal meaning of the idiom parts contributes to the idiom's idiomatic meaning. To avoid confusion with the term *decomposability*, the theory presented here is build on the concept of *idiomatic referentiality*.

As put forward by Nunberg (1978), verb phrase idioms, just as literal verb phrases, refer to (or denote) events, such as states and activities.[4] Given an idiom's denotation, I consider that an idiom part has an *idiomatic referent* if it refers to an element in the idiom's denotation. In this sense, the individual parts of *kick the bucket* do not have an idiomatic referent, because there is no element in its denotation ('die suddenly') to which they can refer, hence the semantic representation is mapped to the idiom as a whole. On the other hand, each part of the idiom *spill the beans* refers to an element in the denotation of the idiom (*spill* denotes 'reveal' and *beans* denotes 'hidden information'), hence both parts are said to have an idiomatic referent. It can be argued that this approach does not differ from an approach in terms of meaning, because it still depends on the denotation assigned to the idiom. However, what is important here is not what the idiomatic referent is, but whether an idiomatic referent can be identified. This point will be elaborated in the next chapter.

[4]In this study, I use the terms *reference* and *denotation* interchangeable as in being a property of a linguistic expression – for instance the word *labrador* denotes/refers to a kind of dog – and not as a speech act, i.e. a relationship that is dependent on the use of the expression.

6.2 THE RESEARCH APPROACH

The aim of this study is to gain better insights into the linguistic behaviour of idioms by analyzing the actual usage of a number of Dutch OBJ1-V idioms. In this section, I will elaborate on the approach taken regarding each aspect of the study and briefly discuss its novelty compared to related work.

6.2.1 IDIOMATIC REFERENT

The theory presented is based on the claim that the variation potential of an idiom depends on whether its individual parts have an idiomatic referent. As discussed in the previous sections, this claim is not novel; it has emerged from the decomposability account as proposed by Nunberg (1978), which has been (partly) adopted by a number of scholars, among which, Van der Linden (1993); Fellbaum (1993); Geeraerts (1995); Stathi (2008). Both Stathi (2008) and Van der Linden (1993) claim that if an idiom part has an idiomatic referent, the use of the idiom is potentially unlimited. If there are any restrictions on its use, it can be explained on the basis of properties that are not necessarily specific to idioms (Van der Linden, 1993, p. 46). However, the extent to which the presence of identifiable idiomatic referents is responsible for the linguistic behaviour of idioms has not yet been systematically investigated for Dutch using corpus data as empirical material. Testing the claim made is the primary goal of this study, which furthermore includes examining whether other factors than the presence of idiomatic referents play a role in the variation potential of idioms.

6.2.2 CORPUS-BASED ANALYSIS

Although the theory is formulated in such a way that it accounts for MWEs is general, the corpus investigation focuses on 25 Dutch OBJ1-V idioms and in particular on idioms of which the definite article *de* or *het* ('the') specifying the idiom noun is the most frequent form in the corpus. The data analysis and interpretation are primarily based on usage data extracted from corpora, in particular the *Twente Nieuws Corpus* (TwNC, Ordelman (2002)),[5] a 400-million-word corpus of newspaper

[5] http://hmi.ewi.utwente.nl/twnc/

texts.

To my knowledge, no corpus-based analysis of idiom variation has been conducted for Dutch. Corpus-based analyses of idioms are, inter alia, available for English (Moon, 1998; Riehemann, 2001; Langlotz, 2006) and for German (Stathi, 2008). With respect to the size of the corpora used, Moon (1998) used a 18-million-word corpus, Riehemann (2001) a 350-million-word corpus, Langlotz (2006) a 100-million-word corpus, and Stathi (2008) a 1-billion-word corpus.

Of these related studies only Riehemann (2001) actually assesses the decomposability view. She focuses mainly on OBJ1-V idioms, which she has classified as decomposable or non-decomposable before conducting the corpus study by formulating paraphrases for the expressions. Riehemann (2001) admits that it is not straightforward to classify all idioms correctly on the basis of paraphrases. She doubts whether there are only two clearly separable categories, but does not elaborate on this statement. Although the analysis has been described exhaustively including many corpus examples to illustrate the findings, the report lacks an interpretation of the overall results.

6.2.3 CORPUS DATA: THE PROS AND CONS

Much has been written on the use of corpus data to test linguistic theories. Where some scholars claim that the use of corpora is not only necessary but also sufficient (e.g. Sinclair (1991) and his followers), others argue that one needs a corpus, but also constructed examples (see e.g. Fillmore (1992); Pullum (2009)).

Corpora provide evidence of language in use reflecting actual usage of a wide variety of speakers, and hence exhibiting phenomena that a single linguist using constructed examples might not be able to come up with (cf. van Noord and Bouma (2009) who illustrate how parsed corpora can be helpful to find new empirical evidence for fairly complicated and subtle linguistic issues). Moreover, corpora are extremely suitable for collecting large amounts of data. Especially corpora that have been annotated with linguistic information, such as e.g. syntactic relations, provide the opportunity to systematically search for certain constructions.

Context plays an important role in the acceptability judgements of data; where constructed examples are often short and may sound pe-

culiar in isolation, corpus data can be studied in broad context, which may be of help when judging a construction. With respect to studying idioms, context can also help to determine whether an expression is used in its literal meaning or in its idiomatic meaning, and moreover to postulate plausible meaning(s).

Not only is a corpus suitable for studying constructions in their context, corpus data can also be taken as an example for constructing test sentences. Creating plausible test sentences is not always easy: one needs to make sure that one only tests the construction examined, and that other factors do not play a role in the well-/ill-formedness of the example. For instance, some linguistic phenomenon may seem ill-formed in a constructed example, but nevertheless occurs in corpora (cf. van Noord and Bouma (2009)). In that case other factors than the phenomenon examined may cause the constructed example to be judged as ill-formed. Basing test sentences on corpus examples can contribute to the naturalness and plausibility of a sentence.

To conclude, corpora can be considered as a very useful source for linguistic research. However, they also have their limitations.

First of all, corpora are not more than large samples of language use. Corpora provide independent data to test a theory, but if a phenomenon is not found in the corpus it does not mean that it cannot occur in language at all. Moreover, if a phenomenon is found in the corpus, it does not necessarily mean that it is a well-formed part of the language.

Second, although annotated corpora can be useful to systematically search certain constructions, annotation errors both lead to noise in the data extracted and to loss of information, i.e. wrongly annotated sentences are either part of the data extracted, while they should not; or they have not been detected, while they should have been.

In this study we make use of a news corpus. Since newspaper articles are often modified forms of official press releases, many sentences occur repeatedly in the corpus. This should be taken into account when referring to the frequencies of a certain construction. Furthermore, one should take into account the source of the example, which is usually the body of a news article, but may also be a header or for instance a poem, which may form a source of wordplay and which may require special treatment.

With respect to the idioms in this study, the frequency overviews

(as presented in the next chapter) show that each specific idiom occurs relatively infrequently in the data (cf. Moon (1998)). The average number of corpus examples for the 25 idioms examined in this study is 414 hits (based on the 400-million-word *Twente Nieuws Corpus* (*TwnC*)). The majority, viz. 390 out of 414 examples, are in just one form and do not show any of the variation types studied here, despite the fact that the examples have been extracted from a news corpus. As observed by Moon (1998), idioms occur more frequently in the journalism genre than in any other genre either written or spoken. If we compare the total number of 10,346 idiom examples found in the *TwNC*, which only consists of newspaper texts, with the numbers found in the *D-coi* corpus (Dutch Language Corpus Initiative),[6] and the *Corpus Gesproken Nederlands* ('Corpus of Spoken Dutch', (CGN)),[7] we can confirm this observation: a total of 98 examples have been found in a subpart of the *D-coi* corpus, which includes 15 million words taken from magazines, books, brochures, proceedings, etc., whereas only 37 examples have been found in the *CGN*, a 9-million-word corpus. This means that there are 4 times more examples in the *TwNC* than in part of the *D-coi* corpus, and even 6.3 times more than in the *CGN*.

Idioms not only occur more often in one genre than in another genre, but frequencies may also differ from domain to domain. Although I do not have any numbers, some idioms occur notably more frequently in the sports domain than in any other domain. The domain in which an idiom is used may also influence its meaning, e.g. although the idiom *de geest geven* generally means 'to die', when used in the sports domain it merely means 'to give up'.

Given these limitations, one cannot and should not solely rely on corpus data to support one's theory, and moreover the corpus examples that have been extracted should be examined thoroughly to verify their usability and well-formedness. In this study, the focus is on testing the hypothesis with corpus data. However, because of the small amount of variation in the data, constructed examples will be used to give a more complete picture of potential idiom variation. The creation of test sentences and their role in the interpretation of the data will be further discussed in Section 7.2.2.

[6] http://lands.let.ru.nl/projects/d-coi/
[7] http://tst.inl.nl/cgndocs/doc_Dutch/start.htm

6.2.4 IDIOMATIC MEANING

It is not trivial to determine whether the individual parts of an idiom have identifiable idiomatic referents. One must know the idiom's idiomatic meaning, but meaning cannot straightforwardly be extracted from corpora. It is not easy to find independent evidence for an idiom's idiomatic meaning. In this dissertation, the idiomatic meaning will be established on the basis of corpus examples after first having categorized the idioms in two groups, viz. (1) idioms the parts of which most probably have an idiomatic referent, and (2) idioms the parts of which are more likely to have no idiomatic referent. In order to make this distinction, I choose one consequence of idiom parts having idiomatic referents as a heuristic for idiom parts actually having idiomatic referents. This initial grouping of idioms forms the basis of the analysis and interpretation of the results. The determination of an idiom's idiomatic meaning is elaborated on in Section 6.3.2.

6.2.5 VARIATION

Lexico-grammatical variation The theory presented only accounts for lexico-grammatical variation, which includes morpho-syntactic variation, such as determiner alternation and inflection of the noun; syntactic variation, such as topicalization, pronominal reference and passivization; and adnominal modification (see the next chapter for details). Other types of variation, such as lexical flexibility of the idiom components (see, inter alia, (Moon, 1998; Langlotz, 2006; Stathi, 2008), will not be studied.

Wordplay I would like to note that the theory presented here accounts for systematic idiom variation, which should be distinguished from so-called *wordplay*. Though no generally accepted definition of the term *wordplay* exists, it is often characterized as a creative use of words to achieve a special communicative effect, e.g. humorous, critical, dramatic, etc. (cf. inter alia Schenk (1994); Langlotz (2006)). In an attempt to differentiate these creative uses from systematic variation, I regard an idiom variant that is the result of a grammatical and regular adaption of the idiom's default form without being intended to create a striking stylistic effect as an example of systematic variation. How-

ever, the distinction between systematic variation and wordplay is not always clear-cut and one should be aware not to judge an idiom variant as wordplay, simply because it does not fit one's theory.

This concludes the discussion of the research approach. In the next section I present the Idiom Variation Potential Hypothesis.

6.3 TOWARDS A THEORETICAL ACCOUNT OF IDIOM VARIATION

This section consists of two parts. I start with introducing the Idiom Variation Potential Hypothesis in Section 6.3.1, which constitutes the basis for the empirical study conducted in the next chapter. In order to make the theory easy to follow, the various possible outcomes of the hypothesis are visualised in so-called concept mappings, which will be elaborated in Section 6.3.2.

6.3.1 IDIOM VARIATION POTENTIAL HYPOTHESIS

As stated, I investigate the claim that an idiom's variation potential depends on whether the idiom parts have idiomatic referents. In order to be able to test this claim, I propose the Idiom Variation Potential Hypothesis, which is formulated in (29).

(29) **Idiom Variation Potential Hypothesis**

a. If an idiom part has an idiomatic referent,
then its variation potential is in principle unrestricted.

b. If an idiom part has no idiomatic referent,
then variation that requires an idiomatic referent is blocked.

Given the hypothesis, two statements need to be elaborated, viz. "in principle" and "variation that requires an idiomatic referent".

I start with the latter statement. What can be inferred from the hypothesis in (29) is that two types of variation are distinguished, viz. (1) variation that only applies to parts that have an idiomatic referent, and (2) variation that does not depend on whether an idiomatic part has an idiomatic referent. An idiomatic part that has an idiomatic referent can

be said to be meaningful, i.e. the idiom part is not semantically empty, because it can be assigned an idiomatic meaning. Hence, in order to determine what types of variation require an idiomatic referent, it should be determined what types of variation require a constituent that has meaning. An example of variation that generally requires constituents to have meaning is modification. In one common type of modification the combination of a modifier and a modifiee restricts the domain that is denoted by the modifiee alone to a subpart of this domain, which is not possible if the modifiee is not denoting, i.e. has no meaning. For example, the noun *book* denotes all the entities that are a book, whereas the combination *beautiful book* denotes all the entities that are a book and that are beautiful.

To summarize, in order for modification to be grammatically plausible, the modified constituent must have meaning. If we apply this condition to the idiom *kick the bucket*, for which I assume that *bucket* has no idiomatic meaning, we must conclude that *bucket* cannot be modified without losing the idiomatic interpretation of the idiom. In other words, the sentence in (26) (for convenience repeated here as (30)) cannot be interpreted idiomatically.

(30) # John kicked the rusty bucket.

On the other hand, the sentence in (28), here repeated as (31), can be interpreted idiomatically, since it is generally assumed that *beans* has the idiomatic meaning 'secrets' and hence the constituent *well-kept beans* can be interpreted as 'well-kept secrets'.

(31) John spilled the well-kept beans.

Other types of variation that require a constituent to have meaning are topicalization, e.g. (32); pronominal reference, e.g. (33); variation in the number of the noun (singular vs. plural) and the use of diminutives; and the use of various types of determiners, such as demonstratives, quantifiers, etc.

(32) That book, he did not read.

(33) He wanted to read that book$_i$, but he could not find it$_i$.

It is beyond the scope of this study and beyond its purpose to elaborate on the distinction between variation that requires constituents to

have meaning and other types of variation. For independent evidence for the existence of this distinction I refer to Odijk (1993) and Schenk (1995).

It should be noted that the modification example given in (31) is an example of so-called *internal modification*, a term introduced by Ernst (1981), who argues that a distinction should be made between *internal modification* and *external modification*.[8] In the case of internal modification, the insertion of an adjective does not only syntactically modify the noun, but also semantically. On the other hand, in the case of external modification, the adjective may syntactically modify the noun, but semantically it is the whole idiom that is modified. Ernst illustrates external modification and internal modification with the examples (34) and (35) repectively (Ernst, 1981, p. 50–51).

(34) With that dumb remark at the party last night, I really kicked the social bucket.

(35) In spite of its conservatism, many people were eager to jump on the horse-drawn Reagan bandwagon.

Since *bucket* in (34) does not have an identifiable idiomatic meaning, the adjective *social* must not be interpreted as modifying *bucket*, but as modifying the meaning of the idiom as a unit, which can be paraphrased as 'Socially, I kicked the bucket' (Ernst, 1981, p. 50). Ernst argues that adjectives that typically occur as external modification are so-called *domain delimiters*, such as economic, political, sociological, etc.

In (35) the adjective *horse-drawn* modifies the noun *bandwagon* both syntactically and semantically. According to Ernst, *horse-drawn* must be interpreted idiomatically in relation to *bandwagon*, meaning that Reagan's political movement is old-fashioned and behind the times.

Nunberg et al. (1994) adopt the dichotomy proposed by Ernst (1981) and moreover note that the external modification phenomenon is not restricted to idioms. This is illustrated with the example in (36a), where the adjective *quick* cannot be interpreted as restricting the meaning of the nominal constituent *cup of coffee* and hence must be interpreted as in (36b).

[8]There are also other types of modification, such as meta-linguistic modification (e.g. *kick the proverbial bucket*). However, since no examples of types other than internal and external modification have been detected in the data, the other types will be ignored here.

(36) a. They drank a quick cup of coffee.
b. They quickly drank a cup of coffee.

Nunberg et al. (1994) furthermore argue that the distinction between internal and external modification seems clear in principle, but is not always easy to make in practice. For example, the phrase *leave no legal stone unturned* can have the interpretation "legally, leave no stone unturned", but it can also have the interpretation "use all legal methods" (Nunberg et al., 1994, p. 500).

Only variation that is assumed to require a constituent that has meaning will be tested systematically in this study. Of course, in order to be able to distinguish between internal and external modification, each corpus example that contains modification will be examined separately.

As observed in the literature (see Odijk (1993); Schenk (1994)) restrictions on passivization do not depend on whether a constituent has meaning, but should be accounted for by independent rules of a language (which may not always be well-understood). The Idiom Variation Potential Hypothesis predicts that if an idiom part has an idiomatic referent, then all types of variation are in principle allowed, but this might not hold for passivization. In this study passivization is approached as being a type of variation that does not necessarily require an idiom part to have an idiomatic referent and hence will be discussed in a separate section (see Section 7.3).

According to the hypothesis, an idiom's variation potential is *in principle* unrestrained if the idiom parts have an idiomatic referent. This point will be illustrated with the Dutch idiom *de boot missen* for which I assume that *boot* has an idiomatic referent that can be described with the concept MOGELIJKHEID ('opportunity') and *missen* has an idiomatic referent that can be described with the concept MISLOPEN ('fail to catch'),[9] which is equal to one of the literal senses of the verb *missen*. Since the idiom parts have an idiomatic referent, it is subject to several types of variation including variation that requires the parts to have idiomatic

[9]Since most idiom parts cannot easily be paraphrased by other words or expressions, inter alia because other words and expressions can have other connotations or have lexical restrictions, idiom parts with an idiomatic referent are described in terms of concepts.

referents. Examples of number alternation and determiner alternation, and determiner alternation and topicalization are given in (37) and (38) respectively.[10]

(37) Erg jammer, want nu missen we allerlei boten.
very pity, because now miss we various boats
id. 'Too bad, because now we miss various opportunities'

(38) Die boot willen de andere educatieve uitgeverijen niet
that boat want the other educational publishers not
missen.
miss
id. 'That opportunity, the other educational publishers do not want to miss.'

The hypothesis states "in principle", because an idiom variant must adhere to the general principles of grammar, pragmatics and discourse, just as any literal construction. In many cases the impossibility of an idiom variant can be explained by independent general principles and therefore does not necessarily violate the hypothesis.

For example, since *boot* has an idiomatic referent, it is expected that variation of the determiner is unrestrained. However, the sentence in (39), where *boot* is specified by the possessive NP *Jans*, cannot be interpreted in its idiomatic meaning, though it is literally well-formed. This fact can be explained as follows: The variant has lost its idiomatic interpretation due to the specifics of the idiomatic referent MOGELIJKHEID; one cannot miss someone else's opportunity, cf. (40), hence *boot* cannot be specified by a possessive NP in this context, i.e. when combined with *missen*, without losing its idiomatic meaning.

(39) # Bob heeft uiteindelijk *Jans* boot gemist.
Bob has in-the-end Jan's boat missed
lit. 'In the end, Bob has missed Jan's boat.'

(40) * Bob heeft uiteindelijk *Jans* mogelijkheid gemist.
Bob has in-the-end Jan's opportunity missed
'In the end, Bob has missed Jan's opportunity.'

[10]The source of each corpus example and of examples taken from the internet are listed in Appendix C.

This concludes the discussion of the Idiom Variation Potential Hypothesis. In the next section, I propose a clear representation that can be used to illustrate the connections between an idiom and both its literal interpretation and its idiomatic interpretation in a way that the idiom's variation potential can easily be deduced.

6.3.2 CONCEPT MAPPING

In order to employ the Idiom Variation Potential Hypothesis, the following information must be available for each idiom:

- whether or not the combination has a literal meaning;
- if it has a literal meaning: the individual properties of the lexical items in their literal meaning;
- whether or not the idiom parts have an idiomatic referent;
- if the idiom parts have an idiomatic referent: the individual properties of these referents.

Based on this list, it can be concluded that on the one hand an expression relates to a so-called *literal domain* that constitutes the literal meaning of the parts and their linguistic properties, and on the other hand an idiom is linked to an *idiomatic domain* comprising the idiom's idiomatic meaning, which is based on the idiomatic referents of its parts if available, and the properties related to the idiomatic meaning.

In order to be able to illustrate the theory in a insightful way, I propose to use so-called *concept mappings*, in which the relation between the idiom and both its literal domain and its idiomatic domain can be represented. Since an expression either has a literal meaning or not, and since the idiom parts either have idiomatic referents or not, we can distinguish four different concept mapping templates, see the illustrations in Figure 6.1 - Figure 6.4.

Figure 6.1 shows the concept mapping template for idioms with a literal meaning and idiomatic referents for the individual parts.[11] The obligatory components of the expression are put in the center of the template, which is illustrated with the placeholders NOUN and VERB.

[11] For expository reasons, I focus on meaning and referents and do not include the linguistic properties of the individual components in the representation.

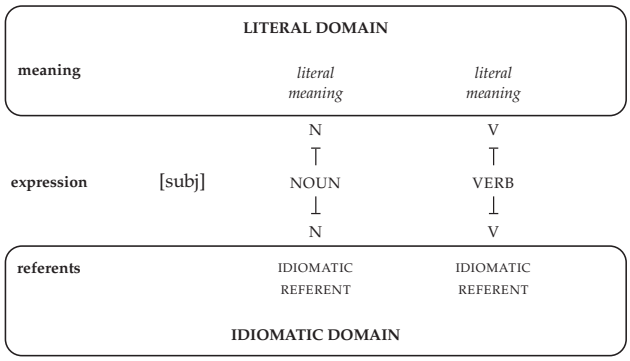


Figure 6.1 *Concept mapping template for idioms with a literal meaning and idiomatic referents for the individual parts.*

Moreover, a slot for the variable subject ([subj]) is represented for completeness sake.[12] The upper part represents the part-of-speech of the components and their relation with the literal domain. In the lower part, the relation with the idiomatic domain is represented, in this case the idiomatic referent of the individual components and the associated parts-of-speech are described.

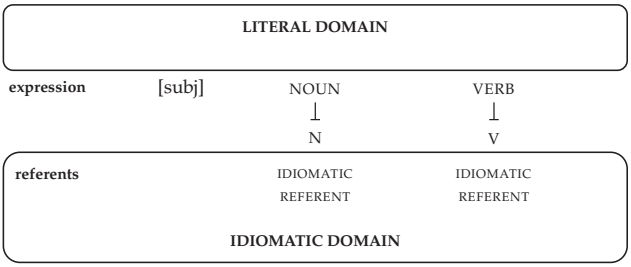


Figure 6.2 *Concept mapping template for idioms with no literal meaning, but with idiomatic referents for the individual parts.*

In Figure 6.2, the concept mapping template is illustrated for expressions with no literal meaning, which therefore have an empty literal domain, but with idiomatic referents for the individual parts.

[12]It should be noted that the subject it is not part of the idiom and its meaning is determined independently of the idiom.

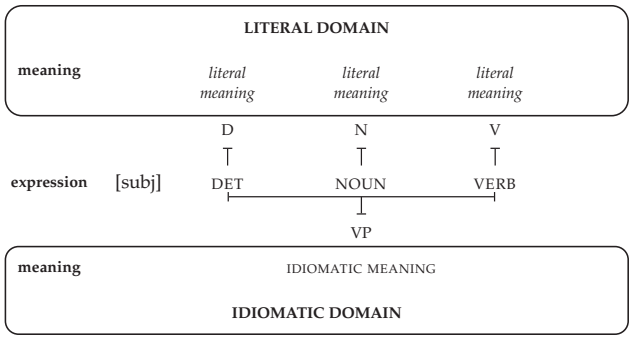


Figure 6.3 *Concept mapping template for idioms with a literal meaning, but without idiomatic referents for the individual parts.*

Figure 6.3 presents the concept mapping template for expressions with a literal meaning, but without idiomatic referents for the individual parts. Since the individual parts have no idiomatic referent, the idiomatic meaning of the whole is represented and associated with the whole VP constituent.

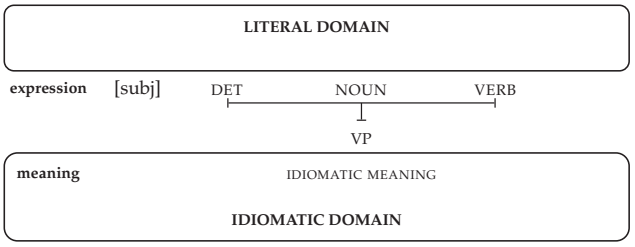


Figure 6.4 *Concept mapping template for expressions with no literal meaning and without idiomatic referents for the individual parts.*

In Figure 6.4, the concept mapping template is given for expressions with no literal meaning, and therefore an empty literal domain, and without idiomatic referents for the individual parts.

In order to represent an idiom's variation potential in a concept mapping, one must choose the appropriate template. First, it must be determined whether the expression has a literal meaning. Idioms may lack a literal meaning because (1) any of the parts does not have a literal meaning, i.e. cannot be used outside the context of the idiom; (2) the combination is syntactically ill-formed; or (3) the combination is se-

mantically ill-formed, i.e. the combination has an unrealistic or even impossible meaning. If any of these cases is applicable, the literal domain is empty. Otherwise the literal domain can be filled by specifying the literal meanings of the individual components.

After specifying the literal domain, it must be established whether the idiom parts have an idiomatic referent. One way to do this is to determine the idiom's idiomatic meaning and to verify that the individual parts refer to parts of the idiomatic meaning. However, deciding on the idiomatic interpretation is not trivial; a certain idiom is often used because its meaning contains specific nuances that cannot be expressed as effectively in other words.[13] For example, it is commonly assumed that *kick the bucket* means 'die', but where *die* can be used in a situation that describes a convicted murderer who has been given a lethal injection, it seems that *kick the bucket* can only be used in cases of a more or less sudden natural death (cf. Nunberg (1978); Stock et al. (1993)). Hence, the idiomatic meaning of an idiom is often so complex and subtle that it cannot simply be captured in a short paraphrase.

In order to determine an idiom's idiomatic meaning, one should use an empirical base that is as broad as possible, for example using dictionaries, intuitions and corpora. With the help of corpora potential idiomatic meanings can be identified by examining idioms in their context. Both the sentential context and a wider context may contain clues that can help to identify the idiom's meaning. Not only contextual clues, but also the lexical components that form the idiom may contain information that implies (part of) an idiom's idiomatic meaning. For instance, in verb phrase idioms, the verb may occur it its literal meaning. An example is the idiom *de boot missen*, in which the idiomatic meaning of the verb *missen* is the same as one of its literal senses, viz. 'fail to catch'.

An idiom component may also be semantically autonomous, i.e. possess an idiomatic meaning that can also appear outside the context of the idiom. An example of such a component is the verb *trekken* ('pull') in the idiom *de kar trekken* (lit. 'pull the cart'). The literal meaning of *trekken* in this combination is 'cause to move by pulling'. In their idiomatic meaning, *kar* refers to PROJECT ('project') and *trekken* refers to LEIDEN ('lead'). This meaning of *trekken* is not idiom-bound, but

[13] It should be noted that this is not only a characteristic of idioms.

also present in other configurations, such as *een project trekken* ('lead a project').

Literal verbs and autonomous verbs not only contribute to the determination of an idiom's idiomatic meaning, but they are also a strong indicator for the presence of idiomatic referents for idiom parts other than the verb.

Idioms may behave like metaphors. They literally refer to particular situations, events, and actions on the literal level. At the same time, these situations, events, and actions may be associated with the idiom's idiomatic meaning when interpreted on a more abstract level. For example, the idiomatic meaning of the expression *hit the panic button* is 'to panic', which can be established by associating the meaning with the event of literally hitting the panic button in certain circumstances (cf. Nunberg (1978)). This process has been described exhaustively in the literature from various perspectives, usually using the terms *transparency* and *motivation* (see inter alia Gibbs and Nayak (1989); Geeraerts (1995); Dobrovol'skij and Piirainen (2005); Stathi (2008)). It is beyond the purpose of this study to elaborate on how the idiom's literal meaning may contribute to its idiomatic meaning (which often varies from speaker to speaker (Dobrovol'skij and Piirainen, 2005)). What is important for this study is whether plausible idiomatic referents can be found for the idiom parts that are compatible with the corpus examples.

CHAPTER 7

DATA ANALYSIS AND INTERPRETATION

The goal of this chapter is to test the Idiom Variation Potential Hypothesis presented in the previous chapter for a number of Dutch OBJ1-V idioms using corpus data as the primary empirical material. The focus is on the examination of idioms of which the definite article *de* or *het* ('the') specifying the idiom noun is the most frequent form in the corpus.

This chapter starts with a description of the methodology pursued in the analysis of the data (Section 7.1). This is followed by the interpretation of the data of 25 OBJ1-V idioms (Section 7.2). Section 7.3 briefly discusses the passivization of idioms. This chapter ends with a summary and a discussion of the overall results in Section 7.4.

7.1 METHODOLOGY

This study analyses actual usage data of 25 OBJ1-V idioms. The analysis encompasses manual examination of the corpus examples, providing quantitative overviews of the results, and interpreting the outcomes in the light of the Idiom Variation Potential Hypothesis.

The data analysis and interpretation are primarily based on usage data extracted from corpora. As discussed in the previous chapter, corpora are very useful because they show facts that one may not find out in another way. However because of the small amount of variation found in the data, constructed examples will also be used to give a more complete picture of potential idiom variation (see Section 7.2.2).

In this section I first discuss the data, i.e. the idioms studied and the

corpus data. This is followed by describing the procedure of the data analysis and interpretation in Section 7.1.2.

7.1.1 DATA

This section describes the idioms studied and the corpus used to extract actual idiom examples.

Idioms The present study concentrates on 25 Dutch idioms that are headed by a verb that takes a direct object that consists of a definite article and a noun. The idioms have been randomly selected from a subset of MWEs in DuELME. This subset includes those expressions that are of the form OBJ1-V and that meet the working definition of idioms, viz. MWEs of which at least the noun has an idiosyncratic meaning.

Corpus data The initial idea was to extract the empirical material from three different corpora, viz. the *TwNC*, the *D-coi* corpus and the *CGN*. However, because of the small number of examples found in *D-coi* and the *CGN*, which moreover did not show notable variation (see Section 6.2.3), it has been decided to focus on the corpus data extracted from the *TwNC*.

The *TwNC* is the same corpus as was used to collect morpho-syntactic information for the candidate MWEs (see Section 3.1). Although the *TwNC* currently has a size of over 500 million words (including newspapers up to the year of 2008), I have used a subcorpus consisting of approximately 400 million words taken from six different Dutch newspapers of the years 1999-2004.

The set of data for each idiom has been extracted from the corpus taking a list of verb-noun pairs, e.g. *missen-boot* ('miss-boat'), as input.[1] The corpus has been syntactically annotated with the Alpino parser, which made it possible to use the syntactic relation between the verb and the noun in the search queries. The output is a list of sentences containing the idiom. As mentioned in the previous chapter, newspaper articles are often modified forms of official press releases, and therefore many sentences occur repeatedly in the corpus. For the purpose of this study, duplicate sentences have been automatically removed from

[1] I would like to thank Gertjan van Noord of the University of Groningen for conducting the search queries and providing me with the relevant corpus data.

the output list. Although this should improve the reliability of the data, it is possible that some short sentences actually do originate from multiple sources and should not have been deleted. Since the elimination was done automatically, sentences that slightly differ from each other, e.g. with different punctuation, had to be deleted manually.

7.1.2 PROCEDURE

The procedure of the corpus study consists of two parts:

1. the data analysis, i.e. examination of the examples sentences extracted from the *TwNC*, and
2. the data interpretation, i.e. generating quantitative overviews of the results for each idiom and discussing the outcomes in the light of the Idiom Variation Potential Hypothesis.

data analysis The corpus examples have been manually analysed. In the first step, each corpus example has been examined as to whether:

- it is an example of a literal use of the combination. See (41) for a literal example of the expression *de boot missen*.

  (41) Zoals vorig jaar, toen ze de boot van Calais naar
  like last year when they the boat from Calais to
  Dover hadden gemist.
  Dover had missed
  lit. 'Just as last year, when they had missed the boat from Calais to Dover.'

- it is an example of a paper headline, a title (of e.g. a book or movie), etc. and hence does not actually represent the normal use of the expression. See (42) for an example.

  (42) Op pagina 12: Heinz, Campell en Diageo missen boot.
  'On page 12: Heinz, Campell en Diageo miss boat.'

- it is an example of an idiom other than the one studied, but with the same noun and verb. An example is the idiom *iemand de duimschroeven aandraaien* ('to tighten the screws on on s.o.') in (43),

which shares the noun *duimschroeven* and the verb *aandraaien* with the idiom *de duimschroeven aandraaien*.

(43) De gemeenten piekeren er niet over om ABN Amro de duimschroeven aan te draaien.
the local-authorities worry there not about to ABN Amro the thumb-screws on to tighten
'The local authorities don't even think about to tighten the thumb-screws on on ABN Amro.'

- it is an example of the expression studied in its idiomatic interpretation, i.e. an idiom variant, e.g. (44).

(44) Door dat niet te doen, dreigt de kerk de boot te missen.
by that not to do threatens the church the boat to miss
id. 'By not doing that, the church threatens to miss the opportunity.'

In general the corpus sentences include sufficient clues to be able to distinguish between an idiom's literal interpretation and its idiomatic interpretation. However, in some cases more context was needed to determine the intended interpretation, and therefore for some examples the wider context in the source has been examined.

Corpus examples other than the ones identified as idiom variants, in total 296 sentences,[2] have been excluded from further analyses. The idiom variants have been carefully examined to detect the following types of variation: determiner variation; for the noun, both number alternation and whether the noun is positive or diminutive; adnominal modification, including premodification by an adjective, and postmodification by a prepositional phrase, a *van*-phrase or relative clause; passivization; topicalization; and pronominal reference. This means that, except for passivization and external modification, only variation that requires a constituent which has meaning has been examined systematically. A limited number of other types of variation have not been

[2]Among which eight paper headlines or book titles, 147 literal examples and 141 examples of another idiom.

examined systematically, since it is assumed that these types can always be applied irrespective of idioms. These types include *verb first*, *verb second* and *verb raising* which occur independently of whether individual constituents have meaning (cf. e.g. Odijk (1993); Schenk (1994)).

data interpretation The next step in the procedure is the interpretation of the data. First, quantitive overviews are generated for each idiom showing the number of corpus examples, if any, for each variation type examined. These frequencies form the basis for the evaluation of the Idiom Variation Potential Hypothesis, which is repeated for convenience in (45).

(45) **Idiom Variation Potential Hypothesis**

a. If an idiom part has an idiomatic referent,
then its variation potential is in principle unrestricted.

b. If an idiom part has no idiomatic referent,
then variation that requires an idiomatic referent is blocked.

To test the hypothesis, it must first be determined whether the idiom parts have an idiomatic referent. One way is to find a plausible idiomatic meaning for the idiom and to show that the idiom parts refer to parts of this meaning. Although corpus examples can be used to establish an idiom's meaning, this meaning cannot be automatically extracted from corpora. To determine the plausibility of idiom parts having idiomatic referents, I select one consequence of idiom parts having idiomatic referents as a heuristic for the idiom parts actually having idiomatic referents. Although basically all the variation types examined require a constituent that has meaning, except for passivization and external modification, one type of variation that occurs relatively frequently in the corpus data is the demonstrative determiner specifying the idiom noun.

In this study, I take the presence of the demonstrative determiner as a heuristic for the noun having an idiomatic referent. This assumption leads to two sets of data, viz. (1) idioms the corpus data of which include one or more examples with a demonstrative determiner, and (2) idioms with no examples with a demonstrative determiner among the corpus data. It is expected that idiomatic referents can be found for the

individual parts of the idioms of the former set of data, and that this is not necessarily possible for the idioms in the latter set of data; 'not necessarily' because of course the non-occurrence of examples with a demonstrative determiner does not imply that the idiom parts cannot have idiomatic referents. Moreover, it is expected that the former set contains (more than one) corpus examples for each type of variation, while the latter set of data contains no variants at all, except for variants showing passivization and external modification.

7.2 INTERPRETATION OF THE DATA

This section tests the Idiom Variation Potential Hypothesis for 25 OBJ1-V idioms containing a definite article departing from two sets of data: (1) idioms with one or more examples with a demonstrative determiner, and (2) idioms with no examples with a demonstrative determiner. Before I present a quantitative overview of the results and interpret the data in the light of the Idiom Variation Potential Hypothesis (in Section 7.2.2), I first illustrate the evaluation of the hypothesis with two examples from each idiom set, viz. *de boot missen*, which shows much variation, and *de geest geven* ('to die'/'to break down'), which contains no variants among its corpus data. More data is examined in Section 7.2.2.

7.2.1 TWO EXAMPLES: *de boot missen* AND *de geest geven*

In this section I will illustrate the hypothesis with two idioms, viz. *de boot missen* and *de geest geven*.

de boot missen

Six examples with a demonstrative determiner specifying the noun in the idiom *de boot missen* have been identified. An example is given in (46).[3]

[3]For expository reasons, a gloss, a literal translation (if possible) and an idiomatic translation (if present) will solely be given for the first example presented of each idiom discussed in this chapter. In the subsequent examples only the idiomatic interpretation is given, unless a gloss and/or a literal translation is necessary for the discussion.

(46) Zo kon je op internet elektronische kranten maken
Hence could you on internet electronic newspapers make
en de hoofdredacteuren wilden die boot niet missen.
and the editors wanted that boat not miss
id. 'Hence, it was possible to create electronic newspapers on the internet and the editors did not want to miss that opportunity.'

This means that it is expected that the idiom parts have an idiomatic referent, and that the idiom's variation potential is in principle unrestricted. Based on the corpus data, I postulate that the idiomatic referent of *boot* is MOGELIJKHEID ('opportunity') and the idiomatic referent of *missen* is MISLOPEN ('fail to catch'), which is equal to its literal meaning. The link between the idiom's literal domain and idiomatic domain is represented in the concept mapping in Figure 7.1.[4]

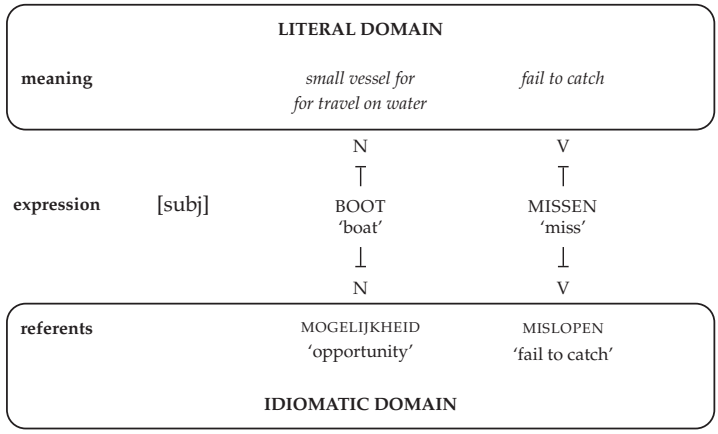


Figure 7.1 *Concept mapping of the idiom* ***de boot missen***

As expected, the data show all kinds of variation occurring in one or more corpus examples. In Table 7.1 the frequencies for this idiom are itemized by determiner type.

[4]The idiomatic referents are represented as concepts in Dutch, while the literal meaning of the components are taken from the English WordNet (http://wordnet.princeton.edu/).

determiner	**premodifier**	BOOT	**postmodifier**	MISSEN
de 533	adj 5	topicalization 1 pro_ref 1	PP 19 *van*-phrase 11 relative clause 1	passive 5
demonstrative 6		topicalization 1		
een 1				
quantifier 2		pl 1 pro_ref 1		

Table 7.1 *The variation frequencies for* ***de boot missen*** *by determiner type*

In (47)-(49), examples are given of number and determiner variation, topicalization, and pronominal reference (pro_ref) respectively. On the idiomatic level, the variation can be applied directly to the idiomatic referent of *boot*, see the translation in the examples given.

(47) Erg jammer, want nu missen we allerlei boten.
id. 'Too bad, because now we miss various opportunities.'

(48) Die boot willen de andere educatieve uitgeverijen niet missen.
id. 'That opportunity, the other educational publishers do not want to miss.'

(49) Ze hebben de boot gemist of zullen weldra die gaan missen, (omdat ze, in tegenstelling tot vrouwen om hen heen, te bescheiden zijn en te lang thee blijven zetten voor hun moeder.)
id. 'They have missed the opportunity or will miss it in a little while, (because, contrary to women around them, they are too modest and continue to make tea for their mother for too long.)'

Since the noun has an idiomatic referent, it is expected to be subject to internal modification. An example of premodification is given in (50), for which an analysis of the adjective modifying internally the idiomatic referent of the noun is appropriate: Financial opportunities are opportunities that for instance may lead to large sums of money.

(50) De docenten en onderzoekers zijn bang de financiële boot te zullen missen.
id. 'The teachers and researchers are afraid that they will miss the opportunity that may lead to large sums of money.'

Furthermore, the idiom *de boot missen* occurs with several types of postmodification, see for example (51) - (54).

(51) Tot nog toe heeft het midden- en kleinbedrijf de boot van het Internet gemist.
id. 'So far, the small business has missed the opportunity of the internet.'

(52) Niemand wil immers de boot van een potentiële bestseller missen.
id. 'After all, nobody wants to miss the opportunity of a potential best seller.'

(53) De partij illustreert hoe Nederland de boot heeft gemist die door China wordt bestuurd.
lit. 'The match illustrates how The Netherlands have missed the boat that is navigated by China.'

(54) Alle banken en beleggers wilden er bij zijn: bang om de boot naar financieel wonderland te missen.
lit. 'All the banks and investors wanted to be there: afraid to miss the boat to financial wonderland.'

In the examples (51) and (52), the noun *boot* is postmodified by a *van*-phrase. In both examples the *van*-phrase is semantically incompatible with the literal sense of *boot*, i.e. if analysed in isolation the NP does not make sense. However, the *van*-phrases are semantically compatible with the idiomatic referent of the noun, i.e. they are perfectly interpretable as modifiers of MOGELIJKHEID.

In example (53) the noun is postmodified by a relative clause that is not only compatible with the literal meaning of the noun, but can also be interpreted metaphorically in a way that is compatible with the idiomatic referent of *boot*; literally, a boat is a type of vehicle that can be navigated, and the person that navigates the boot indicates the direction, while metaphorically 'navigate' can be interprete as in 'lead', which is compatible with MOGELIJKHEID. To be able to interprete (53) in its idiomatic meaning, it is necessary to know the context of the sentence. The sentence is uttered by a badminton coach and *de partij* refers to a badminton match between a Dutch and a Chinese player. In this light, (53) can be interpreted as 'The Netherlands have missed the opportunity that is led by China'.

In example (54) the noun is postmodified by a prepositional phrase headed by *naar* ('toward'). This type of postmodification is typical for a whole range of nouns that literally denote an instrument that can be used for transportation. The preposition *naar* is literally used to indicate a direction in geographical terms, but can be interpreted in the same sense in an abstract domain. Example (54) can be interpreted as 'to miss the opportunity to financial wonderland', where *financial wonderland* can be interpreted in various ways, e.g. as financial advantages in general, but also as a country that is financially attractive.

To conclude, the corpus data of *de boot missen* show much variation and moreover contains at least one example of each type of variation examined. These results support the hypothesis that if idiomatic referents can be identified for the idiom parts, then there is potentially unlimited variation.

de geest geven

A total number of 217 examples have been found in the corpus for the idiom *de geest geven*. An example is given in (55).

(55) Als haar breiwerk eindelijk af is, heeft de oude Jakov
when her knitting finally done is has the old Jakov
voor wie de trui was bedoeld, net de geest gegeven.
for whom the jumper was meant just the soul give
id. 'When her knitting is finally finished, the old Jakov, for whom the jumper was meant, has just died.'

None of the corpus examples show variation that requires an idiomatic referent for the idiom parts. A plausible meaning for *de geest geven* is 'to die' or in other examples, in particular concerning machinery, 'to break down'. Since these meanings do not contain components to which the idiom noun *geest* can refer, it is plausible to assume that the idiom parts do not have idiomatic referents. The concept mapping for this idiom is shown in Figure 7.2.

Since the idiom does not show any variation that requires an idiomatic referent, it can be concluded that the corpus data of this idiom support the Idiom Variation Potential Hypothesis.

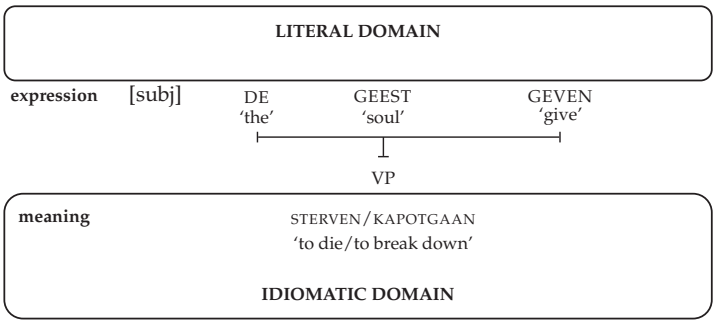


Figure 7.2 *Concept mapping of the idiom* ***de geest geven***

7.2.2 MORE DATA

The next step in the data analysis is to see whether the assumption underlying the Idiom Variation Potential Hypothesis is supported by other idioms in the same way as it is supported by the data of *de boot missen* and *de geest geven*. Before discussing the individual idioms, I first present present a quantitative overview of the overall results.

Figure 7.3 shows the number of corpus examples found for each type of variation per idiom.[5] The table is divided in two parts: The top part lists the idioms for which the number of examples with a demonstrative determiner is not zero and the lower part shows the results of the idioms with zero examples of a demonstrative determiner among the corpus data. The *Count* rows gives the total number of non-empty

[5]The abbreviations in the figure stand for: *TOT* = total number of corpus examples; *de/het* = total number of examples with a definite article; *demon* = total number of examples with a demonstrative determiner; *other* = total number of examples with a determiner other than a definite article or demonstrative determiner; *sg* = total number of examples in which the idiom noun is singular; *pl* = total number of examples in which the idiom noun is plural; *dim* = total number of examples in which the idiom noun is in the diminutive form; *topic* = total number of examples in which the idiom NP is topicalized; *pro_ref* = total number of examples in which the idiom NP is the antecedent of a pronoun; *pre* = total number of examples containing premodification of the idiom noun; *post-PP* = total number of examples containing postmodification of the idiom noun by a prepositional phrase; *van* = total number of examples containing postmodification of the idiom noun by a *van*-phrase; *relat* = total number of examples containing postmodification of the idiom noun by a relative clause; *pass* = total number of examples showing passivization.

Idioms with one or more examples with a demonstrative determiner

		determiner			inflection			syntactic		modification			
Idiom	**TOT**	**de/het**	**demon**	**other**	**sg**	**pl**	**dim**	**topic**	**pro_ref**	**pre**	**post-PP**	**van**	**relat**	**pass**
de kar trekken	493	464	23	6	493	-	-	-	1	14	3	21	-	6
de dans ontspringen	735	723	11	1	735	-	-	-	-	14	1	10	-	-
de boot afhouden	443	434	9	-	443	-	-	-	-	1	-	1	-	-
de handschoen opnemen	152	144	8	-	152	-	-	2	-	3	-	3	4	1
de ban breken	163	154	7	2	163	-	-	-	-	4	-	8	-	11
de boot missen	542	533	6	3	541	1	-	2	2	5	19	11	1	5
de bal terugkaatsen	138	133	5	-	138	-	-	2	-	-	-	1	-	4
de trom roeren	113	102	3	8	113	-	-	-	-	5	-	9	-	3
het boetekleed aantrekken	385	382	1	2	385	-	-	-	-	2	-	1	-	-
het roer omgooien	537	536	1	-	537	-	-	2	-	4	-	1	-	28
Count	**10**	**10**	**10**	**6**	**1**	**0**		**4**	**2**	**9**	**3**	**10**	**2**	**7**

Idioms with no examples with a demonstrative determiner

		determiner			inflection			syntactic		modification				
Idiom	**TOT**	**de/het**	**demon**	**other**	**sg**	**pl**	**dim**	**topic**	**pro_ref**	**pre**	**post-PP**	**van**	**relat**	**pass**
de aftocht blazen	184	179	-	5	184	-	-	-	-	6	-	-	-	3
de bakens verzetten	280	250	-	30	1	279	-	-	-	5	-	3	-	37
de benen nemen	661	660	-	1	-	661	-	-	-	-	-	-	-	-
de boventoon voeren	1414	1413	-	1	1414	-	-	1	-	13	-	-	-	1
de duimschroeven aandraaien	71	71	-	-	-	71	-	-	-	3	-	2	-	12
de geest geven	217	217	-	-	217	-	-	-	-	-	-	-	-	-
de kroon spannen	711	711	-	-	711	-	-	-	-	2	-	2	-	-
de lakens uitdelen	424	420	-	4	-	424	-	-	-	2	-	-	-	6
de mouwen opstropen	276	217	-	59	-	276	-	-	-	-	-	-	-	20
de noodklok luiden	645	638	-	7	642	3	-	-	-	3	-	-	-	22
de plank misslaan	185	184	-	1	184	1	-	-	1	1	-	1	-	8
de stormbal hijsen	56	56	-	-	56	-	-	-	-	-	-	-	-	7
het ijs breken	155	155	-	-	155	-	-	-	-	-	-	-	-	8
het onderspit delven	868	867	-	1	868	-	-	-	-	2	-	-	-	4
het spits afbijten	498	498	-	-	498	-	-	-	-	-	-	-	-	14
Count	**15**	**15**	**0**	**9**	**3**	**0**		**1**	**1**	**9**	**0**	**4**	**0**	**12**

Figure 7.3 *Variation frequencies per idiom based on corpus data.*

cells for each variation type.[6]

Various conclusions can be drawn from the overall results. First of all, as expected, the non-occurence of variation is dominant in the lower part of the chart. In other words, idioms without examples with a demonstrative determiner exhibit less variation than idioms with one or more examples with a demonstrative determiner. For both data sets we can calculate the percentage of variation by adding the number of non-empty cells and divide this total by the total number of cells. The idioms with demonstrative show 54% of variation and the data set with idioms without demonstrative show 26% of variation (a proportion of approximately 2:1). If we leave out passivization and premodification (see below), then the data set with idioms with demonstrative show 47.5% of variation against 15% of variation for the data set with idioms without demonstrative determiner (a proportion of approximately 3:1).

Second, it is remarkable that the variant in which the definite article specifies the idiom noun is by far the most frequent use in the corpus. Except for the idiom *de mouwen opstropen* (lit. 'to roll up the sleeves', id. prepare to work hard') of which the definite article occurs in 80% of the examples, the use of the definite article in other idioms lies between 88% and 100%. The function of determiners in idioms and especially the frequent use of the definite article is discussed in detail in Fellbaum (1993).

What can furthermore be observed from the table in Figure 7.3 is that whether or not an idiom can occur in the passive does not seem to relate to whether the idiom parts have idiomatic referents, i.e. factors independent of the status of the idiom noun play a role in the passivization of idioms (cf. inter alia Van der Linden (1993); Schenk (1994)). This point will be elaborated in Section 7.3.

Another type of variation that is dominant in both parts of the table is premodification. As discussed in the previous chapter, idiom nouns do not need to have idiomatic referents to be suitable for modification. However, there is a distinction between internal and external modification. The former type can only occur with idiom nouns that have an idiomatic referent. This means that idioms of which the idiom noun does not have an idiomatic referent can only be modified externally, i.e

[6]Except for the totals of the number of the noun, for which I only count the total number of idioms of which both the number of singular examples and the number of plural examples are not zero.

the modifier does not semantically modify the idiom noun, but must be interpreted as modifying the idiom as a whole. The internal versus external modification dichotomy can only be tested by examining the individual examples containing modification. In the discussion of the individual idioms, special attention will be paid to the modification of idiom nouns without idiomatic referents.

In the remainder of this section, the results will be interpreted in the light of the Idiom Variation Potential Hypothesis. I will start with discussing the set of idioms with one or more demonstrative determiners. It is expected that these idioms have idiomatic referents for the individual parts, hence plausible referents will be sought using the corpus examples.

Next the set of idioms with no demonstrative determiner among the corpus data will be discussed. As mentioned, it is not necessarily the case that the idioms in this set do not have idiomatic referents. Also for these idioms I will postulate on the basis of the corpus data whether the individual parts have idiomatic referents. If this is the case it will be examined if an example can be provided in which the noun can be specified by a demonstrative determiner without losing its idiomatic meaning. Moreover, as a consequence of the hypothesis it is expected that if the idiom has idiomatic referents, other types of variation are present among the corpus data. On the other hand, if no plausible idiomatic referents can be found for the idiom parts, no examples of variation are expected to be part of the corpus data. Each occurrence of idiom variation where non-occurrence is expected will be discussed. Furthermore, examples with premodification will be examined to determine whether the adjective semantically modifies the idiom noun or the idiom as a whole.

As noted before, the overall frequencies are very low. Especially, topicalization, pronominal reference and postmodification by a relative clause occur very infrequently in the data, with a maximum of four occurrences per idiom. The non-occurence of a variation type where occurrence is expected is not necessarily counter-evidence. Constructing plausible examples can help to decide whether some type of variation is indeed impossible and provide a more complete overview of potential idiom variation. Although the focus is on corpus data, examples will be constructed (1) for variants that are expected, but not observed

in the corpus data, and (2) for some variants that are not expected and that also have not been observed. To avoid theoretical bias in the acceptability judgements, a total of 30 examples (merely constructed examples) have been presented to a panel of eight linguists, all native speakers of Dutch and three of whom have no knowledge of the theory tested. Constructing plausible examples is not easy, especially since contextual factors play an important role and may influence the acceptability judgements, in particular when judging borderline examples. The constructed examples are modeled on corpus examples, but also the corpus data extracted often lack sufficient context.

It is not my intention to discuss each idiom as elaborately as has been done with *de boot missen*. The primary goal is to determine whether plausible idiomatic referents can be found for the idiom parts and accordingly to test the Idiom Variation Potential Hypothesis with the corpus data to begin with and if necessary using constructed examples. Notable corpus examples and the subset of constructed examples that have been judged by the panel will be discussed in the interpretation of the data. For the sake of completeness, Appendix B lists the examples for each type of variation that have been found in the corpus, but that are not presented in the main text. Moreover the appendix contains the constructed examples that have not been judged by the panel. To visualize the outcome of the Idiom Variation Potential Hypothesis in a insightful way, the link between the idiom components and the idiom's literal domain and idiomatic domain will be illustrated in a concept mapping.

de kar trekken

A total of 23 examples with a demonstrative determiner are among the corpus data, see (56) for an example. The example in (56) has been judged by the panel: six subjects judged it as well-formed in its idiomatic reading, while one subject judged it as being ill-formed and one subject finds it a borderline example.

(56) Zoals de zaken nu liggen, blijf ik bij het team en zal ik
as the things now are stay I with the team and shall I
met hem die kar trekken.
with him that cart pull

lit. 'As things are now, I stay with the team and I will pull that cart with him.'
id. 'As things are now, I stay with the team and I will lead that project with him.'

Based on the corpus examples I assume that de idiom parts of *de kar trekken* have idiomatic referents. The idiom's literal meaning and the idiomatic referents are represented in the concept mapping in Figure 7.4.

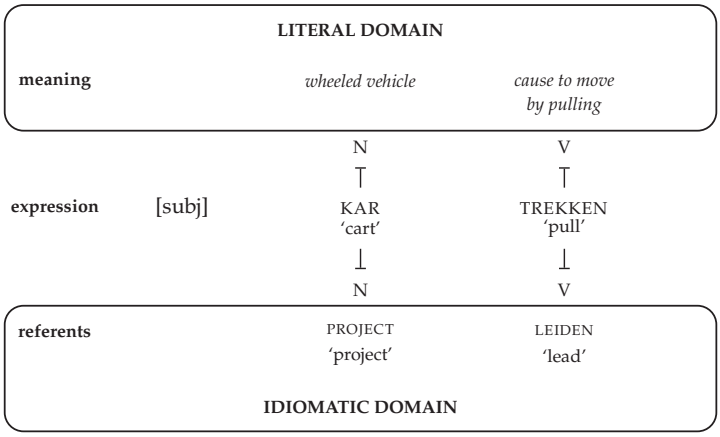


Figure 7.4 *Concept mapping of the idiom* ***de kar trekken***

The corpus data are compatible with the Idiom Variation Potential Hypothesis, see Appendix B.1 for some examples. However, not all types of variation that are expected have actually been found in the data and therefore some examples have been constructed to see whether they are plausible, see again Appendix B.1. According to my intuitions, all examples are well-formed and hence compatible with the hypothesis.

Notable are the corpus examples in which the idiom noun is premodified by an adjective that is compatible with both the literal meaning of *kar* and its idiomatic referent, see e.g. (57).

(57) Dan ben ik er toch wel trots op dat ik de zware kar zelf trek.
lit. 'Then I am proud that I pull the heavy cart myself.'
id. 'Then I am proud that I lead the hard project myself.'

One example of pronominal reference was found in the corpus, see (58). Based on the context, it can be concluded that this is an example of the idiom in its idiomatic interpretation. It is however an open question whether this variant is an example of systematic variation or whether it should be regarded as wordplay; if we assume that *kar* cannot be interpreted as PROJECT outside the idiom, i.e. without the verb *trekken*, then the pronoun *hem* which refers to *de kar* cannot be interpreted in the NP's idiomatic sense, since *hem* is not the direct object of *trekken* but of the verb *duwen*.

(58) (De voortrekkersrol die Nederland tien jaar geleden op de VN-top in Rio de Janeiro speelde behoort tot het verleden.)

"We trekken de kar niet meer, we helpen hem af en toe
we pull the cart no more we help it once and while
duwen", zegt professor Hans Opschoor.
push say professor Hans Opschoor

'(The role of pioneer that The Netherlands played ten years ago at the VN-top in Rio de Janeiro is past history.)
"We do not lead the project anymore, we help to push it further once in a while", says Professor Hans Opschoor.'

de dans ontspringen

A total of 11 examples with a demonstrative determiner have been found in the corpus. An example is given in (59).

(59) Jerome Courtailler ontspringt deze dans omdat hij nog
Jerome Courtailler originates-from this dance because he still
onder het huidige strafrecht zal worden berecht.
by the current law shall be tried
id. 'Jerome Courtailler escapes this punishment, because he will still be tried by the current criminal law.'

Although the data do not show many types of variation, I conclude that the idiom parts have idiomatic referents, based on the corpus examples, among which the examples in Appendix B.2. However, it is not easy to define the concept to which the idiom noun refers. This is a typical example of a noun that denotes a rather unspecified concept. One

property of the idiomatic referent is that it has a negative connotation, but furthermore it receives its specific interpretation by the context in which it it used. This means that the idiomatic referent of *dans* in this idiom can best be described as IETS ONAANGENAAMS ('something unpleasant'). The literal meaning of the idiom and the idiomatic referents of its parts are shown in Figure 7.5.

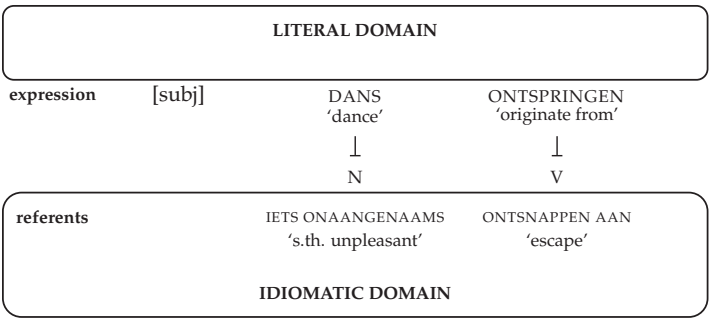


Figure 7.5 *Concept mapping of the idiom* ***de dans ontspringen***

As a consequence of the idiom parts having idiomatic referents, internal modification is predicted to be possible. An example with modification is given in (60), for which an internal reading is plausible, see the translation.

(60) Nixon had zich schuldig gemaakt aan een strafbaar feit,
Nixon had himself guilty made to a punishable fact
en hij kon alleen maar de strafrechtelijke dans
and he could only just the legal dance
ontspringen doordat zijn opvolger hem pardonneerde.
originate-from because his successor him excused
id. 'Nixon had committed a legal offence, and he could only escape the criminal prosecution, because his successor excused him.'

Some examples have been constructed for types of variation that have not been found in the corpus, see the Appendix B.2. Since all examples are well-formed in their idiomatic meaning, I conclude that the data of this idiom are in agreement with the Idiom Variation Potential Hypothesis.

de boot afhouden

In total, nine examples with a demonstrative determiner are among the corpus data, inter alia example (61).

(61) Aanvankelijk hield Vasalis deze boot af, maar in latere jaren
initially kept Vasalis this boat off but in later years
heeft ze Reve kennelijk toch enkele malen verzen
has she Reve apparently yet a-few times poems
toegestuurd.
sent
lit. 'Initially, Vasalis kept off this boat, but apparently she has sent poems to Reve a few times in later years.'
id. 'Initially, Vasalis warded off this undesirableness, but apparently she has sent poems to Reve a few times in later years.'

Based on the fact that multiple examples of a demonstrative determiner have been found in the corpus, it can be assumed that the idiom parts have an idiomatic referent. Again we are dealing with an idiom, the noun of which denotes a rather unspecified concept. One property of the idiomatic referent is that it has a negative connotation. The idiomatic referent of *boot* in this idiom can best be described as IETS ONGEWENSTS ('something undesirable'), as is shown in the concept mapping in Figure 7.6.

Besides the nine examples with a demonstrative determiner, the data show one example of premodification, see (62), and one example of postmodification by a *van*-phrase, see (63). The translations show that both examples can be interpreted with *boot* referring to 'something undesirable', and are therefore compatible with the hypothesis.

(62) Zij ziet drie redenen waarom met name de christen-democraten de extremistische boot moeten afhouden.
id. 'She sees three reasons why in particular the Christian Democrats must ward off the extremist undesirableness.'

(63) Uit vertrouwelijke gesprekken blijkt dat Bakker de boot van de liberalisering afhield.
id. 'From confidential conversations, it seems that Bakker warded off the undesirableness of liberalization.'

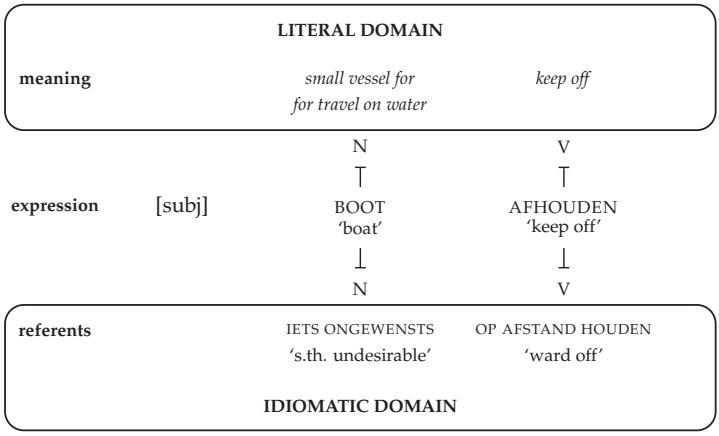


Figure 7.6 *Concept mapping of the idiom* ***de boot afhouden***

No other examples of variation have been found in the data. Some constructed examples are given in Appendix B.3. The constructed example in (64) was among the set of examples that was judged by the panel. First I would like to note that it is not without reason that just four examples of pronominal reference have been found among the total number of corpus data of all idioms examined; as discussed above, it is the question whether the pronoun that refers to the idiom NP can be the direct object of another verb than the idiom verb, but repetition of the idiom verb might sound odd. This is confirmed by the acceptability judgements of the panel: four subjects have judged example (64) as borderline example, against three who find the example well-formed and one who finds the example ill-formed.

(64) De Nederlandse overheid zou die boot moeten afhouden, maar zij ziet liever dat die vanuit Brussel wordt afgehouden.
id. 'The Dutch government should ward off that undesirableness, but they prefer that it will be ward off from Brussels.'

According to my intuitions, all constructed examples as given in Appendix B.3 are well-formed and in agreement with the Idiom Variation Potential Hypothesis.

de handschoen opnemen

A total of eight examples with a demonstrative determiner have been found among the corpus data, see (65) for an example. Based on, inter alia, the example in (65), I assume that the idiom parts of *de handschoen opnemen* have idiomatic referents.

(65) De overheid wordt in het manifest ruimschoots
the government gets in the manifesto amply
aangesproken en het kabinet wil die handschoen
addressed and the cabinet wants that glove
opnemen.
pick-up
id. 'The government is addressed amply in the manifesto and the cabinet wants to take up that challenge.'

The idiom's literal meaning and the idiomatic referents are represented in the concept mapping in Figure 7.7.

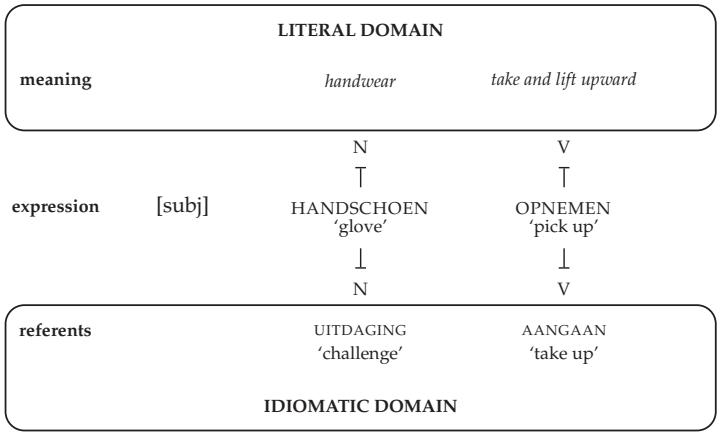


Figure 7.7 *Concept mapping of the idiom* ***de handschoen opnemen***

Interesting are the examples of premodification and postmodification, in (66) and (67) respectively, in which another idiom containing the noun *handschoen*, viz. *iemand de handschoen toewerpen* ('to throw down the gauntlet'), is part of the modifier.

(66) Het is onbegrijpelijk, dat CDA en VVD geweigerd
it is incomprehensible that CDA and VVD refused
hebben de hun toegeworpen handschoen op te nemen en
have the them thrown-down glove up to pick and
volstaan hebben met eenvoudig naar de traditie te
confine have with simply to the tradition to
verwijzen.
point
id. 'It is incomprehensible that CDA and VVD have refused to take up the challenge issued and that they have confined themselves to simply pointing to the tradition.'

(67) Ina Brouwer nam de handschoen op die Bolkestein haar had
Ina Brouwer took the challenge up that Bolkestein her had
toegeworpen, maar stelde een voorwaarde die de uitdager
thrown-down but stated a condition that the challenger
moeilijk kan weigeren.
hardly can refuse.'
id. 'Ina Brouwer took up the challenge that Bolkestein had issued, but stated a condition that can hardly be refused by the challenger.'

Both examples are interpretable with the modifier modifying the idiom noun as denoting UITDAGING, see the translations. Corpus examples of topicalization and postmodification by a *van*-phrase and some plausible constructed examples of other types of variation are given in Appendix B.4. All examples are compatible with the hypothesis.

de ban breken

A total of seven examples with a demonstrative determiner have been found in the corpus data. An example is given in (68).

(68) (De gevestigde partijen durfden niet vrijuit over de problemen die de immigratie meebracht te praten, uit vrees met typen als Glimmerveen en Janmaat te worden geassocieerd.)

Pas begin jaren negentig heeft VVD-leider Bolkestein deze
just begin years nineties has VVD-leader Bolkestein this
ban gebroken, ...
spell broken

id. '(The established parties were afraid to speak freely about the problems that the immigration involves, for fear of being associated with people like Glimmerveen and Janmaat.)
Only at the beginning of the nineties VVD-leader Bolkestein has ended this fear, ...'

Although at first sight and based on another idiom containing the noun *ban*, viz. *in de ban zijn/raken van iets* ('be/fall under the spell of s.th.'), it reasonable that the idiomatic referent of *ban* is BETOVERING ('spell'). However, this concept does not seem to fully the cover the meaning. A better description for *ban* in this idiom is something more vague such as 'a (slightly negative) state that exists for a long time and preferably should be ended'. The precise interpretation can most often be determined from the context in which the idiom is used. The concept mapping of *de ban breken* is given in Figure 7.8.

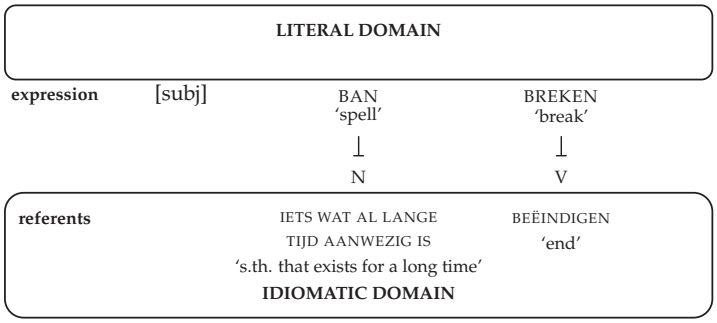


Figure 7.8 *Concept mapping of the idiom* ***de ban breken***

Examples of premodification are merely examples of modifiers of time, which can be interpreted as semantically modifying the idiom noun, see e.g. (69).

(69) Voor Den Haag was dat het sein om het tij te keren en de 12 jaar durende ban te breken.
id. 'For The Hague it was the sign to turn the tide and to end the 12-year-old spell.'

More examples, both corpus examples and constructed examples, are given in Appendix B.5. All examples are compatible with the hy-

pothesis that if the idiom parts have idiomatic referents that then the idiom's variation potential is in principle unrestricted.

de bal terugkaatsen

Although seven examples with a demonstrative determiner have been found in the corpus, not many other types of variation are present. Example (70) is one of the corpus examples with a demonstrative determiner and has also been judged by the panel. Although it can be argued that the example needs more context to be better interpretable, six subjects find this example well-formed, one finds it ill-formed and another one judged it as borderline example.

(70) Sinds Eureko's lobby in Brussel kaatsen de Polen die bal
since Eureko's lobby in Brussels hit the Poles that ball
net zo hard terug.
just as hard back
lit. 'Since Eureko's lobby in Brussels, the Poles hit back that ball just as hard.'
id. 'Since Eureko's lobby in Brussels, the Poles rebound that matter just as hard.'

Based on the panel's and my own intuitions, I conclude that this example is well-formed in its idiomatic reading and the idiom parts have idiomatic referents, which are represented in the concept mapping in Figure 7.9.

The corpus data for this idiom include one example of modification, viz. postmodificaton by a *van*-phrase, see (71).

(71) In het debat na afloop van de hoorzitting kaatste Frans Timmer van Unilever de bal van de transparantie terug naar de actiegroepen.
id. 'During the debate after the hearing, Frans Timmer of Unilever rebounded the issue of transparency to the action group.'

This example is perfectly interpretable with the idiom noun denoting KWESTIE. To test whether other types of variation are also possible for this idiom, I have constructed some examples, which are given in Appendix B.6. All examples are well-formed and compatible with the hypothesis.

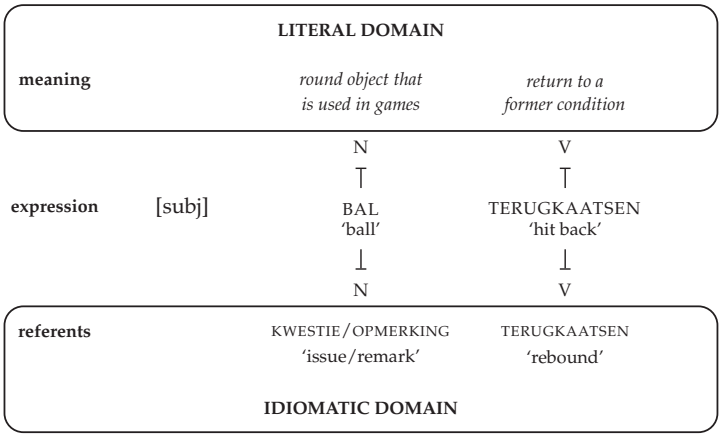


Figure 7.9 *Concept mapping of the idiom* ***de bal terugkaatsen***

de trom roeren

Three examples containing a demonstrative determiner are among the corpus data. One example is given in (72).

(72) De generale staf in Moskou roert die trom oorverdovend
the general staff in Moskou moves that drum deafening
hard.
loud
id. 'The general staff at Moskou spreads that message very loud.'

The idiom *de trom roeren* is probably the least well-known idiom studied; five of the panel members indicated not to know the meaning of this idiom. With no knowledge of an idiom's meaning, it requires a set of examples preferably in broad context to get a feeling of how the idiom must be interpreted. Not knowing the idiom's meaning can be a possible explanation of why sentence (73) was judged as well-formed only once, while four subjects find the example ill-formed and three subjects have judged it as borderline example.

(73) Zij heeft jarenlang die trom geroerd en zal hem nog langer blijven roeren zolang het kabinet niet overstag gaat.
id. 'She has spread that message for years, and she will continue spreading it as long as the cabinet does not come round.'

I tentatively assume that the idiom parts have idiomatic referents, see the concept mapping in Figure 7.10. However, more evidence from speakers who know the idiom is required to draw firm conclusions.

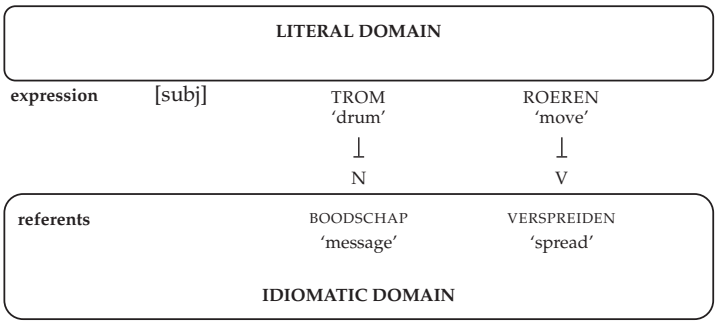


Figure 7.10 *Concept mapping of the idiom* ***de trom roeren***

Two example of modification are given in (74) and (75) and are interpretable with the parts having idiomatic referents.

(74) In beide landen wordt nu driftig de nationalistische trom geroerd.
id. 'In both countries, the nationalistic message is being spread vehemently.'

(75) En conflict is wat de nationalisten nodig hebben, om de trom van de zelfstandigheid krachtiger te kunnen roeren.
id. 'And conflict is what the nationalists need to be able to spread the message of independence more powerful.'

The corpus data do not show much variation, and therefore some examples have been constructed. According to my intuitions, these examples are well-formed in their idiomatic meaning and in agreement with the hypothesis.

het boetekleed aantrekken

Only one example with a demonstrative determiner has been found in the corpus for the idiom *het boetekleed aantrekken*. The example, given in (76), is part of a poem and may not be reliable.

(76) Je zegt: ik trek dat boetekleed wel aan.
you say I put that hairshirt just on
lit. 'You say: I will put on that hairshirt'

I am not sure whether the idiom parts have idiomatic referents. The difficult part in determining this is that the literal meaning of the noun *boetekleed* contributes to the idiom's idiomatic meaning, i.e. literally the idiom denotes the event of putting on a hairshirt as a sign of repentance and atonement (which is common is some religious circles). It is however the question whether *boetekleed* actually refers to 'sign of repentance' in its idiomatic interpretation, and hence can be said to have an idiomatic referent, as depicted in Figure 7.11, or that the individual parts do not have idiomatic referents, as represented in Figure 7.12.

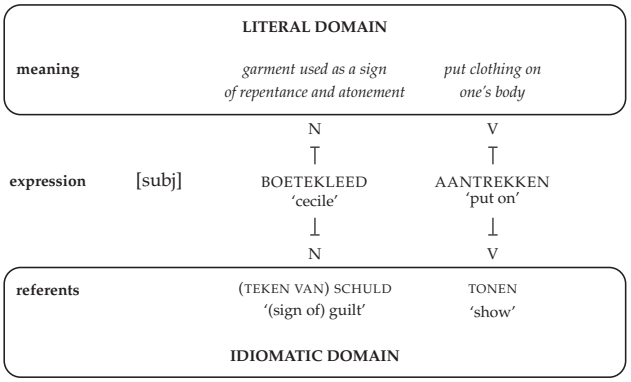


Figure 7.11 *Concept mapping of the idiom* ***het boetekleed aantrekken*** *with idiomatic referents for the individual parts.*

The corpus data, which do not show much variation, do not give a decisive answer, neither do the panel judgements of the examples given in (77) and (78). Example (77) is ill-formed for six subjects, and just two subjects find it well-formed in its idiomatic meaning. On the other hand, example (78) was judged as well-formed by seven subjects and as borderline example by one subject.

(77) Heel wat renners die de gezondheidscontrole niet gepasseerd waren (en zij allen ontkenden uiteraard in hoge mate, geen van

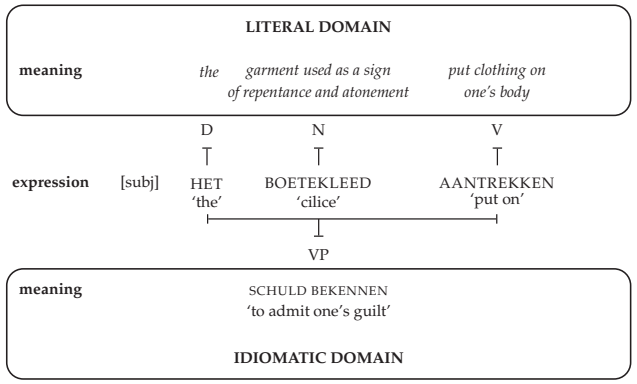


Figure 7.12 *Concept mapping of the idiom* ***het boetekleed aantrekken*** *with no idiomatic referents for the idiom parts.*

hen trok een boetekleed aan).
lit. 'Many riders who had not passed the health check (and of course they all denied to a great extent, none of them put on a hairshirt.'

(78) Uiteindelijk trok hij het boetekleed aan en hij zal het vast nog vaker aan moeten trekken.
lit. 'In the end, he put on the hairshirt and he certainly shall have to put it on more often.'

The judgements given by the panel about these examples form a problem for my theory. On the one hand, the judgements about example (77), which shows determiner alternation, support the view that the idiom parts do not have an idiomatic referent. On the other hand, the judgements about example (78), which is an example of pronominal reference, support the view of the idiom parts having idiomatic referents. At this point, I can only speculate about the difference in judgements; it is possible that (77) is judged as ill-formed for other reasons than the change of the determiner within the idiom NP, whereas a possible argument for the acceptability of example (78) might be that this type of variation is plausible with idioms such as *het boetekleed aantrekken* that literally denote an event that contributes to the idiom's idiomatic meaning. Further research will be needed to show whether this idiom is in fact a counterexample of the hypothesis, or that it is an example of

another type of idiom. I will return to this point when discussing the overall results in Section 7.4.

As can be seen in the table in Figure 7.3, the corpus examples do not show much variation. Two examples of premodification are among the corpus data, which are given (79) and (80). Example (79) is, according to my intuitions, an example close to wordplay involving the literal meaning of *boetekleed*. In (80) the adjective *politieke* ('political') can only be interpreted as modifying the idiom as a whole, but not in the sense of 'in the political domain', but in the sense of 'with respect to the policy'. The one example with a *van*-phrase is shown in (81). I am not sure how to interpret this sentence, and stipulate that it is not a correct use of the idiom, not necessarily because of the *van*-phrase, but based on the context, which does not seem compatible with the meaning of the idiom.

(79) Voor de aandeelhoudersvergadering had bestuursvoorzitter Peter Bakker van TPG gisteren wel zijn meest deemoedige boetekleed aangetrokken.
lit. 'For the shareholders meeting, the chair of the board of TPG, Peter Bakker, had put on his most humble hairshirt yesterday.'
id. 'For the shareholders meeting, the chair of the board of TPG, Peter Bakker, admitted guilt in a most humble way yesterday.'

(80) We vroegen hem het politieke boetekleed aan te trekken, maar dat gebeurt tegenwoordig kennelijk nooit meer.
lit. 'Regarding the policy, we asked him to admit guilt, but apparently that does not happen anymore nowadays.'

(81) Chipmachinefabrikant ASM Lithography en automatiseerder Getronics voerden vanochtend afwisselend de lijst van verliezers aan. Het koersverlies van ASM Lithography, die de laatste weken vaker het boetekleed van de getergde tech-sector mocht aantrekken, schommelde rond 4,5 procent op 24,83 euro.
lit. 'Chip manufacturer ASM Lithography and computer firm Getronics alternately led the list of losers this morning. The fall in prices of ASM Lithography, who ought to put on the hairshirt of the plagued technical sector, fluctuated around 4.5 percent at 24.83 euro.'

het roer omgooien

One example of a demonstrative determiner has been found in the corpus, see the example in (82).

(82) (Tot die tijd deden we ons best om een goede schoen te maken. Maar de echte dynamiek ontbrak.)

Dat roer hebben we omgegooid.
that helm have we shifted.

lit. '(Until then, we did our best to make a good shoe. But the real dynamics were lacking.)
That helm, we have shifted.'

According to my intuitions, the sentence in (82) is not a correct use of the idiom. Not the example in (82) but another example with a demonstrative determiner, viz. the sentence in (84) which has been found on the internet, has been presented to the panel, just as the examples in (83) and (85). Example (83) does not show any variation and is well-formed for all the subjects. The sentence in (84) was judged by three subjects as well-formed and by three other subjects as ill-formed in its idiomatic reading, while two subjects have judged it as borderline example. Three subjects find example (85) well-formed, while two subjects find it ill-formed, and three subjects have judged it as borderline example.

(83) De Bijenkorf had enkele jaren daarvoor al succesvol
the Bijenkorf had some years before already successfully
het roer omgegooid.
the helm shifted

(84) Ik benijd en heb groot respect voor die mensen die dat
I envy and have great respect for those people who that
roer helemaal kunnen omgooien en hun eigen weg inslaan.
helm totally can shift and their own way take

(85) De Hadeln wil het roer van Berlijn omgooien.
De Hadeln wants the helm of Berlin shift

I tentatively assume that both (84) and (85) are ill-formed in their idiomatic reading and that the individual parts of the idiom *het roer*

omgooien have no idiomatic referents, but that a meaning is assigned to the idiom as a whole. This view is represented in the concept mapping in Figure 7.13.

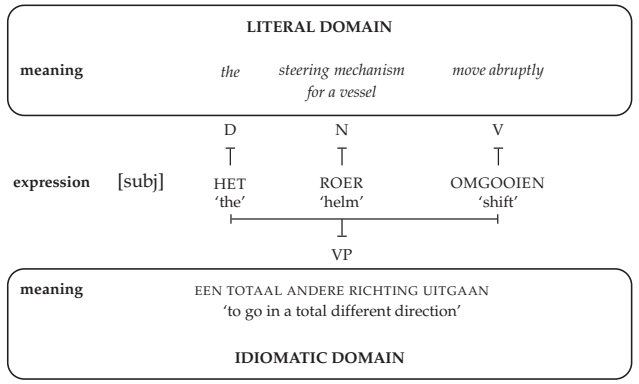


Figure 7.13 *Concept mapping of the idiom* ***het roer omgooien***

However, as the judgements of the examples reveal, for some speakers variation is acceptable. Just as the idiom *het boetekleed aantrekken* this idiom is also a problem for my theory, but on the other hand it may also be evidence for the existence of another type of idiom; again we are dealing with an idiom that literally denotes an event that contributes to idiom's idiomatic meaning. The expression *het roer omgooien* literally denotes the action of shifting the helm so that the course (or direction) of the ship changes. An explanation for the acceptability of the examples above for some speakers might be that *roer* is used as referring directly to RICHTING ('direction'). See Section 7.4 for further discussion on this point.

The corpus data do not show much variation. An example of premodification is given in (86). According to my intuitions, only an external reading is possible (see the translation *id1*), but for others an internal reading as in *id2* may also be plausible.

(86) Niet alleen liet hij het publiek dinsdag maar liefst anderhalf uur wachten, hij stelt zijn achterban ook steevast op de proef door bijna per cd het muzikale roer rigoureus om te gooien.
id1. 'Not only did he let the audience wait for no less than one hour and a half, he also tried his supporters' mettle by musically

changing the direction almost with every cd.'
id2. 'Not only did he let the audience wait for no less than one hour and a half, he also tried his supporters' mettle by changing the musical direction almost with every cd.'

This concludes the interpretation of the set of idioms the data of which contain at least one example with a demonstrative determiner. A discussion of the overall results will follow after the upcoming part in which I analyse the idioms with no demonstrative determiner among the corpus examples.

de aftocht blazen

Although no examples of a demonstrative determiner specifying the idiom noun are among the corpus data, this does not necessarily mean that the idiom parts have no idiomatic referents. Whereas the combination as a whole is idiosyncratic and does not have a literal meaning, the noun *aftocht* literally means 'retreat' and can be said to also have this meaning in its idiomatic interpretation, see the concept mapping in Figure 7.14.

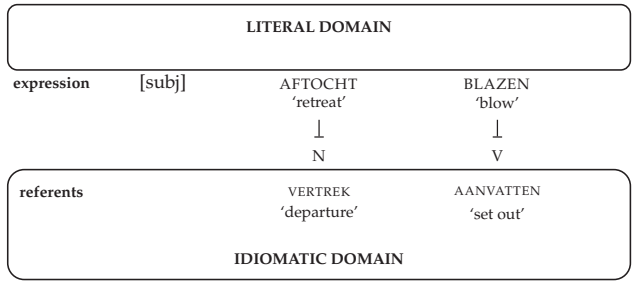


Figure 7.14 *Concept mapping of the idiom* ***de aftocht blazen***

According to my intuitions, an example containing a demonstrative determiner, such as in (87), is perfectly well-formed.

(87) Het is sneu dat hij na jarenlange revalidatie, bij zijn
it is unfortunate that he after years-of revalidation at his
rentree op het veld zo'n aftocht moet blazen.
comeback on the field such departure must blow

id. 'It is a pity that, after years of revalidation, he has to set out such a departure at his comeback on the field.'

Other examples are given in (88)-(91). The examples (88) and (89) have been found in the corpus and are examples of determiner alternation and premodification. The examples in (90), which is an example of pronominal reference, and (91), which is an example of postmodification by a *van*-phrase, have been constructed. All examples are well-formed and compatible with the hypothesis.

(88) Jeugdidool Lleyton Hewitt moest zijn vernederende aftocht nog blazen toen Kristie Boogert op Wimbledon voor een kleine sensatie zorgde.
id. 'The young's idol Lleyton Hewitt still had to set out his humiliating departure, when Kristie Boogert caused a little sensation at Wimbledon.'

(89) Maar deze keer moest Forrester een pijnlijke aftocht blazen.
id. 'But this time, Forrester had to set out a painful departure.'

(90) Het is sneu dat hij na jarenlange revalidatie, bij zijn rentree op het veld een aftocht moet blazen die zo vernederend is.
id. 'It is a pity that, after years of revalidation, he had to set out a departure that is such humiliating.'

(91) Jeugdidool Lleyton Hewitt moest de aftocht van de vernedering nog blazen toen Kristie Boogert op Wimbledon voor een kleine sensatie zorgde.
id. 'The young's idol Lleyton Hewitt still had to set out the departure of the humiliation, when Kristie Boogert caused a little sensation at Wimbledon.'

de bakens verzetten

Since no examples with a demonstrative determiner have been found in the corpus, I have presented the constructed example in (92) to the panel; only one subject finds this example well-formed in its idiomatic meaning, whereas two subjects judged it as ill-formed and five subjects judged it as borderline example.

(92) Nu zij zelf aan de macht zijn, verzetten ze die
now they themselves at the control are move they those
bakens tamelijk ingrijpend.
beacons rather radical
lit. 'Now that they are in control, they move those beacons rather radical.'

This idiom can be compared with the idiom *het roer omgooien*, which literally denotes the action of shifting the helm so that the course of the ship changes. The idiom *de bakens verzetten* literally denotes the action of moving the beacons to change the sea route (for instance in the situation of changing tides). When interpreting this action in a more abstract domain it can be said that if one moves the beacons, one sets out a new course. The individual idiom parts have no idiomatic referents, since the noun *bakens* does not refer to any of the parts in the denotation. The concept mapping for this idiom is represented in Figure 7.15.

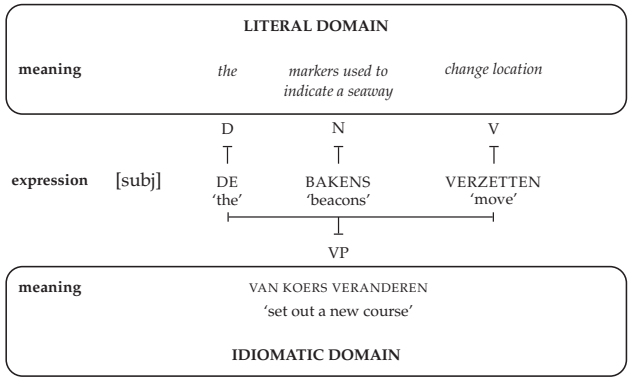


Figure 7.15 *Concept mapping of the idiom* ***de bakens verzetten***

Since the individual parts have no idiomatic referents, it is expected that no variation that requires an idiomatic referent occurs in the corpus. However, variation does occur, see (93)-(95) for some examples,[7]

[7]Note that for this idiom, 21 examples with a possessive pronoun have been found in the corpus data, see e.g. (94). This type of variation often occurs irrespective of whether the idiom parts have idiomatic referents (cf. e.g. *de/zijn hielen lichten* (lit. 'to lift the/one's heels', id. 'to take to one's heels') and therefore it is not necessarily an argument in favor of the idiom parts having idiomatic referents.

and just as with the idiom *het roer omgooien*, it is reasonable to believe that the judgements about various examples will differ from speaker to speaker depending not only on how the idiom's meaning is stored in the mental lexicon, but also on how the idiom's idiomatic meaning is associated with its literal meaning. More research to this type of idioms is needed to establish whether such idioms are counterexamples of the theory presented here or that they constitute a class that is not yet covered by the theory (see Section 7.4).

(93) Maar het is nog steeds een goede mogelijkheid om een paar
but it is yet always a good opportunity to a few
bakens te verzetten, intellectueel én in het eigen leven.
beacons to move, intellectually and in the own life.

(94) Het bureau zegt dat de kleinere winkels hun bakens
the department says that the smaller shops their beacons
moeten verzetten.
must move
id. 'The department says that the smaller shops must set out a new course.'

(95) In ieder geval is het nuttig om de grote kostenposten goed op een rij te zetten en om (waar nodig) tijdig de financiële bakens te verzetten.
id. 'Anyhow, it is useful to list the large debit items well and (if necessary) to financially set out a new course in time.'

Not only the example in (92), but also example (96) was judged by the panel: four subjects find it well-formed, three find it ill-formed and one has judged it as a borderline example. It seems that in this example the idiom is used in another meaning than 'to set out a new course', viz. in the sense of 'to push back the frontiers'. This is a plausible meaning of the idiom, since beacons are literally markers that are used to set the frontiers of a seaway. More research is needed to determine whether this idiom has indeed multiple senses.

(96) (De Japanner Takeru Kobayashi heeft een vrijwel niet te overtreffen nieuw record gevestigd: 50 hotdogs eten, broodje en al, in 12 minuten.)

"Hij heeft de bakens van de sport voor altijd verzet",
he has the beacons of the sport for ever moved
oordeelde Tom Maher van de Internationale Bond voor
judged Tom Maher of the International Federation for
Wedstrijdeten.
Competitive-Eating

id. '(The Japanese Takeru Kobayashi has set a new record that can hardly be surpassed: eating 50 hot dogs, with roll and all, within 12 minutes.)
"He has pushed back the frontiers of the sport once and for all", judged Tom Maher of the International Federation for Competitive Eating.'

de benen nemen

An example of this idiom is given in (97), which was judged as well-formed in its idiomatic meaning by all of the subjects.

(97) Voordat ze de benen namen, overmeesterden ze twee
before they the legs took overpowered they two
bewakers en stalen enkele auto's.
guards and stole a-few cars

id. 'Before they ran off, they overpowered two guards and stole a few cars.'

The idiom's idiomatic meaning is 'to run off' and there is no reason to assume that the idiom parts have idiomatic referents. The concept mapping of this idiom is represented in Figure 7.16.

As a consequence of the hypothesis, it is expected that examples showing types of variation that require the idiom parts to have idiomatic referents are not possible. An example with a demonstrative determiner has been constructed, see (98), but is only possible (with some fantasy) when interpreted literally. The example in (99) is the only example among the corpus data that shows variation that requires the noun to have an idiomatic referent. This example should be regarded as wordplay: *de kleine, snelle jongen* refers to a famous sprinter, and it seems that the use of *beide benen* uses the literal meaning of *benen* to create a stylistic effect: the famous sprinter both literally takes both (of his)

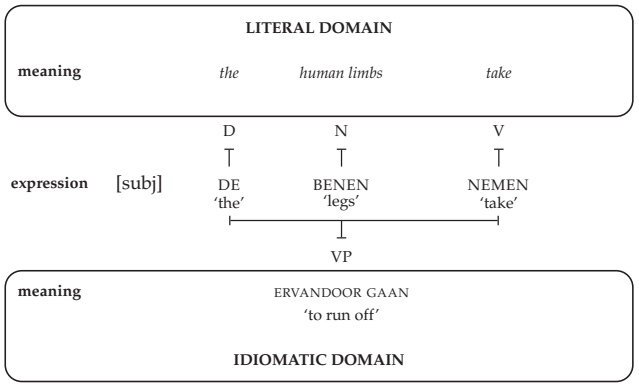


Figure 7.16 *Concept mapping of the idiom* ***de benen nemen***

legs, which he will bring into action for another country, and he runs off to another country.

(98) # Hij heeft de wet gebroken en heeft toen die benen
he has the law broken and has then those legs
genomen.
taken

(99) # Gelukkig waren er mensen in andere landen die hem
Luckily were there people in other countries who him
hielpen. Dat vond de kleine, snelle jongen zo leuk
helped. That found the little, fast boy so pleasant
dat hij beide benen nam en naar het kleine land in het
that he both legs took and to the little country in the
noorden vol met aardige mensen liep.
north full with nice people walked

Since the idiom does not show any systematic variation that requires an idiomatic referent, it can be concluded that the data of this idiom support the Idiom Variation Potential Hypothesis.

de boventoon voeren

The example in (100) is an example found in the corpus and showing no special type of variation. No examples with a demonstrative de-

terminer are among the corpus data. The example in (101) has been constructed and judged by the panel: six subjects find it ill-formed and two subjects judged it as borderline example.

(100) Op de bovenste verdieping waar vooral tassen zijn
on the top floor where especially bags are
uitgestald, voert grijs de boventoon.
exposed leads grey the dominant-tone
id. 'On the top floor, where especially bags are exposed, grey dominates. '

(101) * Het CDA en de VVD voeren nu die boventoon, en
the CDA and the VVD lead now that dominant-tone and
die staan bij mij niet bekend om hun politieke
those are with me not known because-of their political
vernieuwing.
renewal

The idiom's idiomatic meaning is 'to dominate', and the idiom parts do not have idiomatic referents, see the concept mapping in Figure 7.17.

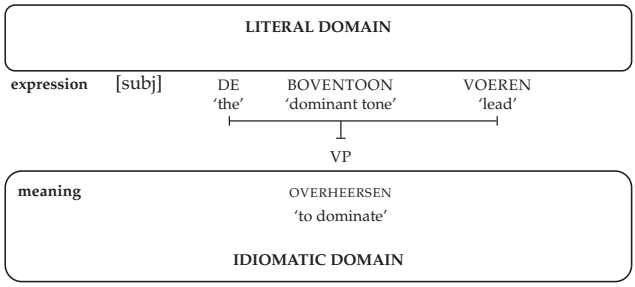


Figure 7.17 *Concept mapping of the idiom* ***de boventoon voeren***

As hypothesized the idiom cannot occur in types of variation that require the idiom parts to have idiomatic referents. Although the corpus examples do not show much variation, one variant has been found showing determiner alternation, see example (102), and another variant in which the idiom NP is topicalized, see (103). According to my intuitions, both examples are ill-formed and should not be regarded as counterexamples of the theory.

(102) * Anders dan Renate Rubinstein en Hugo Brandt Corstius
other then Renate Rubinstein and Hugo Brandt Corstius
voerde hij geen boventoon in felle publieke debatten,
led he no dominant-tone in fierce public debates,
zoals die over de kruisraketten.
like those about the cruise-missiles

(103) * De boventoon voeren de neonbalken in alle kleuren,
the dominant-tone lead the neon-lights in all colors,
die de aanwezigheid van casino's en gokhallen
that the presence of casino's and gambling-joints
kenbaar maken aan het publiek.
known make to the public

A total of 13 examples with premodification are among the corpus data. Since the idiom parts do not have idiomatic referents, only an external reading is predicted to be possible. The adjectives can be divided in three types: (1) so-called domain delimiters (see Section 6.3.1), e.g. example (104); (2) adjectives that are compatible with the literal meaning of *boventoon* ('dominant tone'), which is present in this expression, even though the idiom as a whole does not have a literal meaning (see (105) for an example); and (3) other types of adjectives, see the the example in (106) and (107).

(104) Bijna een eeuw lang stond de Maasstad bekend als het Rode Bolwerk, waar eerst de SDAP en na de Bevrijding de PvdA de politieke boventoon voerde.
id. 'Almost for a century, the Maasstad was known as the Rode Bolwerk, where first the SDAP and after the liberation the PvdA dominated politically.'

(105) De deur is nog niet dicht of ik hoor het jonge stel schreeuwen. Vooral zij voert de schelle boventoon. Even later wordt het verbale geweld buiten, pal onder mijn raam, voortgezet.
id. 'The door has yet not closed or I hear the young couple scream. Especially she dominates with a shrill voice. Later the verbal violence is continued outside, just below my window.'

(106) De kleur rood voert de absolute boventoon, gevolgd door roze.
id. 'The color red absolutely dominates, followed by pink.'

(107) Daarnaast zit een vak Duitse Halbstarken, met sjaals en petjes, trommels en tamboerijn, ook keurig, zeker wel, maar duidelijk de verbale boventoon voerend.
id. 'Next to that is a section of German Halbstarken, with scarfs, drums and tambourine, also nice, for sure, but obviously dominating verbally.'

All the adjectives can only be interpreted as modifying the whole idiom and not as modifying the idiom noun, see the idiomatic translations. It should be noted that the adjective *schelle* ('shrill') in (105) can be interpreted as semantically modifying *boventoon*, but not in this combination, i.e. when *boventoon* is combined with *voeren*, but only for example in a combination such as in (108).

(108) Door een verkeerde afstelling van de stereo, is er steeds een schelle boventoon te horen.
lit. 'Because of a wrong adjustment of the stereo, one constantly hears a shrill dominant tone.'

I conclude that the examples are compatible with the Idiom Variation Potential Hypothesis.

de duimschroeven aandraaien

Both the examples in (109) and (110) have been presented to the panel. The example in (109) was judged as well-formed by all eight subjects, while three subjects find example (110) well-formed, two subjects find it ill-formed, and three have judged it as borderline example.

(109) Bang om stroppen te lijden draaien banken de
afraid to raw-deals to suffer tighten banks the
duimschroeven aan.
thumb-screws on
lit. 'Afraid to suffer from raw deals, banks tighten the thumb screws.'
id. 'Afraid to suffer from raw deals, banks increase the pressure.'

(110) Daarna werden die duimschroeven nog sterker
afterwards were the thumb-screws yet stronger
aangedraaid, met alle desastreuze gevolgen van dien.
tightened with all disastrous consequences accordingly

The disagreement between the subjects about (110) can perhaps be explained by how the idiom is internalized, but without further analysis and additional data, I must conclude that it forms a problem for my theory. Again we are dealing with an idiom that literally denotes an event that can be associated with the idiom's idiomatic meaning when interpreted in a more abstract domain. This may lead to the plausibility of some types of variation, but more research is needed to draw accurate conclusions.

I tentatively assume that the idiom parts do not have idiomatic referents, see the concept mapping in Figure 7.18.

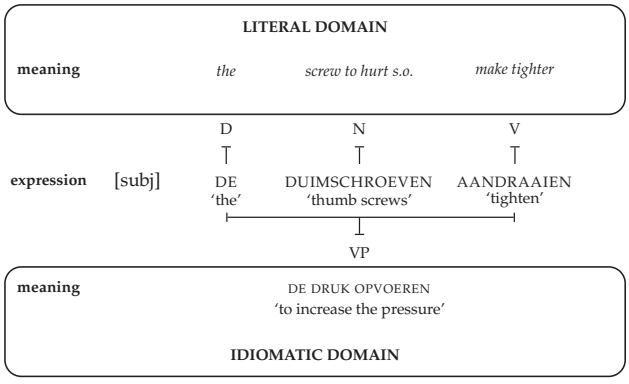


Figure 7.18 *Concept mapping of the idiom* ***de duimschroeven aandraaien***

As a consequence, examples of modification are predicted to be only interpretable as external modification. An example of premodification is given in (111), for which an internal reading is not possible (see the translation for the external reading). The example in (112) is an example of postmodification by a *van*-phrase, and it is the question whether this variant is an example of the the idiom *de duimschroeven aandraaien* or of its near-equivalent *iemand de duimschroeven aandraaien* ('to tighten the screws on on s.o.'). Either way, I do not see how the *van*-phrase can be interpreted as modifying the idiom noun, but only as modifying

the idiom as a whole which comes close to the idiom *iemand de duimschroeven aandraaien*, see the translation.

(111) Daarna werden de financiële duimschroeven nog sterker aangedraaid, met alle desastreuze gevolgen van dien.
id. 'Next, the pressure had been increased financially, with all disastrous consequences accordingly.'

(112) Draai de duimschroeven van de Iraakse dictator Saddam Hussein aan en verlicht tegelijkertijd het lijden van zijn onderdanen.
id. 'Increase the pressure on the Iraqi dictator Saddam Hussein and at the same time relieve the suffering of his subjects.'

This idiom will be part of the discussion in Section 7.4.

de kroon spannen

The example in (113) has been found in the corpus and judged as well-formed in its idiomatic meaning by seven subjects, while one subject judged it as borderline example.

(113) Zoals dat hoort spant de zanger van de groep de kroon.
like that belongs stretches the singer of the group the crown
id. 'The singer of the group beats everything, just like it should be.'

The idiom's idiomatic meaning is 'to beat everything' and there is no reason to assume that the idiom parts have idiomatic referents. The concept mapping of this idiom is represented in Figure 7.19.

According to the hypothesis, variation that requires idiomatic referents is blocked. Besides four examples of modification, the corpus examples do not show any variation. The constructed example in (114) is ill-formed and hence compatible with the hypothesis.

(114) * De concentratie die de musici van het Ives Ensemble aan de dag leggen is al bewonderenswaardig, maar Sunds fysieke zelfbeheersing spant die kroon.
the concentration that the musicians of the Ives Ensemble to the day lay is already admirably, but Sund's physical self-control stretches that crown

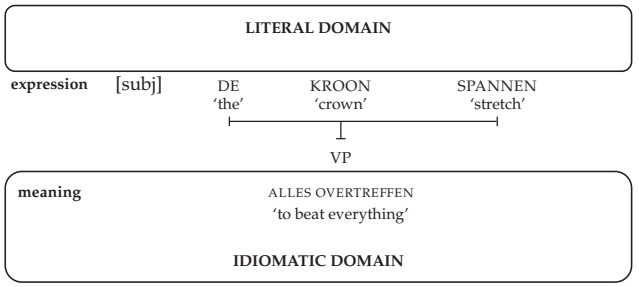


Figure 7.19 *Concept mapping of the idiom* ***de kroon spannen***

The example in (115) is an example of premodification by a geographical adjective, and can only be interpreted externally. The examples in (116) and (117) are examples of postmodification by a *van*-phrase, and intuitively I find these examples ill-formed although the modifiers can be interpreted externally as domain delimiters, see the translations. Either way, an internal reading of the *van*-phrase semantically modifying the idiom noun is impossible, and hence compatible with the hypothesis.

(115) De overnames dit jaar van American Bankers Insurance door Fortis en Wang Global door Getronics spanden de Nederlandse kroon.
id. 'This year's taking-overs of the American Bankers Insurance by Fortis and Wang Global by Getronics beat everything within the Netherlands'

(116) (Vredesoperaties gaan allang niet meer om de vrede, maar zijn omgevormd tot een aantal klussen die we aankunnen zonder kleerscheuren op te lopen. Het zijn klussen met een hoog 'doen alsof'-gehalte.)

Essential Harvest spant de kroon van deze nieuwe
Essential Harvest stretches the crown of this new
benadering.
approach

id. '(For a long time, peace operations are not about peace anymore, but have been transformed into a number of jobs that we

can do without being hurt. Those are jobs with a high proportion of faking.)
Essential Harvest is the best in this new approach.'

(117) Haar rampenverhaal spant de kroon van de voorstelling.
id. 'Her disaster story beats everything in the show.'

de lakens uitdelen

The example in (118) is an example found in the corpus and not showing a special type of variation.

(118) Echtgenote en manager Sharon Osbourne lijkt de enige
spouse and manager Sharon Osbourne seems the only-one
die het hoofd koel houdt en is duidelijk degene die de
who the head cool keeps and is obviously the-one who the
lakens uitdeelt.
sheets hands-out'
id. 'Spouse and manager Sharon Osbourne seems the only one who keeps a cool head and she is obviously the boss.'

No examples with a demonstrative determiner specifying the idiom noun are among the corpus data, even though four examples with a determiner other than a demonstrative or definite article have been found, see e.g. the examples in (119) and (120). A constructed example with a demonstrative is given in (121). Both example (119) and (121) have been judged by the panel. Example (119) was judged as ill-formed by six subjects and as borderline example by two subjects, while one subject finds example (121) well-formed, five subjects find it ill-formed and two have judged it as borderline example.

(119) # En mama deelde tot haar dood, enkele jaren geleden,
and mama handed until her death a-few years ago
vele lakens uit.
many sheets out

(120) # Tien danssloeries laten zich behandelen als huisvuil,
ten dance-sluts let themselves threat like trash
maar dat betekent niet dat hun mannen geen lakens
but that means not that their men no sheets

krijgen uitgedeeld.
get hand-out

(121) # Hij heeft wel vaker dat soort lakens uitgedeeld.
he has actually before that kind-of sheets hand-out

Geeraerts (1995) regards this idiom as having idiomatic referents, where *lakens* refers to 'orders' and *uitdelen* refers to 'give'.[8] If this is the case, then the sentences in (119)-(121) must be well-formed examples, which, according to my intuitions, is not true and this is confirmed for the examples (119) and (121) by the panel judgements. Although it is plausible that the idiom denotes 'to give orders', the idiom parts do not refer to any part of this meaning, but the meaning is assigned to the idiom as a whole. The concept mapping of *de lakens uitdelen* is given in Figure 7.20.

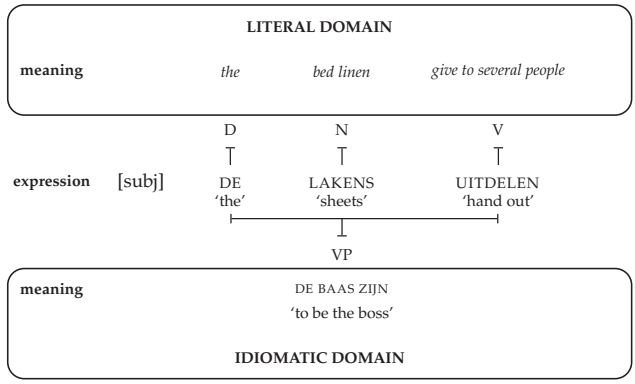


Figure 7.20 *Concept mapping of the idiom* ***de lakens uitdelen***

In this light, only an external reading of the adjective in (122) is possible, which is suggested by the translation.

(122) In de perceptie van het Israëlische publiek deelt Sharon de
in the perception of the Israeli public hands Sharon the
politieke lakens niet meer uit, maar de extreem-rechtse
political sheets not anymore out, but the extreem-right

[8]It should be noted that Geeraerts (1995) uses another terminology than *idiomatic referentiality*, viz. *isomorphism*.

vleugel van de partij.
wing of the party
id. 'In the perception of the Israeli, Sharon is politically not the boss anymore, but the extreem right wing of the party is.'

Although the corpus data do show some variation which would be a problem for my theory, I argue that the variants given in (119)-(121) are ill-formed, and that the example in (122) can only be interpreted as external modification.

de mouwen opstropen

Although no example with a demonstrative determiner has been found in the corpus, 54 examples show a determiner other than a definite article, 35 of which contain a possessive pronoun. The use of possessives is common in expressions involving body parts or clothing. The use of the possessive, which is subject-bound, does not necessarily mean that the noun has idiomatic referents, nor does the absence of a determiner,[9] see the examples (123)-(125). On the other hand, examples with a demonstrative determiner are ill-formed, see e.g. (126).

(123) Organisaties voor de rechten van de mens over de
organisations for the rights of the human around the
hele wereld stroopten de mouwen op.
whole world rolled the sleeves up
lit. 'Human rights organizations around the world rolled up the sleeves.'
id. 'Human rights organizations around the world prepared to work hard.'

(124) De Cloe hoopte dat Dijkstal de rest van de kabinetsperiode zijn mouwen gaat opstropen.
lit. 'De Cloe hoped that Dijkstal will roll up his sleeves during the rest of the cabinet term.'
id. 'De Cloe hoped that Dijkstal will prepare to work hard during the rest of the cabinet term.'

[9]For this idiom apparently general rules of the construction overrule the idiom-specific requirement that a determiner must be present.

(125) In een ander land zou men zeggen: jongens, mouwen opstropen en aan de slag om het noodlot te keren.
lit. 'In another country one would say: guys, roll up sleeves and get going to turn the fate.'
id. 'In another country one would say: guys, prepare to work hard and get going to turn the fate.'

(126) # Bij een probleem stropen militairen die mouwen op om
with a problem roll soldiers those sleeves up to
er wat aan te doen.
there something on to do

I conclude that the idiom parts do not have idiomatic referents, see the concept mapping in Figure 7.21. As a consequence, no variation that requires the idiom parts to have idiomatic referents is expected, which is confirmed by the data.

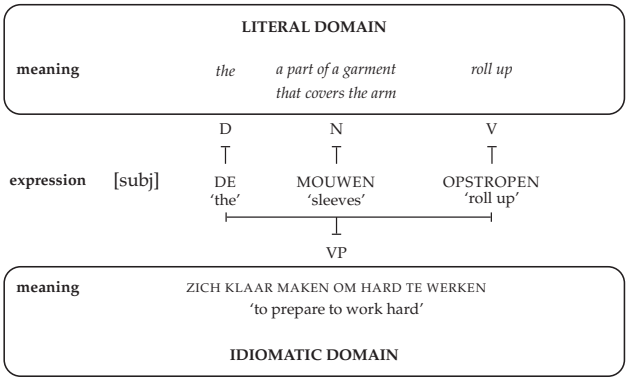


Figure 7.21 *Concept mapping of the idiom* ***de mouwen opstropen***

de noodklok luiden

The idiom *de noodklok luiden* can be placed in the same list as the idioms *het boetekleed aantrekken*, *het roer omgooien* en *de bakens verzetten*; it is another example of an expression that literally denotes an event that can be interpreted in the abstract domain. The idiom *de noodklok luiden* literally denotes the event of sounding the alarm bell to warn of

danger. Interpreted on an abstract level, it can be assumed that *de noodklok luiden* denotes 'to warn of problems'. However, I hypothesize that the idiom parts do not refer to any of the parts of this denotation and moreover do not have idiomatic referents. The idiom's literal meaning and the idiomatic referents are represented in the concept mapping in Figure 7.22.

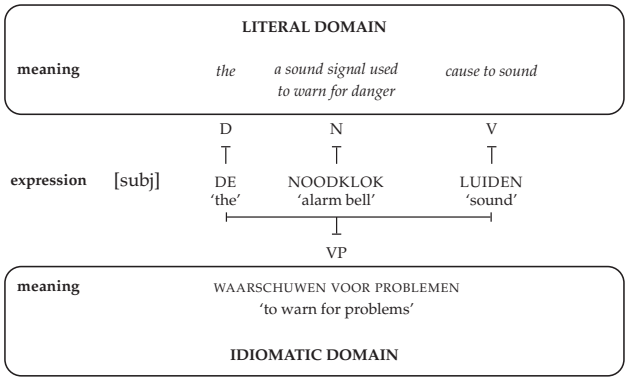


Figure 7.22 *Concept mapping of the idiom* ***de noodklok luiden***

As a consequence of the hypothesis, no variation that requires the idiom parts to have idiomatic referents is possible, but again this idiom might be of another type that allows some types of variation. The examples in (127) and (128) have been presented to the panel. The constructed example in (127), which contains a demonstrative determiner, was judged as well-formed in its idiomatic meaning by one subject, as ill-formed by three subjects and as borderline example by four subjects, while the corpus example in (128) was judged as well-formed by one subject, as ill-formed by five subjects and as borderline example by two subjects. These judgements are in line with the judgements for other idioms of this type, and most probably caused by the fact that the literal meaning of *noodklok* contributes to the idiomatic meaning of the whole. For example, the sentence in (127) can be interpreted literally (if one assumes that there is a special alarm bell for doctors that can be caused to sound). However, there is no actual *noodklok* in the idiomatic interpretation. This point will be elaborated in Section 7.4.

(127) De landelijke Huisartsen Vereniging luidde onlangs die
the national GP association rang recently that
noodklok omdat zij een groeiend tekort aan huisartsen
alarm-bell because she a growing shortage of GPs
vreest, mede als gevolg van het nieuwe belastingplan.
fears, also a consequence of the new tax-plan

(128) Het zou dan ook met een sisser zijn afgelopen, als niet de
it would then also with a squib be end, if not the
Utrechtse hoogleraar toxicologie Willem Seinen alsnog van
Utrecht Professor toxicology Willem Seinen still of
NRC Handelsblad de gelegenheid kreeg zijn noodklok te
NRC Handelsblad the opportunity got his alarm-bell to
luiden.
ring

Assuming that the idiom parts do not have idiomatic referents, only external modification is possible. Two example of premodification are given in (129) and (130).

(129) Er is alarmerend nieuws van Noorse onderzoekers die de ecologische noodklok luiden: de ijskap van de Noordpool zal nog voor 2050 gesmolten zijn.
id. 'There is alarming news of the Norwegian researchers who warn of ecological problems: the ice cap of the North Pole will be melted yet before 2050.'

(130) Vitesse is sinds het vertrek van oud-voorzitter Karel Aalbers gewend geraakt aan bestuurders en politici die eens in de zoveel tijd weer eens de financiële noodklok komen luiden.
id. 'Sinds the leaving of ex-chair Karel Aalbers, Vitesse got used to managers and politicians who warn of financial problems once in a while '

Although in the translations the modifiers semantically modify the noun *problems*, this does not mean that (129) and (130) are examples of internal modification and that *noodklok* must refer to 'problems'. That *noodklok* probably does not refer to 'problems' can be illustrated with a modifier such as *grote* ('big'), which can occur as modifier of *problemen* ('problems'), but not of *noodklok* without the expression losing its idiomatic interpretation, cf. (131) and (132).

(131) Er is alarmerend nieuws van Noorse onderzoekers die
there is alarming news of Norwegian researchers who
voor grote problemen waarschuwen.
for big problems warn
'There is alarming news of the Norwegian researchers who warn of big problems.'

(132) # Er is alarmerend nieuws van Noorse onderzoekers
there is alarming news of Norwegian researchers
die de grote noodklok luiden.
who the big alarm-bell ring

de plank misslaan

Again, even though the corpus data do not contain any examples with a demonstrative determiner, this does not necessarily mean that the idiom parts cannot have idiomatic referents. I assume that the idiom parts of the idiom *de plank misslaan* do have idiomatic referents, although the evidence presented may not be as strong as for other idioms.

The precise concept to which the noun refers is not entirely clear, i.e. it often gets its specific meaning in the context in which it is used. The concept mapping of this idiom is represented in Figure 7.23.

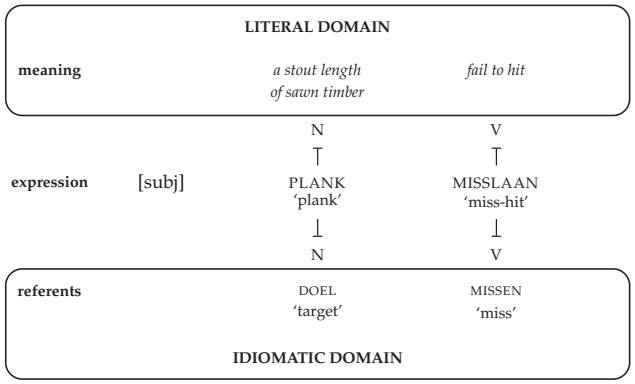


Figure 7.23 *Concept mapping of the idiom* ***de plank misslaan***

According to the hypothesis, the variation potential of this idiom is

in principle unrestricted. The corpus data do not show much variation. The examples in (133) and (134) have been taken from the internet and are well-formed in their idiomatic meaning. Some may disagree about example (133), which has been presented to the panel: four subjects find it well-formed, three subjects find it ill-formed, and one subject has judged it as borderline example.

(133) Ik vind het eerder pijnlijk om te zien dat iemand zoveel
I find it rather painful to to see that someone so-much
moeite doet om grof en lollig te zijn, en steeds maar
effort does to crude and funny to be and always just
weer die plank misslaat.
again that plank miss-hit
id. 'I find it rather painful to see that someone takes so much trouble to be crude and funny, and misses that goal every time.'

(134) Vijf minuten voor de eredivisie, waarin Jan Mulder weer elke plank misslaat. En voor de rest buitenlands voetbal.
id. 'Five minutes for the Dutch premier league, in which Jan Mulder misses again each target. And furthermore international soccer.'

Two examples of modification are given in (135) and (136). Although, for both examples it is predicted that they can be interpreted as semantically modifying the idiom noun, an external reading of example (135) is more plausible.

(135) Ondanks alle goede bedoelingen is het opvallend hoe vaak de gemeente de culturele plank misslaat.
id. 'Despites the best intentions, it is striking how many times the local council misses the aim at the cultural domain.'

(136) In "Allemaal Harvard", over de bachelors-masterstructuur en over de voor- en nadelen van brede studieprofielen, wordt de plank van de academische vorming weer eens jammerlijk misgeslagen.
id. 'In "All Harvard", about the bachelor-master division and about the advantages and disadvantages of broad study profiles, the aim of the academic education has been miserably missed.'

Another example that has been presented to the panel is the sentence in (137): six subjects find this example well-formed in its idiomatic meaning, one subject finds it ill-formed and another one has judged it as borderline example. Although the example is interpreted in its idiomatic meaning, it is the question whether it is an example of systematic variation. As mentioned in the discussion of the idiom *de kar trekken*, it is unclear whether the pronoun *hem* can be interpreted in its idiomatic sense, because it is not the direct object of the idiom verb (viz. *raken*), but of another verb, that if combined with the idiom noun can only be interpreted as a literal expression, i.e. the combination *de plank raken* is not an idiom.

(137) (Dat komt omdat Will Smith musiceert zoals hij acteert: grappig, luchtig en zonder de pretentie diepe wijsheden over de liefde te debiteren.)

Je kunt de plank niet misslaan als je hem niet probeert te
you can the plank not miss-hit if you it not try to
raken, en dus is er geen enkele reden om een hekel
hit and hence is there no single reason to a dislike
aan de zanger/rapper Will Smith te hebben.
to the singer/rapper Will Smith to have

id. '(That is because Will Smith makes music like he performs: funny, lightly and without the pretention to utter deep profundities.)
You cannot miss the target if you do not try to hit it, and hence there is not a single reason to hate the singer/rapper Wil Smith.'

de stormbal hijsen

The idiom *de stormbal hijsen* resembles the idiom *de noodklok luiden* in that they both denote an action that can be performed to warn of danger, and when interpreted in a more abstract domain, to warn of problems. The idiom *de stormbal hijsen* occurs less frequently in the corpus and shows less variation than *de noodklok luiden*. In fact, *de stormbal hijsen* does not show any variation at all. A corpus example is given in (138).

(138) We hebben niet te maken met zulke enorme afwijkingen dat we straks de stormbal moeten hijsen.
lit. 'We are not dealing with such large differences that we need to hoist the storm cone next.'
id. 'We are not dealing with such large differences that we need to warn of problems next.'

The example in (139) has been constructed and presented to the panel. It is remarkable that this example was judged similar to example (127), which contains the idiom *de noodklok luiden* with a demonstrative determiner. Three subjects find (139) ill-formed, while five subjects have judged it as borderline example.

(139) Verzekeringsmaatschappijen hesen als eerste die stormbal,
Assurance-companies hoist as first that storm-cone,
want zij zouden worden aangesproken de schade te
because they would be tackled the damage to
vergoeden die het gevolg was van het
repay that the consequence was of the
millenniumprobleem.
millenium-problem

Just as with the idiom *de noodklok luiden*, I conclude that the idiom parts of *de stormbal hijsen* do not have idiomatic referents, see the concept mapping in Figure 7.24.

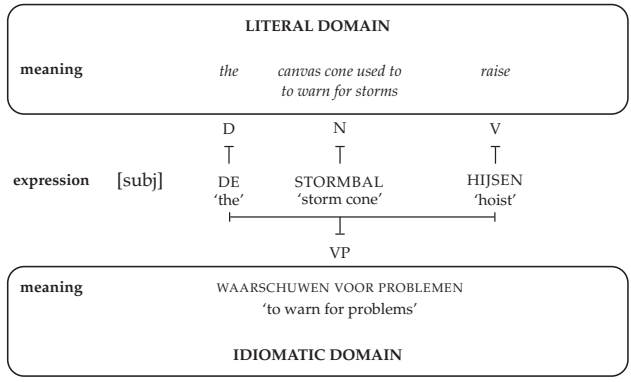


Figure 7.24 *Concept mapping of the idiom* ***de stormbal hijsen***

Since the corpus data do not show any variation, it can be concluded that the data are compatible with the hypothesis. However, more research is needed to determine how this idiom precisely relates to idioms such as *de noodklok luiden* and *het boetekleed aantrekken*.

het ijs breken

The corpus data of the idiom *het ijs breken* do not include any type of variation. A corpus example is given in (140).

(140) Dijkstal beschikt toch over alle vaardigheden om het ijs te
Dijkstal disposes yet of all skills to the ice to
breken in ongemakkelijke situaties.
break in awkward situations
lit. 'Dijkstal yet has all the skills to break the ice in awkward situations.'
id. 'Dijkstal yet has all the skills to break up a tense atmosphere in awkward situations.'

Riehemann (2001) paraphrases the English equivalent *break the ice* as 'to end the silence/tension', and categorizes it as being decomposable, i.e. *break* means 'end' and *ice* means 'silence/tension'. Although it is possible that *het ijs breken* denotes 'end the tension', it is the question whether *ijs* refers to 'tension'.

Two constructed examples have been judged by the panel. Example (141) was judged as well-formed in its idiomatic meaning by four subjects, as ill-formed by one subject and as borderline example by three subjects, while four subjects find example (142) well-formed, two subjects find it ill-formed, and two subjects have judged it as borderline example.

(141) Vaak gaat een eerste gesprek met veel spanning
often goes a first conversation with much tension
gepaard. Door een opmerking over het weer te
accompanied by a remark about the weather to
maken kun je dat ijs wellicht breken.
make can you that ice perhaps break

(142) In eerste instantie lukte het niet om het stugge ijs
at first resort succeeded it not to the stiff ice
tussen beide landen te breken.
between both countries to break

Based on these judgements, it could be argued that the interpretation of this idiom is not straightforward. According to my intuitions the examples (141) and (142) are ill-formed. I assume that the idiom parts do not have idiomatic referents, as represented in Figure 7.25. However, more constructed examples need to be judged to determine whether there is just one possible analysis for this idiom and to determine whether the speaker variation is systematic.

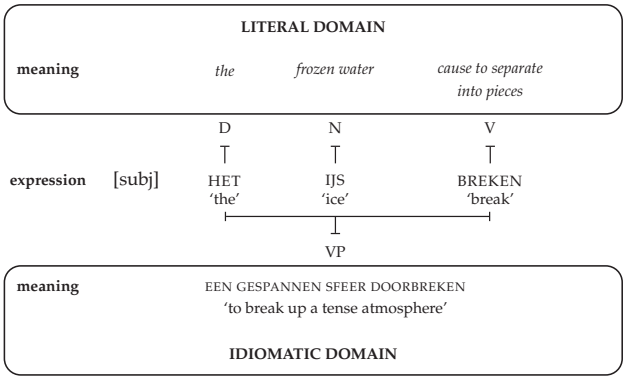


Figure 7.25 *Concept mapping of the idiom* ***het ijs breken***

het onderspit delven

The example in (143) is an example found in the corpus and showing no special type of variation.

(143) En voorlopig delft de werknemer het onderspit.
and for-now digs the employee the ONDERSPIT
id. 'And for now, the employee comes off worst.'

No examples with a demonstrative determiner are among the corpus data. The constructed example in (144), which contains a demonstrative determiner, was judged by the panel: seven subjects find it ill-formed and one subject judged it as borderline example.

(144) * Als Bradley ook in New Hampshire dat onderspit delft,
If Bradley also in New Hampshire that ONDERSPIT digs
zullen de kiezers zich gaan afvragen of hij wel
will the voters themselves go ask if he really
sterk genoeg is als kandidaat.
strong enough is as candidate

The idiom's idiomatic meaning is 'to come off worst ' and there is no reason to assume that the idiom parts have idiomatic referents. The concept mapping of this idiom is represented in Figure 7.26. It should be noted that the noun *onderspit* cannot be used outside the idiom, i.e. without the verb *delven*.

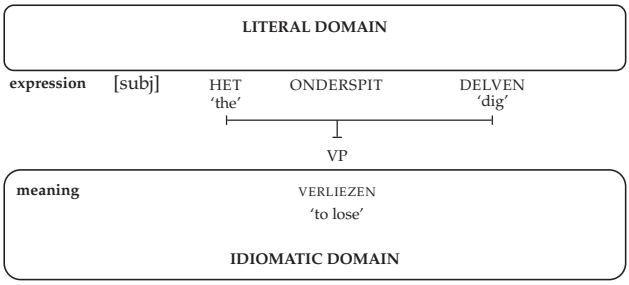


Figure 7.26 *Concept mapping of the idiom* ***het onderspit delven***

Since the idiom parts do not have idiomatic referents, it is expected that no variation that requires idiomatic referents is possible. Besides two examples with premodification, the data contain one variant with a determiner other than the definite article, see the example in (145). I find this example ill-formed; it seems like a wrong combination of *het onderspit delven* and the idiom *zijn eigen graf graven* ('to dig one's own grave').

(145) * Ze delven hun eigen onderspit.
they dig their own ONDERSPIT

Both examples of premodificaton given in (146) and (147) can only be interpreted as external modification, as also suggested by the translation.

(146) In werkelijkheid heeft Mengelberg jaar in jaar uit het mentale
in reality has Mengelberg year in year out the mental
onderspit gedolven in treiterpartijen en ordeverstoringen.
ONDERSPIT dug in tormenting and order-violation
id. 'In reality, Mengelberg has come off worst mentally year after year in tormenting and violation of the order.'

(147) Op dit moment treedt Toneelgroep Carver op met een
at this moment appears theatre-group Carver on with a
voorstelling over muizenlief, maar vooral -leed, want
performance about mouse-love but moreover -harm since
ten slotte delven de kleine onderduikers het gruwelijke
eventually dig the little animals the horrible
onderspit.
ONDERSPIT
id. 'At this moment, the theatre group Carver appears with a performance about mouse love, but moreover mouse harm, since eventually the little animals come off worst horribly.'

de/het spits afbijten

When used literally, the idiom noun *spits* must be specified by the definite determiner *de*, whereas in this idiom it can both be specified by *de* and *het*, see (148) and (149) for examples. In fact, the majority of the idioms, viz. 421 out of 498, contain the article *het*.

(148) De Rotterdamse skyline mag de spits afbijten.
the Rotterdam skyline may the peak bite-off
id. 'The skyline of Rotterdam is allowed to start.'

(149) Engel Vastgoed beet dinsdag het spits af met een gang naar
Engel Vastgoed bit tuesday the peak off with a trip to
de rechter.
the judge
id. 'Engel Vastgoed was on Tuesday the first with a trip to the judge.'

The idiom's idiomatic meaning is 'to be the first (to)/to start' and there is no reason to assume that the idiom parts have idiomatic referents. The concept mapping of this idiom is represented in Figure 7.27.

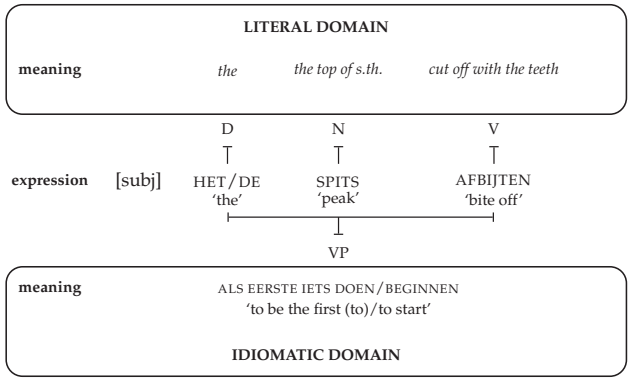


Figure 7.27 *Concept mapping of the idiom* ***de/het spits afbijten***

As a consequence of the hypothesis, variation that requires idiomatic referents is blocked. The corpus examples do not show any variation, which is compatible with the hypothesis. Moreover the constructed example in (150) was judged as ill-formed by six subjects, whereas one finds it well-formed in its idiomatic meaning and one has judged it as borderline example.

(150) # Jij mag vandaag die spits afbijten.
lit. 'You are allowed today to bite off that peak.'

This concludes the discussion of the data. In the remainder of this section, I first discuss the passivization of the idioms studied, which is followed by a summary and brief discussion of the overall results.

7.3 PASSIVIZATION

As discussed, passivization does not depend on whether a constituent has meaning, and in terms of the Idiom Variation Potential Hypothesis, does not require the idiom parts to have idiomatic referents. Whether or not an idiom is subject to passivization depends on various factors, which are not always well-understood. Van der Linden (1993) discusses the occurrence of idioms in passive constructions in more detail. He states that the non-passivizability of a large group of idioms can be accounted for by normal rules of passivizability. Some verbs – e.g. aux-

iliaries, verbs of possession, perception verbs, *to know*, etc. – simply do not passivize, which means that the idioms in which these verbs occur, do not passivize either. The non-passivizability of the remaining group of idioms is accounted for in terms of the intransitivity of the expression as a whole; an idiom loses its transitivity if it both has no idiomatic referents for the idiom parts and if there is no link between the idiom's idiomatic meaning and its literal meaning.[10] Based on the data examined in this study, I must conclude that this explanation works well for some examples, but is not completely unproblematic.

Let us turn to the passivizability of the idioms studied here. As shown in Figure 7.3, corpus examples containing passivization have been found for 19 of the 25 idioms, see Appendix B.8 for examples. This means that six idioms either cannot passivize or that there are just no examples of passivization among the corpus data. Two of these idioms, viz. *de boot afhouden* and *de dans ontspringen*, have been classified as having idiomatic referents for the idiom parts. Constructed examples of passivization are given in (151) and (152).

(151) In eerste instantie werd de boot afgehouden.
at first resort was the boat kept-off
id. 'At first, the undesirableness had been warded off.'

(152) Vorig jaar werd de dans nog door hem ontsprongen.
last year was the dance yet by him originated-from
id. 'Last year the unpleasantness was again escaped by him.'

According to my intuitions, (151) is perfectly well-formed in its idiomatic meaning. The example is (152) is a borderline example, which can also be concluded on the basis of the judgements of the panel: four subjects find the example well-formed, three subjects find it ill-formed, and one subject has judged it as borderline example. However, the contrastive judgements might also be caused by the formulation of the

[10]The link between intransitivity and non-passivizability is confusing, since it seems to conflict with the statement that "while in English it is not possible to form passive constructions on the basis of intransitive verbs, in Dutch some intransitive verbs can occur in so-called impersonal passive constructions." (Van der Linden, 1993, p. 32). For the discussion here, I will depart from Van der Linden's hypothesis that the non-passivizability of idioms relates to the simultaneous absence of idiomatic referents for the idiom parts and of a link between the idiom's idiomatic meaning and its literal meaning.

sentence. For example the sentence in (153) is more convincing as an example of passivization of *de dans ontspringen*. For both idioms I conclude that they are subject to passivization, but that just no examples have been found in the corpus.

(153) Natuurlijk jeuken je handen wel eens en moet je knarsentandend aanzien hoe soms weer eens de dans wordt ontsprongen.
id. 'Of course your fingers are itching some times and do you have to watch how now and then the persecution is escaped.'

Three idioms have been classified as having no idiomatic referents for the idiom parts. The examples in (154) and (155) are ill-formed, and since they have no idiomatic referents for the parts and because there is no link between the idiom's idiomatic meaning and its literal meaning, these idioms support the hypothesis stated by Van der Linden (1993).

(154) * Uiteindelijk werden de benen genomen door Jan.
finally were the legs taken by Jan

(155) * Uiteindelijk werd de geest gegeven door Jan.
finally was the soul given by Jan

However, other examples of idioms the parts of which have no idiomatic referents and for which there is no link between the idiomatic meaning and the literal meaning, are the idioms *het onderspit delven* and *de kroon spannen*. Some well-formed examples of passivization of the idiom *het onderspit delven* have been found in the corpus data, see Appendix B.8. Even though no examples of passivization have been found in the corpus for *de kroon spannen*, many examples can be found on the internet, see e.g. (156), which has also been presented to the panel: four subjects find the example well-formed, while the other four subjects have judged it as borderline example. Given the many plausible examples found on the internet, it can be concluded that *de kroon spannen* can be passivized and that perhaps more research is needed to determine whether examples other than *het onderspit delven* and *de kroon spannen* can be found against the hypothesis formulated by Van der Linden (1993).

(156) Terwijl bij de heren de kroon werd gespannen door Wouter
whereas by the men the crown was tightened by Wouter
Boog, ...
Boog
id. 'While in the men's league Wouter Boog had beaten everyone, ...'

The last idiom without passivization among its corpus data is *het boetekleed aantrekken*. The example in (157) has been found on the internet and is according to my intuitions well-formed. Although more research is needed to determine whether the idiom parts have idiomatic referents, it can be concluded that just no examples of passivization have been found in the corpus, but that passivization is possible.

(157) In het septembernummer van het blad Milieudefensie
in the September-edition of the magazine Milieudefensie
wordt het boetekleed aangetrokken.
is the hairshirt put-on
id. 'Guilt has been admitted in the September edition of the magazine Milieudefensie'

To conclude, in the end just two of the 25 idioms cannot be passivized, viz. *de benen nemen* and *de geest geven*. Although this can be due to the simultaneous absence of idiomatic referents for the idiom parts and of a link between the idiom's idiomatic meaning and its literal meaning, as reasoned by Van der Linden (1993), this claim is not sufficient, if we take into account the passivization of *de kroon spannen*. Hence, other factors might be responsible here.

7.4 SUMMARY AND DISCUSSION

A total of 25 OBJ1-V idioms have been examined in the light of the Idiom Variation Potential Hypothesis. For each idiom it has been decided whether the individual parts have an idiomatic referent. However, this decision needs to be taken manually and is not always straightforward. As a first step, the idioms were divided into two groups: (1) idioms the corpus data of which include one or more examples with a demonstrative determiner, and (2) idioms with no examples with a demonstrative determiner among the corpus data. This was done,

since it is expected that idioms with examples containing a demonstrative determiner most probably have idiomatic referents for the idiom parts, while idioms without demonstrative determiners among the data would probably have not.

In the subsequent step, corpus examples were analysed to determine the idiom's idiomatic meaning and to establish whether the parts of the idiom refer to parts of this meaning, and hence can be said to have idiomatic referents.

As hypothesized, idioms of which the individual parts have an idiomatic referent have in principle an unlimited variation potential. On the other hand, if the idiom parts do not have idiomatic referents then variation that requires the idiom parts to have idiomatic referents is blocked.

Based on the detailed interpretation of the data in this chapter, I conclude that the idioms studied can be divided into three groups:

1. Idioms the parts of which have idiomatic referents and for which the data support the hypothesis: *de kar trekken*, *de dans ontspringen*, *de boot afhouden*, *de handschoen opnemen*, *de ban breken*, *de boot missen*, *de bal terugkaatsen*, *de trom roeren*, *de aftocht blazen*, *de plank misslaan*.

2. Idioms the parts of which do not have idiomatic referents and for which the data support the hypothesis: *de benen nemen*, *de boventoon voeren*, *de geest geven*, *de kroon spannen*, *de lakens uitdelen*, *de mouwen opstropen*, *de stormbal hijsen*, *het ijs breken*, *het onderspit delven*, *het spits afbijten*.

3. Idioms for which it is not clear whether the idiom parts have idiomatic referents and for which the corpus examples and judgements of the panel do not give decisive answers: *het boetekleed aantrekken*, *het roer omgooien*, *de bakens verzetten*, *de duimschroeven aandraaien*, *de noodklok luiden*.

The latter group forms a problem for the theory presented. Although the existence of this type of idioms has already been observed by Nunberg (1978), who refers to this type as *abnormally decomposable* (see Section 6.1.4), it has not been taken into account in the Idiom Variation Potential Hypothesis. Studying the variation potential of these

idioms requires a different approach than the one taken here. An idiom's variation potential highly depends on how its interpretation is internalized in the mental lexicon, which may differ from speaker to speaker. This plays an even more important role with this type of idioms, because the literal meaning of the individual parts and the association with the event which the idiom denotes can be interpreted in different ways, which often varies from speaker to speaker. Based on the examination of this set of idioms, I hypothesize the following:

1. Either the idiom is interpreted as having idiomatic referents for the individual parts and as a consequence of the hypothesis it is expected that the variation potential of these idioms is in principle unlimited,

2. or the idiom is interpreted as not having idiomatic referents for the individual parts and as a consequence of the hypothesis it is expected that variation that requires idiomatic referents is blocked.

3. In the latter case, it is plausible that both the contribution of the literal meaning of the parts to the idiom's idiomatic meaning and the association with the event the idiom denotes are reason for a more flexible use of these type of expressions, i.e. the allowance of some types of variation.

More data and moreover more research is needed to gain better insights into the variation potential of this type of idioms taking into account inter alia the role of systematic speaker variation. Furthermore, it needs to be further investigated whether idioms such as *de stormbal hijsen* and *de trom roeren*, which also denote an event that may be associated with the idiom's idiomatic interpretation, are part of the third class of idioms, even when the data examined in this study point otherwise.

CHAPTER 8

CONCLUSION

The second part of this dissertation studied the variation potential of idioms. It was hypothesized that the absence of idiomatic referents restricts an idiom's variation potential in that some types of variation, viz. those that require constituents to have meaning, are blocked. On the other hand, if the idiom parts have idiomatic referents then the variation potential of the idiom is unlimited, provided that the general principles of grammar, pragmatics and discourse are not violated.

In order to test the hypothesis, I followed a corpus-based approach examining the actual usage data of 25 OBJ1-V idioms. Sometimes the corpus data was not sufficient to draw accurate conclusions, and therefore constructed examples have been used in addition.

The empirical analysis has led to the following conclusions:

1. Three types of idioms can be distinguished:

    - Idioms the parts of which have idiomatic referents.
    - Idioms the parts of which do not have idiomatic referents.
    - Idioms the parts of which literally contribute to the idiom's idiomatic meaning, but for which it is unclear whether the parts actually refer to parts of the idiomatic meaning and hence whether they have idiomatic referents.

2. An idiom's variation potential primarily depends on whether its parts have idiomatic referents. For the third idiom type, more research is needed to determine the effect that the contribution of

the literal meaning to the idiomatic meaning has on the idiom's variation potential.

3. Of the two types of modification that are distinguished

   (a) internal modification is only possible if the idiom noun has an idiomatic referent, whereas

   (b) external modification is in principle possible with all idioms irrespective of whether the parts have idiomatic referents.

4. Whether or not an idiom can passivize does not depend on whether the idiom parts have idiomatic referents; the data did not show any idioms the parts of which have idiomatic referents that cannot passive, but did contain idioms without idiomatic referents for the parts that can passivize.

I conclude this part with a discussion in Section 8.1 which is followed by an overview of the main consequences of this study for the lexical representation of MWEs in DuELME (Section 8.2).

8.1 DISCUSSION

The focus of this study was on testing the Idiom Variation Potential Hypothesis for 25 idioms taking a corpus-based approach. The main advantage of using corpus data as empirical material is that idioms can be studied in their actual use. On the other hand, the number of examples found per idiom is low; although a mean of 414 examples per idiom should be sufficient when solely focusing on idiom meaning, for this study, where the primary focus is on examining idiom variation it is not enough; especially since 390 (or 94%) of these 414 examples are in one form and do not show any of the variation types studied here. The majority of the examples with variation show passivization (with a mean of eight examples per idiom), which is followed by examples with a determiner other than definite article or demonstrative determiner (mean = 5.2), examples with premodification (mean = 4.6), examples with demonstrative determiner (mean = 3), examples with postmodification by a *van*-phrase (mean = 3), examples with postmodification by a PP (mean = 0.92), examples with topicalization (mean =

0.36), examples with number variation of the noun (mean = 0.24), examples with postmodification by a relative clause (mean = 0.2), and lastly examples with pronominal reference (mean = 0.16).

The Idiom Variation Potential Hypothesis predicts that the corpus data show more variation for idioms the parts of which have idiomatic referents than for idioms the parts of which have no idiomatic referents. However, corpora do not fully reflect the variation potential of an idiom. To test (1) whether variants that are predicted, but not observed in the data are well-formed, and (2) whether variants that are not predicted and also not observed in the data are indeed ill-formed, I have created a number of test sentences. Constructing test sentences is not always easy, since one needs to make sure that one tests the right aspect. Corpus data have been used as a basis for the construction of test sentences. However, it is not without reason that some types of variation do not occur frequently in the corpus. It was particularly difficult to construct examples with pronominal reference, postmodification by a PP and postmodification by a relative clause. In the case of pronominal reference and postmodification by a relative clause, the idiom noun must be compatible with two predicates, and it is not clear whether the use of two different predicates must be regarded as wordplay, as might be the case for example (58) (partly repeated here in (158)), where the pronoun *hem*, which refers to *de kar*, is the direct object of the *duwen* and not of the idiom verb *trekken*. However, using the same predicates may sound odd, see e.g. (159) for an example of pronominal reference.

(158) "We trekken de kar niet meer, we helpen hem af en toe duwen", zegt professor Hans Opschoor.
'"We do not lead the project anymore, we help to push it further once in a while", says professor Hans Opschoor.'

(159) Als jij geen zin hebt om die kar langer te trekken, dan trekken wij hem wel.
'If you do not want to lead that project any longer, then we lead it.'

Looking at the data, postmodification by a PP only occurs as internal modification, and hence with idioms the noun of which has an idiomatic referent. Although not explicitly examined in this study, it seems that the PP not only needs to be compatible with the idiomatic

referent but also with the literal meaning of the noun. These requirements limit the number of plausible test sentences, and more research is needed to examine the precise requirements of this type of variation.

It has been concluded that of the 25 idioms examined, the data of 20 idioms support the Idiom Variation Potential Hypothesis. For the other five idioms, the data can be interpreted in two ways:

1. either the use of these idioms depends on how the idiom is stored in a speaker's mental lexicon, i.e. as either with or without idiomatic referents for the idiom parts,

2. or the idioms belong to a different class that allows some types of variation independent of whether the idiom parts have idiomatic referents, which can perhaps be explained by the contribution of the literal meaning of the parts to the idiom's idiomatic meaning and the association with the event that the idiom denotes.

Either way, further research must show (1) whether the variation potential of these idioms solely depends on how the idiom is internalized, which may vary from speaker to speaker, or that additional factors play a role, (2) whether all five idioms behave in the same way, and (3) whether there are more idioms that belong to this type.

What remains is the question whether the Idiom Variation Potential Hypothesis also holds for MWEs other than OBJ1-V idioms containing a definite article. The question mainly concerns other types of verb phrase combinations including one or more nouns that are not used literally, because (1) the variation types studied all involve nominal constituents, and (2) if the noun is used literally, such as *blunder* in the MWE *blunder maken*, then it is expected that it is not restricted in its variation potential. I will leave this point for future research and conclude this part with a discussion of the consequences of the Idiom Variation Potential Hypothesis for the representation of MWEs.

8.2 THE REPRESENTATION OF MWES REVISED

The present study has several consequences for the lexical representation of MWEs in DuELME. Let us start with the pattern descriptions. DuELME contains three pattern descriptions of the form OBJ1-V with a fixed determiner specifying the idiom noun, see Table 8.1.

Pattern	Description
EC1	Expressions headed by a verb, taking a direct object consisting of a fixed determiner and a modifiable noun.
EC3	Expressions headed by a verb, taking a direct object consisting of a fixed determiner and a limited modifiable noun.
EC7	Expressions headed by a verb, taking a direct object consisting of a fixed determiner and a unmodifiable noun.

Table 8.1 *MWE pattern descriptions of the form OBJ1-V with a fixed determiner.*

First of all, although not explicitly indicated in the previous part, the degree of modifiability as specified in the pattern descriptions concerns internal modification, i.e. the semantic modifiability of the idiom noun. However, as concluded in the present study, internal modification is only possible if the idiom parts have idiomatic referents. As addressed in the discussion of Part I (see Section 5.1), limited modifiability of the noun was introduced, since some of the idiom nouns seemed modifiable, although not unlimitedly. Although corpus data only show a finite set of possible modifiers, based on this study, it is expected that limited modifiability follows from general principles, i.e. it is expected that it is caused by the requirement that the modifier must be compatible with the properties of the idiomatic referent in order to be plausible. Furthermore, because the variation potential of idioms, which includes determiner alternation, depends on whether the idiom parts have idiomatic referents, it is expected that determiners are only fixed when the idiom parts do not have idiomatic referents. This means that with respect to the pattern descriptions in DuELME, it is sufficient to have one pattern description instead of the three descriptions given in Table 8.1, and hence to have one equivalence class for expressions headed by a verb, taking a direct object consisting of a determiner and a noun.

As a consequence many of the MWEs probably need to be assigned a different pattern. Take for example the idiom *de boot missen*, which has been assigned the pattern *EC1*, based on the properties presented in the data record of *mis#boot* (see Section 3.1). However, after the thorough study of its corpus data it has been concluded that this idiom has idiomatic referents and that its variation potential is in principle

unlimited, hence it should not be assigned a pattern with a fixed determiner position. Besides the patterns in Table 8.1, DuELME contains another pattern description for OBJ1-V idioms, viz. *EC133* for expressions headed by a verb, taking a direct object consisting of a modifiable noun. The pattern only contains a position for the head of the NP, which means that this position is a fixed part of the expression and must be filled, whereas the rest of the NP is flexible and can be formed according to the basic principles of language. This means that this pattern can be used for idioms the parts of which have idiomatic referents, and hence can be assigned to the idiom *de boot missen*.

The idiom *de geest geven* on the other hand has been assigned the pattern *EC1*, which does not need to be changed, provided that it is indicated that patterns specifying a fixed determiner imply that the idiom parts do not have idiomatic referents.

Not only the pattern allocation needs to be reviewed, also the design of the MWE descriptions needs to be adapted so that idiomatic referents (and their properties) can be specified. Furthermore, it should be explicitly encoded in the lexical entry of an MWE whether it can passivize. In Table 8.2, the basic MWE description has been extended with (1) the ID_REFS-field, which is filled for *de boot missen*, but not for *de geest geven*; and (2) with the PASSIVIZATION-field.

EXPRESSION	boot missen	de geest geven
CL	boot missen	de geest[sg] geven
ID_REFS	mogelijkheid mislopen	n.a.
PATTERN_NAME	ec133	ec1
PASSIVIZATION	yes	no
EXAMPLE	hij heeft de boot gemist	hij heeft de geest gegeven

Table 8.2 *Two examples of revised MWE descriptions.*

Again, future research must show whether more MWE patterns need to be adapted or added to cover the variation potential of other types of MWEs in DuELME.

CHAPTER 9

GENERAL CONCLUSION AND OUTLOOK

Multiword expressions are very interesting and complex phenomena that can be studied from various perspectives. The main objectives of this dissertation were:

1. To study the lexical representation of MWEs aiming at reusability and use in NLP, and

2. to study the variation potential of MWEs using corpus data as the primary empirical material.

The first objective has led to the creation of DuELME, a Dutch Electronic Lexicon of Multiword Expressions which contains the lexical descriptions of over 5,000 MWEs and which can be integrated into Dutch NLP systems with a minimal amount of manual effort. The design, implementation and population of DuELME has been described in the first part of this dissertation.

In part II, the focus was on a small subset of MWEs taken from DuELME for which the corpus data extracted were carefully analysed to detect variation. The results have been interpreted in the light of the Idiom Variation Potential Hypothesis, which postulates that the variation potential of idioms the parts of which have idiomatic referents is in principle unrestricted, while variation that requires idiomatic referents for the parts is blocked if the idiom parts have no idiomatic referents. The hypothesis has been confirmed for 20 of the 25 idioms studied, yielding two classes of idioms, viz. (1) idioms with idiomatic referents for the individual parts, and (2) idioms without idiomatic referents for

the individual parts. For five idioms the data do not support the hypothesis and for these idioms more research is required to get a better understanding of their variation potential.

I would like to end this dissertation with some suggestions for future research.

9.1 FUTURE RESEARCH

There are a number of aspects relating to the lexical representation and variation potential of MWEs that have been left open for further research. Some of them have already been mentioned in the concluding chapters of the parts. These will be repeated here and some additional topics will be addressed.

DuELME has been designed in such a way that a wide variety of MWEs can be included. It describes a set of essential properties needed for successful conversion of the standard representation to representations required by specific systems, and forms a good basis for an even more complete description of MWEs. Several extensions and improvements of the resource are possible.

First of all, the standard lexical representation does not include the encoding of semantic properties, except for a small set of features used to encode restrictions on free subjects and direct objects. Moreover, no meaning has been assigned to the individual MWEs. The lack of representing semantic properties and meaning in DuELME may not be directly a problem, because systems such as Alpino do not (yet) contain a module for semantic analysis. However, for future purposes adding semantic information to DuELME would certainly enrich the resource.

Deeper insights into of semantic properties of MWE parts can also help to decide whether some combinations, especially noun-verb combinations, are lexically determined, i.e. whether the choice for the specific items is unpredictable and cannot be derived from the semantic properties of these items, and hence should be regarded as true MWEs that should be included in DuELME.

As mentioned in the conclusion of Part I (Chapter 5), a small number of MWEs in DuELME, viz. approximately 375, have been left for further research. These expressions need to be further examined to determine their precise pattern. Moreover it needs to be determined

whether this pattern is unique or that more MWEs in this "rest" class have the same pattern.

The knowledge gained in Part II of this dissertation must be incorporated in DuELME. As a consequence, the MWE pattern allocation needs to be reviewed and MWE descriptions should be adapted to be able to encode an MWE's variation potential (see Section 8.2). Furthermore, it should be examined whether the Idiom Variation Potential Hypothesis also holds for MWEs other than the OBJ1-V idioms studied in Part II of this dissertation, and what the possible consequences are for the representation of these other MWEs in DuELME.

A small group of the idioms studied need to be further examined in order to gain better insights in their variation potential. It concerns those idioms that literally denote domain specific situations, events and actions that can be associated with the idiom's idiomatic meaning when interpreted on a more abstract level. For these idioms it is unclear to what extent other factors than idiomatic referentiality, such as e.g. the literal meaning of the idiom parts, play a role in the idiom's variation potential. Important for further examination is the influence of speaker variation on the use of these expressions and on the use of MWEs in general. Although a subset of examples have been presented to a panel of native Dutch speakers, this set appeared to be too small to draw accurate conclusions regarding the systematicity of speaker variation.

The title of this dissertation is *Unraveling Multiword Expressions*; the goal was not only to untangle large amounts of data to separate true MWEs from false MWEs, but also to find regularity in the variation potential of idioms. As concluded, for the majority of the idioms studied the variation potential can be predicted on the basis of the properties of the individual parts yielding two classes of idioms. For the other idioms it remained unclear whether their variation potential is predictable on the basis of their properties, hence more research is needed. But then again, it is the (apparent) irregularity and unpredictability of natural language that keeps many linguists busy.

APPENDIX A

FORMAT OF THE DATA RECORDS

Besides the tuple extracted, a data record contains a list of attributes and values, which differs per extracted pattern. This appendix describes the data record format for each pattern and illustrates it with an example of an actual data record. It should be noted that some examples have been slightly adapted for display reasons.

A.1 NP_V

verb#noun root form of the candidate expression separated by #

frame subcategorization frame assigned by the Alpino parser

freq absolute frequency of the tuple

corpus corpus size

hd head of the candidate expression

subject subject information (with a maximum of 10 values)

compl1 complement

hd1 head of *compl1*

dep1 dependency label of *compl1* (value: *obj1*)

mor1 number information of *hd1* (values: *sg pl*)

dim1 diminutive information of *hd1* (values: *dim nodim*)

det1 determiner information of *hd1*

premod1 premodifier information of *hd1* (with a maximum of 10 values)

postmod1 postmodifier information of *hd1* (with a maximum of 10 values)

```
mis#boot
frame    transitive 426,
freq     426
corpus   500M words
hd       mis
subject  die 35,we 35,ze 31,je 25,Nederland 23,wie 15,zij 11,niemand 11,
compl1   boot
hd1      boot
dep1     obj1 426,
mor1     sg 424,pl 2,
dim1     nodim 426,
det1     de 413,NO 3,deze 3,die 2,welk 1,allerlei 1,geen 1,elk 1,zijn 1,
premod1  NO 420,digitaal 2,financieel 1,pan_Aziatisch 1,laat 1,
postmod1 NO 400,van 9,naar 7,in 2,tegen 1,boot 1,AD 1,met 1,voor 1,
```

A.2 (NP)_PP_V

verb#(NP)#preposition#noun root form of the candidate expression separated by #. The variable noun phrase (NP) is either *nul*, i.e. there is no variable NP, or *np*, i.e. there is a variable NP.

frame subcategorization frame assigned by the Alpino parser

freq absolute frequency of the tuple

corpus corpus size

hd head of the candidate expression

subject subject information (with a maximum of 10 values)

comp1 first complement. This takes either the value NO if there is no variable NP, or a list of the 10 most occurring values.

dep1 dependency label of *comp1* (value: *obj1*)

comp2 second complement

dep2 dependency label of *comp2*

hd2 head of *comp2*

hdcomp2 head of the complement of *comp2*

mor2 number information of *hdcomp2* (values: *sg pl*)

dim2 diminutive information of *hdcomp2* (values: *dim nodim*)

det2 determiner information of *hdcomp2*

premod2 premodifier information of *hdcomp2* (with a maximum of 10 values)

postmod2 postmodifier information of *hdcomp2* (with a maximum of 10 values)

```
sta#nul#onder#druk
frame     nonp_copula 4833,ld_pp 819,so_nonp_copula 140,intransitive 127,
freq      5987
corpus    500M words
hd        sta
subject   die 232,hij 140,koers 134,relatie 117,resultaat 117,prijs 109,
comp1     NO 5987,
dep1      obj1
comp2     onder druk
dep2      predc 4946,ld 613,mod 428,
hd2       onder
hdcomp2   druk
mor2      sg 5987,
dim2      nodim 5987,
det2      NO 5915,een 26,de 17,welk 6,zo'n 6,geen 4,enig 3,veel 2,die 2,
premod2   NO 4457,groot 695,zwaar 447,neem_toe 88,enorm 51,sterk 36,
postmod2 NO 4987,van 525,door 243,na 40,vanwege 29,als 29,in 20,
```

A.3 NP_NP_V

verb#noun#noun root forms of the candidate expression separated by #.

frame subcategorization frame assigned by the Alpino parser

freq absolute frequency of the tuple

corpus corpus size

hd head of the candidate expression

subject subject information (with a maximum of 10 values)

comp1 first complement

det1 determiner information of *comp1*

premod1 premodifier information of *comp1* (with a maximum of 10 values)

postmod1 postmodifier information of *comp1* (with a maximum of 10 values)

mor1 number information of *comp1* (values: *sg pl*)

dim1 diminutive information of *comp1* (values: *dim nodim*)

dep1 dependency label of *compl1* (value: *obj1*)

hd2 head of the second complement

mor2 number information of *hd2* (values] *sg pl*)

dim2 diminutive information of *hd2* (values] *dim nodim*)

det2 determiner information of *hd2*

premod2 premodifier information of *hd2* (with a maximum of 10 values)

postmod2 postmodifier information of *hd2* (with a maximum of 10 values)

dep2 dependency label of *hd2* (value: *obj2*)

```
bind_aan#bel#kat
frame    ninv(np_np,part_np_np(aan)) 121,
freq     121
corpus   500M words
hd       bind_aan
subject  die 53,iemand 12,hij 10,initiatief_groep 7,fractievoorzitter 3,
comp1    bel 121,
det1     de 121,
premod1  NO 121,
postmod1 NO 121,
mor1     sg 121,
dim1     nodim 121,
dep1     dr1(obj1) 121,
hd2      kat
mor2     sg 121,
dim2     nodim 121,
det2     de 121,
premod2  NO 121,
postmod2 NO 121,
dep2     dr2(obj2) 121,
```

A.4 A_N

adjective#noun root forms of the candidate expression separated by #

freq absolute frequency of the tuple

corpus corpus size

hd head of the candidate expression

hdmod head of the modifier

dep dependency information of the whole candidate expression

mor1 number information of *hd* (values: *sg pl*)

dim1 diminutive information of *hd* (values: *dim nodim*)

det1 determiner information of the whole candidate expression

premod1 premodifier information of the whole candidate expression (with a maximum of 10 values). The first value is the adjective that forms the candidate expression.

postmod1 postmodifier information of the whole candidate expression (with a maximum of 10 values)

```
open#dag
freq     161
corpus   80M words
hd       dag
hdmod    open
dep      mod 89,obj1 33,su 22,pc 10,ld 4,predc 2,obj2 1,
mor1     sg 130,pl 31,
dim1     nodim 161,
det1     de 94,een 34,NO 19,deze 3,haar 3,die 2,hun 1,of 1,zulk 1,
premod1  open 151,jaarlijks 4,eerste 3,landelijk 1,jaar 1,jaarlijke 1,
postmod1 NO 94,van 38,in 7,voor 6,op 6,die 2,bij 2,naast 1,11 1,Oude 1,
```

A.5 N_PP

noun#preposition#noun root forms of the candidate expression separated by #.

freq absolute frequency of the combination noun1#prep followed by the absolute frequency of the whole candidate expression

corpus corpus size

hd head of the candidate expression

det1 determiner information of *hd*

premod1 premodifier information of *hd* (with a maximum of 10 values)

postmod1 postmodifier information of *hd* (with a maximum of 10 values)

mor1 number information of *hd* (values: *sg pl*)

dim1 diminutive information of *hd* (values: *dim nodim*)

compl complement

hd1 head of the complement

hdcomp head of the complement of *hd1* followed by

1. A number representing the difference between the relative frequency of the noun with the highest relative frequency and the average relative frequency of each noun in the set of all nouns, and
2. The label *VAR* or *FIXED*: *VAR* if the difference is smaller than 0.8, assuming that the noun with the highest frequency is a variable direct object of the preposition, and *FIXED* if the difference is bigger than 0.8, assuming that the noun with the highest frequency is a fixed direct object of the preposition.

det2 determiner information of *hdcomp*

premod2 premodifier information of *hdcomp* (with a maximum of 10 values)

postmod2 postmodifier information of *hdcomp* (with a maximum of 10 values)

mor2 number information of *hdcomp* (values: *sg pl*)

dim2 diminutive information of *hdcomp* (values: *dim nodim*)

```
raad#van#bestuur
freq     5745 2583
corpus   ca.160M words
hd       raad
det1     de 5112,een 384,NO 99,zijn 42,deze 24,Fokker 18,geen 15,soort 9,
premod1  NO 5040,nieuw 69,Europees 66,wijs 27,heel 27,ook 27,eigen 18,
postmod1 van 5739,die 3,council 3,
mor1     sg 5568,pl 177,
dim1     nodim 5745,
compl    van bestuur
hd1      van
hdcomp   bestuur 0.45 (VAR)
det2     NO 2577,het 6,
premod2  NO 2583,
postmod2 NO 1536,van 987,die 6,waarvan 6,Publieke 6,waaronder 6,met 3,
mor2     sg 2583,
dim2     nodim 2583,
```

A.6 P_N_P

preposition#noun#preposition root forms of the candidate expression separated by #

freq absolute frequency of the tuple

corpus corpus size

hd head of the candidate expression

compl complement

det determiner information of the noun

premod premodifier information of the noun

postmod postmodifier information of the noun

mor number information of the noun

dim diminutive information of the noun

```
in#plaats#van
freq    10242
corpus  ca.400M words
hd      in
compl   plaats van
det     NO 9811,de 404,een 7,zijn 7,hun 2,die 2,het 2,deze 1,600 1,elk 1,
premod  NO 10196,eerste 33,tweede 3,ander 2,meest 1,smerig 1,divers 1,
postmod NO 9835,van 398,om 2,maar 2,voor 1,Jan 1,Pordenone 1,in 1,op 1,
mor     na 9835,sg 400,pl 7,
dim     nodim 10242,
```

APPENDIX B

CORPUS DATA AND CONSTRUCTED EXAMPLES

Appendix B.1-B.7 lists, for a number of idioms, examples that have not been presented in the main text, but that provide additional evidence for the Idiom Variation Potential Theory. Appendix B.8 lists for each idiom a corpus example showing passivization.

B.1 *de kar trekken*

CORPUS EXAMPLES

(160) Hij polemiseerde erover en droeg oplossingen aan en toch is hij vaak laf en inconsequent genoemd omdat hij als het erop aankwam geen enkele kar wilde trekken.
id. 'He polemized about it and came out with solutions and yet he has often been called cowardly and inconsistent, because when it came to it he did not want to lead a single project.'

(161) Ik vind het juist een teken van kracht en van vertrouwen als twee mensen samen besluiten dat één van hen de financiële kar zal trekken en de ander zich meer richt op de zorg in het gezin.
id. 'I think it is just a sign of power and trust when two people decide together that one of them leads the financial project and the other one focuses more on the care within the family.'

(162) Als enige overgebleven international moest hij de kar met jonge talenten trekken.

id. 'Being the only left-over international he had to lead the project with young potentials.'

(163) Al snel wist De Poel dat hij de kar van Netwerk niet langer moest trekken.
id. 'De Poel knew quickly that he should not lead the project of Network any longer.'

CONSTRUCTED EXAMPLES

(164) "Pas in maart zal helemaal duidelijk zijn wie welk karretje trekt", zegt De Wever.
id. '"Not until March, it will be totally clear who leads which little project", says de Wever.'

(165) Die kar zou de overheid moeten trekken.
id. 'That project, the government should lead.'

(166) Hij heeft vaker een kar getrokken die hem voor een enorme uitdaging stelde.
id. 'He has led a project before that presented him with an enormous challenge.'

(167) De overheid zou die kar moeten trekken, maar zij ziet liever dat die door anderen getrokken wordt.
id. 'The government should lead that project, but she rather sees that it is led by others.'

B.2 *de dans ontspringen*

CORPUS EXAMPLES

(168) Mede dankzij die doorgestoken Deense kaart heeft hij, als
also thanks-to that pierced Danish card has he, as
een van de weinige echte top-nazi's, de dans om de
one of the few real top-Nazis, the dance around the
galg weten te ontspringen.
gallows manage to originate-from
id. 'Partly due to that Danish frame-up, he has, as one of the few real top-Nazis, escaped the unpleasantness of being hanged.'

(169) Nederland ontsprong toen net de dans van de dalende productie, maar ontkwam niet aan de werkloosheidsstijging.
id. 'The Netherlands escaped then just the unpleasantness of decreasing production, but did not escape the increase of unemployment.'

CONSTRUCTED EXAMPLES

(170) Hij blijft erin slagen elke dans te ontspringen.
id. 'He keeps on succeeding in escaping each unpleasant thing.'

(171) Die dans is de mijne volgens mij ontsprongen, want afgezien van wat spuitwerk is mijn auto nooit gerestaureerd.
id. 'That unpleasantness, mine has escaped, I believe, because apart from some spray paint, my car has never been renovated.'

(172) Ze zijn die dans ontsprongen en ze zullen die nog vaker ontspringen.
id. 'They have escaped that unpleasantness and they will escape it more often.'

(173) Als enige ontsprong de baby de verschrikkelijke dans, die in de gruwelijke vorm van het concentratiekamp Sobibor in Polen alle naaste familieleden in één moorddadige klap uitroeide.
id. 'Being the only one, the baby escaped the terrible unpleasantness, that in the horrible shape of the concentration camp Sobibor in Poland, all fellow family members extirpated in one fell swoop.'

B.3 *de boot afhouden*

CONSTRUCTED EXAMPLES

(174) Aanvankelijk hield zij alle boten af, maar later heeft ze verschillende schrijvers kennelijk toch enkele verzen toegestuurd.
id. 'At first, she warded off all the things that are undesirable, but later she has apparently sent various writers yet some poems.'

(175) Die boot hield Vasalis aanvankelijk af, maar later heeft ze Reve kennelijk toch enkele verzen toegestuurd.

id. 'That undesirableness, Vasalis warded off at first, but later she has apparently sent Reve yet some poems.'

(176) Aanvankelijk hield Vasalis de boot naar financieel wonderland af.
id. 'At first, Vasalis warded of the undesirableness that leads to financial wonderland.'

(177) Sommige partijen willen iedere boot afhouden die vanuit Brussel komt.
id. 'Some parties want to ward off every undesirableness that comes from Brussel.'

B.4 *de handschoen opnemen*

CORPUS EXAMPLES

(178) Die handschoen neem ik graag op.
id. 'That challenge, I willingly take up.'

(179) Als (demissionaire) premier nam hij al gauw de handschoen van Heinsbroek op en is hij blijven hameren op het belang van een maatschappelijk en politiek debat over normen en waarden.
id. 'As (outgoing) prime minister he soon took up the challenge of Heinsbroek and he continued to hammer away at the importance of a social and political debate about standards and values.'

CONSTRUCTED EXAMPLES

(180) Er komt een tijd dat hij geen enkele handschoen meer opneemt.
id. 'There will be a time that he does not take up any challenge anymore.'

(181) Hij nam al eerder die handschoen op en hij zal die nog vaker opnemen.
id. 'He took up that challenge before and he will take it up more often.'

B.5 *de ban breken*

CORPUS EXAMPLES

(182) Maar vervolgens bracht invaller Houwing de 2-2 op het
but next brought substitute Houwing the 2-2 on the
scorebord en brak daarmee een ban.
scoreboard and broke with-that a spell

(183) Als de gewezen eerste man terecht zou staan, zou de
if the former first man tried should be should the
ban van de angst worden gebroken.
spell of the fear be broken

CONSTRUCTED EXAMPLES

(184) Die ban brak VVD-leider Bolkestein pas begin jaren
that spell broke VVD-leader Bolkestein just begin years
negentig.
nineties

(185) Pas begin jaren negentig heeft VVD-leider Bolkestein die
just begin years nineties has VVD-leader Bolkestein that
ban gebroken, maar eigenlijk had die al jaren eerder
spell broken but actually had that already years before
gebroken moet worden.
broken must be

(186) Pas begin jaren negentig heeft VVD-leider Bolkestein de
just begin years nineties has VVD-leader Bolkestein that
ban gebroken die al jaren eerder gebroken had moeten
spell broken that already years before broken had must
worden.
be

B.6 *de bal terugkaatsen*

CORPUS EXAMPLES

(187) Maar ook daarvan zal Niederer niet erg onder de indruk zijn,
denk ik, want die bal kaatst hij gewoon terug.

id. 'But Niederer shall also not be very impressed by that, I think, because he just rebounds that issue.'

CONSTRUCTED EXAMPLES

(188) Van Gaal probeerde de bal slim terug te kaatsen, maar daar trapte de journalist niet in en hij kaatste hem meteen weer terug.
id. 'Van Gaal tried to smartly rebound the issue, but the journalist did not fall for that and he rebound it immediately.'

(189) Met dat antwoord kaatste hij die venijnige bal weer terug.
id. 'With that answer, he rebound that spiteful remark.'

(190) Met dat antwoord kaatste hij de bal terug die hem venijnig werd toegespeeld.
id. 'With that answer, he rebound the remark that had been spitefully slipped to him.'

B.7 *de trom roeren*

CORPUS EXAMPLES

(191) Wie steeds dezelfde trom roert, verandert soms het opiniekliimaat en daarmee soms ook de cultuur van bestuur en beleid.
id. 'Who always spreads the same message, changes sometimes the public opinion and with that sometimes also the culture of management and policy.'

CONSTRUCTED EXAMPLES

(192) Die trom roeren ze in Moskou oorverdovend hard.
id. 'That message, they spread very loud in Moskou.'

(193) Zij roert de trom die lange tijd niet geroerd werd.
id. 'She spreads the message that had not been spread for a long time.'

B.8 CORPUS EXAMPLES OF PASSIVIZATION

(194) Vorig jaar werd de kar getrokken door de drie zuidelijkste
last year was the cart pulled by the three south-most
provincies: Zeeland, Noord-Brabant en Limburg.
regions Zeeland Noord-Brabant and Limburg
id. 'Last year the project had been led by the three most-southern regions: Zeeland, Noord-Brabant and Limburg.'

(195) Toch werd de handschoen opgenomen tegen de
yet was the glove picked-up against the
professionele teams van Australië, Zwitserland, Verenigde
professional teams of Australia Switzerland United
Staten en vooral Duitsland.
States and especially Germany
id. 'Yet the challenge was taken up against the professional teams of Australia, Switzerland, USA and especially Germany.'

(196) Maar omstreeks 1850 werd de ban gebroken.
but around 1850 was the spell broken
id. 'But around 1850 the spell was broken.'

(197) Niet alleen na het einde van de Tweede Wereldoorlog werd
not only after the end of the second World-War was
de boot gemist, maar vooral toen in dat paleis te Batavia.
the boat missed, but especially then in that palace at Batavia
id. 'The opportunity had not only been missed at the end of World War II, but especially then in that palace in Batavia.'

(198) Op de kritiek van kille managementbureaus werd vanuit
at the criticism of chilly management-agencies, was from
de dienst meteen de bal teruggekaatst.
the service immediately the ball hit-back
id. 'The criticism of chilly management agencies was responded to immediately from out the service.'

(199) In beide landen wordt nu driftig de nationalistische
in both countries is now vehemently the nationalistic
trom geroerd.
drum moved
id. 'In both countries, the nationalistic message is being spread

vehemently.'

(200) Begin jaren negentig werd het roer weer omgegooid.
beginning years nineties was the helm again shifted
id. 'At the beginning of the nineties, a total different direction had been gone into again.'

(201) Donderdag werd de aftocht geblazen.
Thursday was the retreat blown
id. 'On Thursday, the departure was set out.'

(202) Daarna werden de bakens niet verzet, ook niet toen het
thereafter were the beacons not moved, also not then it
daarna slecht ging.
thereafter wrong went
id. 'After that a new course had not been set out, not even when it went wrong afterwards.'

(203) Het was een mooie bijeenkomst, waar de boventoon
it was a beautiful meeting where the dominant-tone
werd gevoerd, door oude Britten met vet, grijsgeel
was led by old Brits with greasy grey-yellow
haar.
hair
id. 'It was a beautiful meeting, where the old Brits with greasy, grey-yellow hair dominated.'

(204) Begin jaren tachtig werden met de bezuinigingen van
begin years eighties were with the savings of
Deetman de duimschroeven aangedraaid.
Deetman the thumb-screws tightened
id. 'At the beginning of the nineties, the pressure was increased with the savings of Deetman.'

(205) Bij Feyenoord worden de lakens uitgedeeld door mensen
at Feyenoord are the sheets handed-out by people
die nooit hebben gevoetbald.
who never have played-soccer
id. 'At Feyenoord the boss is played by people who have never played soccer.'

(206) Er zijn dagen waarop de mouwen moeten worden
there are days on-which the sleeves need be
opgestroopt.
rolled-up
id. 'There are days on which one needs to be prepared to work hard.'

(207) Onlangs werd op een symposium de noodklok geluid.
recently was at a symposium the alarm-bell rung
id. 'Recently at a symposium, warnings for problems were sent.'

(208) Waarom zo'n ophef over een boek waarin de plank zo
why such fuss about a book in-which the plank so
wordt misgeslagen.
is missed
id. 'Why is there such a fuss about a book which is completely besides the point.'

(209) Op het hoofdkantoor van ABN Amro werd twee weken
at the head-office of ABN Amro were two weeks
geleden de stormbal gehesen.
ago the storm-cone hoisted
id. 'Two weeks ago warnings for problems were sent at the ABN Amro head office.'

(210) In Balkenendes Haagse appartement werd het ijs gebroken.
in Balkendende's Hague apartment was the ice broken
id. 'The ice was broken in Balkenende's Hague apartment.'

(211) Maar even zovele malen werd het onderspit gedolven.
but just as-many times was the ONDERSPIT dug
id. 'But just as many times one had lost.'

(212) Op Nederland 3 wordt de spits afgebeten door Buitenhof.
on Nederland 3 is the peak bitten-off by Buitenhof
id. 'Buitenhof starts on Nederland 3.'

APPENDIX C

CORPUS EXAMPLES SOURCES

This appendix lists the sources of the examples taken from the *TwNC* given the unique file names or internet given URLs.

(37) nrc20001006.xml
(38) nrc20011006.xml
(41) parool19990105.xml
(42) ad20000713.xml
(43) nrc19991110.xml
(44) trouw20001209.xml
(46) trouw20001228.xml
(47) nrc20001006.xml
(48) nrc20011006.xml
(49) vk19970711.xml
(50) nrc20011101.xml
(51) parool19990517.xml
(52) parool20010911.xml
(53) trouw19990805.xml
(54) ad19990209.xml
(55) volkskrant19991227.xml
(56) parool20010412.xml
(57) volkskrant20010321.xml

(58) nrc20020903.xml
(59) volkskrant20011030.xml
(60) nrc20000303.xml
(61) parool19991105.xml
(62) trouw20040618.xml
(63) nrc20000324.xml
(65) ad20001115.xml
(66) trouw20020528.xml
(67) vk19971125.xml
(68) trouw20000219.xml
(69) ad20040212.xml
(70) volkskrant20020327.xml
(71) nrc20010608.xml
(72) ad19990604.xml
(74) volkskrant20000613.xml
(75) vk19970913.xml
(76) trouw20001107.xml
(77) trouw20000506.xml
(79) nrc20040408.xml
(80) volkskrant19990408.xml
(81) nrc20010312.xml
(82) parool20020622.xml
(83) parool20020410.xml
(84) http://exprezziess.hyves.net/blog/16627168/Dagelijkse_Gedachte_6_3_08/Mi8x/?pageid=5T3N9IB4ZU888KSSG&PHPSESSID=0716743da76f9d336a2ecc559d0bdd14
(85) volkskrant19990211.xml
(86) trouw20000127.xml
(88) parool20000628.xml
(89) volkskrant20000225.xml

(93) nrc20011011.xml
(94) volkskrant19991019.xml
(95) trouw20000311.xml
(96) ad20010706.xml
(97) trouw20001002.xml
(99) ad19990916.xml
(100) parool19990921.xml
(102) vk19970110.xml
(103) ad20020309.xml
(104) ad20020316.xml
(105) ad19991104.xml
(106) ad20020215.xml
(107) trouw19991115.xml
(109) nrc20011025.xml
(111) volkskrant20000330.xml
(112) ad20010704.xml
(113) ad20000404.xml
(115) parool19990730.xml
(116) parool20010823.xml
(117) trouw20001009.xml
(118) ad20020503.xml
(119) trouw19990126.xml
(120) trouw20031031.xml
(122) ad20040504.xml
(123) nrc20010409.xml
(124) vk19971009.xml
(125) ad20020615.xml
(127) ad20000105.xml
(128) parool20020104.xml
(129) volkskrant20010203.xml

(130) nrc20030422.xml

(134) http://ajax.netwerk.to/forums/HTML/forum17/604-41.php

(135) nrc20020405.xml

(136) nrc20020615.xml

(137) volkskrant19990730.xml

(138) vk19970225.xml

(140) parool20020824.xml

(143) vk19970210.xml

(145) volkskrant20020528.xml

(146) volkskrant19990219.xml

(147) volkskrant20020208.xml

(148) ad20020327.xml

(149) volkskrant20041110.xml

(153) http://www.volkskrant.nl/archief_gratis/article628006.ece/Babbelende_homeopaat_is_niet_tegen_te_houden

(156) http://217.18.75.118/nl/archief/artikel/sport-kort/10518/reacties

(157) http://www.sdnl.nl/lastpak1.htm

(160) volkskrant20011116.xml

(161) trouw20020315.xml

(162) volkskrant20010618.xml

(163) ad20030607.xml

(164) http://www.standaard.be/Artikel/Detail.aspx?artikelId=dmf10022009_105

(168) parool20010317.xml

(169) parool20001227.xml

(173) http://www.genpage.nl/dresden/sara%20dresden/saradresdenkrant/saraentoneel.htm

(178) volkskrant20010531.xml

(179) parool20031206.xml

(182) volkskrant20000907.xml

(183) nrc20000904.xml
(187) trouw19991111.xml
(191) volkskrant20020323.xml
(194) vk19971217.xml
(195) trouw19991115.xml
(196) volkskrant20001215.xml
(197) trouw20001014.xml
(198) trouw19990723.xml
(199) volkskrant20000613.xml
(200) vk19970624.xml
(201) volkskrant20010728.xml
(202) nrc20030906.xml
(203) parool20010127.xml
(204) trouw20000603.xml
(205) volkskrant20040410.xml
(206) trouw20020529.xml
(207) parool20000624.xml
(208) vk19971110.xml
(209) volkskrant20010118.xml
(210) volkskrant20041109.xml
(211) ad19990529.xml
(212) volkskrant20011030.xml

BIBLIOGRAPHY

Abeillé, A. (1995), The flexibility of French idioms: A representation with Lexicalised Tree Adjoining Grammar, *in* Everaert et al. (1995), pp. 15–42.

Abeillé, A. and Schabes, Y. (1989), Parsing idioms in lexicalized TAGs, *Proceedings of the 4th conference on European chapter of the Association for Computational Linguistics*, ACL, pp. 1–9.

Baptista, J., Correia, A. and Fernandes, G. (2004), Frozen sentences of Portuguese: Formal descriptions for NLP, *Proceedings of the EACL 2004 Workshop on Multiword Expressions: Integrating Processing*, EACL, pp. 72–79.

Cacciari, C. and Tabossi, P. (eds) (1993), *Idioms: Processing, structure, and interpretation*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Català, D. and Baptista, J. (2007), Spanish adverbial frozen expressions, *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, ACL, pp. 33–40.

Chafe, W. L. (1968), Idiomaticity as an anomaly in the Chomskyan paradigm, *Foundations of Language* **4**, 109–127.

Chomsky, N. (1981), *Lectures on government and binding*, Foris Publications, Dordrecht.

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. and Flickinger, D. (2002), Multiword expressions: Linguistic precision and reusability, *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, ELRA, pp. 1941–7.

de Groot, H. et al. (1999), *Idioomwoordenboek: Verklaring en herkomst van uitdrukkingen en gezegden*, Van Dale Lexicografie, Utrecht.

Dobrovol'skij, D. and Piirainen, E. (2005), *Figurative Language: Cross-Cultural and Cross-Linguistic Perspectives*, Elsevier.

Dormeyer, R. and Fischer, I. (1998), Building lexicons out of a database for idioms, *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC 1998)*, ELRA, pp. 833–838.

Dormeyer, R., Fischer, I. and Keil, M. (1998), A database for verbal idioms, *EURALEX 1998 Proceedings*, pp. 99–109.

Ernst, T. (1981), Grist for the linguistic mill: Idioms and "extra" adjectives, *Journal of Linguistic Research* **1**(3), 51–68.

Everaert, M., van der Linden, E.-J., Schenk, A. and Schreuder, R. (eds) (1995), *Idioms: Structural and Psychological Perspectives*, Lawrence Erlbaum Associates, Hove, UK.

Fellbaum, C. (1993), The determiner in english idioms, *in* Cacciari and Tabossi (1993), pp. 271–295.

Fellbaum, C., Geyken, A., Herold, A., Koerner, F. and Neumann, G. (2006), Corpus-Based Studies of German Idioms and Light Verbs, *International Journal of Lexicography* **19**(4), 349–361.

Fillmore, C. J. (1992), "Corpus linguistics" or "computer-aided armchair linguistics", *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, Mouton de Gruyter, Berlin, pp. 35–60.

Fraser, B. (1970), Idioms within a transformational grammar, *Foundations of Language* **6**, 22–42.

Geeraerts, D. (1995), Specialisation and reinterpretation in idioms, *in* Everaert et al. (1995), pp. 57–73.

Gibbs, R. and Nayak, N. (1989), Psycholinguistic studies on the syntactic behaviour of idioms, *Cognitive Psychology* **21**(1), 100–138.

Granger, S. and Meunier, F. (eds) (2008), *Phraseology: an interdisciplinary perspective*, John Benjamins Publishing Company, Amsterdam.

Grégoire, N. (2007a), Design and implementation of a lexicon of Dutch multiword expressions, *Proceedings of the ACL 2007 Workshop on A Broader Perspective on Multiword Expressions*, ACL, pp. 17–24.

Grégoire, N. (2007b), Dutch Electronic Lexicon for Multiword Expressions: GUI user manual, Published as part of the DuELME documentation (available via the Dutch HLT agency: www.tst.inl.nl).

Grégoire, N. (2007c), MWE lexicon for Dutch: Alpino conversion, STEVIN IRME internal report published on http://www-uilots.let.uu.nl/irme/.

Grégoire, N. (2007d), MWE lexicon for Dutch: Encoding protocol, Published as part of the DuELME documentation (available via the Dutch HLT agency: www.tst.inl.nl).

Grégoire, N. (2007e), MWE lexicon for Dutch: Rosetta conversion, STEVIN IRME internal report published on http://www-uilots.let.uu.nl/irme/.

Gross, M. (1986), Lexicon-grammar. the representation of compound words., *Proceedings of COLING 1986*, pp. 1–6.

Gross, M. (1996), Lexicon-grammar, *in* K. Brown and J. Miller (eds), *Concise Encyclopedia of Syntactic Theories*, Pergamon, Cambridge, pp. 244–259.

Haeseryn, W., Romijn, K., Geerts, G., de Rooij, J. and van den Toorn, M. (1997), *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff and Wolters Plantyn, Groningen en Deurne.

Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I. and van der Wouden, T. (2003), CGN syntactische annotatie.

Katz, J. (1973), Compositionality, idiomaticity, and lexical substitution, *in* S. R. Anderson and P. Kiparsky (eds), *A Festschrift for Morris Halle*, Holt, Rinehart, and Winston, New York, pp. 357–376.

Krenn, B. (2000a), CDB - a database of lexical collocations, *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, ELRA.

Krenn, B. (2000b), *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*, PhD thesis, German Research Center for Artificial Intelligence and Saarland University Dissertations in Computational Linguistics and Language Technology.

Kuiper, K., McCann, H., Quinn, H., Aitchison, T. and van der Veer, K. (2003), SAID: A syntactically annotated idiom dataset, Linguistic Data Consortium, LDC2003T10, Pennsylvania.

Langlotz, A. (2006), *Idiomatic Creativity: A Cognitive Linguistic Model of Idiom-Representations and Idiom-Variation in English*, John Benjamins Publishing Company, Amsterdam.

Laporte, E. and Voyatzi, S. (2008), An electronic dictionary of French multiword adverbs, *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, LREC, pp. 31–34.

Martin, W. and Maks, I. (2005), *Referentie Bestand Nederlands: Documentatie*.

Mel'čuk, I. (1995), Phrasemes in language and phraseology in linguistics, *in* Everaert et al. (1995), pp. 167–232.

Moon, R. (1998), *Fixed Expressions and Idioms in English: A Corpus-Based Approach*, Oxford Studies in Lexicography and Lexicology, Clarendon Press, Oxford.

Moreno, R. E. V. (2007), *Creativity and Convention : The Pragmatics of Everyday Figurative Speech*, John Benjamins Publishing Company, Amsterdam.

Nunberg, G. (1978), *The Pragmatics of Reference*, Indiana University Linguistics Club, Bloomington, Ind.

Nunberg, G., Sag, I. and Wasow, T. (1994), Idioms, *Language* **70**, 491–538.

Odijk, J. (1993), *Compositionality and syntactic generalizations*, PhD thesis, Katholieke Universiteit Brabant.

Odijk, J. (2003), Towards a standard for multi-word expressions. ISLE Project Report.

Odijk, J. (2004a), Multiword expressions in NLP, Course presentation, LOT Summerschool, Utrecht.

Odijk, J. (2004b), A proposed standard for the lexical representation of idioms, *EURALEX 2004 Proceedings*, Université de Bretagne Sud, pp. 153–164.

Odijk, J. (2005), Standard lexical representation of multi-word expressions, Research proposal published at http://www-uilots.let.uu.nl/irme.

O'Grady, W. (1998), The syntax of idioms, *Natural Language and Linguistic Theory* **16**, 279–312.

Ordelman, R. (2002), Twente nieuws corpus (TwNC).

Pullum, G. (2009), Computational linguistics and generative linguistics: The triumph of hope over experience, *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, ACL, pp. 12–21.

Richter, F. and Sailer, M. (2002), Cranberry words in formal grammar, *in* C. Beyssade, O. Bonami, P. C. Hofherr and F. Corblin (eds), *Empirical Issues in Syntax and Semantics*, Vol. 4, Presses de l'Université Paris-Sorbonne, pp. 155–172.

Riehemann, S. (1997), Idiomatic constructions in HPSG, Presented at the 4th International Conference on HPSG.

Riehemann, S. (2001), *A Constructional Approach to Idioms and Word Formation*, PhD thesis, Stanford University.

Rosetta, M. T. (1994), *Compositional Translation*, Kluwer Academic Publishers, Dordrecht.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. and Flickinger, D. (2001), Multiword expressions: A pain in the neck for NLP, LinGO Working Paper, (2001-03).

Sailer, M. (2003), *Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar*, PhD thesis, Eberhard-Karls-Universitat Tübingen.

Schenk, A. (1994), *Idioms and collocations in compositional grammars*, PhD thesis, University of Utrecht.

Schenk, A. (1995), The syntactic behavior of idioms, *in* Everaert et al. (1995), pp. 253–272.

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

Stathi, E. (2008), Lexical and grammatical properties of idioms: A corpus-based approach to figurative verb phrases, unpublished PhD thesis.

Stock, O., Slack, J. and Ortony, A. (1993), Building castles in the air. some computational and theoretical issues in idiom comprehension, *in* Cacciari and Tabossi (1993), pp. 229–337.

Van der Linden, E.-J. (1993), *A categorial, computational theory of idioms*, PhD thesis, Katholieke Universiteit Brabant.

van der Wouden, T. (1997), *Negative Contexts: Collocation, polarity and multiple negation*, Routledge, London.

van Gestel, F. (1995), En bloc insertion, *in* Everaert et al. (1995).

van Noord, G. (2006), At Last Parsing Is Now Operational, *TALN06 Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pp. 20–42.

van Noord, G. and Bouma, G. (2009), Parsed corpora for linguistics, *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, ACL, pp. 33–39.

van Noord, G., Schuurman, I. and Vandeghinste, V. (2006), Syntactic annotation of large corpora in STEVIN, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, ELRA, pp. 1811–14.

Villada Moirón, B. (2006), Evaluation of a machine learning algorithm for MWE identification. Decision trees, STEVIN IRME internal report published on http://www-uilots.let.uu.nl/irme/.

Villada Moirón, B. (2007), A task-based evaluation of the ECM database. Effect on parsing performance., STEVIN IRME internal report published on http://www-uilots.let.uu.nl/irme/.

Villavicencio, A. and Copestake, A. (2002), On the nature of idioms, LinGO Working Paper No. 2002-04.

Villavicencio, A., Copestake, A., Waldron, B. and Lambeau, F. (2004), The lexical encoding of MWEs, *Proceedings of the ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, ACL, pp. 80–87.

Wasow, T., Sag, I. and Nunberg, G. (1983), Idioms: An interim report, *in* S. Hattori and K. Inoue (eds), *Proceedings of the XIIIth International Congress of Linguists*, pp. 102–115.

Weinreich, U. (1967), Problems in the analysis of idioms, *in* J. Puhvel (ed.), *Substance and Structure of Language*, University of California Press, Berkeley, pp. 23–81.

SAMENVATTING

Centraal in deze dissertatie staat het fenomeen *meerwoordexpressie*. Meerwoordexpressies (MWEs) zijn combinaties van woorden met taalkundige eigenschappen die niet voorspelbaar zijn uit de eigenschappen van de individuele woorden of de normale manier waarop ze gecombineerd worden. Deze taalkundige eigenschappen beperken zich niet tot een specifiek taalkundig niveau van beschrijvingen, zie de voorbeelden in (1)-(4).

(1) a. blunder maken/begaan/*doen/*slaan
    b. flater *maken/*begaan/*doen/slaan

(2) het loodje/*lood leggen

(3) in opdracht van

(4) met de handen in het haar zitten

In alle voorbeelden worden specifieke lexicale elementen gebruikt. De keuze van deze elementen is onvoorspelbaar en niet afleidbaar uit de semantische eigenschappen van de individuele woorden. Neem bijvoorbeeld de expressie in (1). Hoewel *flater* en *blunder* synoniemen zijn, zijn de werkwoorden waarmee ze kunnen optreden verschillend en onvoorspelbaar. Daarom zijn de combinaties *blunder maken*, *blunder begaan* en *flater slaan* MWEs. Een voorbeeld van een MWE met idiosyncratische morfologische eigenschappen is (2), waarin het zelfstandig naamwoord *loodje* verplicht de diminutief vorm heeft. De expressie in (3) is een MWE, omdat *opdracht* een enkelvoudig telbaar zelfstandig naamwoord is dat volgens de Nederlandse grammatica over het algemeen

voorafgegaan moet worden door een determinator, maar de expressie is *in opdracht van* en niet **in de opdracht van*.

Door hun onvoorspelbare eigenschappen vormen MWEs een groot probleem voor natuurlijke taalverwerking. Niet alleen de grammatica van een computationeel systeem moet om kunnen gaan met MWEs, maar er moet ook een grote hoeveelheid lexicale beschrijvingen van MWEs beschikbaar zijn die compatibel zijn met de vereisten van de grammatica. Hoewel onderzoek naar de behandeling van MWEs in specifieke systemen nog altijd gaande is, is het onwenselijk om voor ieder computationeel systeem een voor dat systeem specifiek lexicon te creeëren.

DuELME is ontwikkeld in het kader van herbruikbaarheid. DuELME is een acroniem voor *Dutch Electronic Lexicon of MultiWord Expressions* ('Nederlands Elektronisch Lexicon van Meerwoordexpressies'). DuELME bevat zo theorie- en implementatieneutraal mogelijke beschrijvingen van ruim 5.000 Nederlandse MWEs. Deze beschrijvingen zijn zo gerepresenteerd dat ze met een minimale hoeveelheid werk kunnen worden geïntegreerd in verschillende natuurlijke taalverwerkingssystemen.

Deel I van dit proefschrift beschrijft het ontwerp, de implementatie en het vullen van DuELME. Het ontwerp van DuELME is gebaseerd op de *Equivalentieklassenmethode* ('Equivalence Class Method' (ECM)) (Odijk, 2003). De ECM is een innovatieve methode die abstraheert van de syntactische structuur van MWEs en alleen vereist dat MWEs met dezelfde syntactische structuur gegroepeerd worden in zogenoemde *Equivalentieklassen* ('Equivalence Classes' (ECs)). Het idee achter de ECM is dat MWEs met dezelfde syntactische structuur op dezelfde manier behandeld worden in natuurlijke taalverwerkingssystemen. Door deze MWEs dus te groeperen kunnen grote aantallen MWEs op een eenvoudige manier worden geïncorporeerd in een specifiek systeem: aan de hand van een handmatig uitgevoerde conversie van één MWE uit een bepaalde EC, kunnen alle andere MWEs uit diezelfde EC volledig automatisch worden geconverteerd.

De oorspronkelijke ECM, zoals geïntroduceerd door Odijk (2003), bestaat uit een standaard lexicale representatie voor MWEs en een voorstel voor een conversieprocedure voor het incorporeren van de standaard representatie in natuurlijke taalverwerkingssystemen. Eén van de doelen van dit proefschrift is de oorspronkelijke ECM te verfijnen

en de aangepaste versie te implementeren in DuELME.

Een belangrijke verbetering van de oorspronkelijke ECM is het parameteriseren van de ECs. De oorspronkelijke ECM bevat zowel een EC *MWEp1* zoals beschreven in (5) als een een EC *MWEp2* zoals beschreven in (6).

(5)

MWE patroon	**beschrijving**
MWEp1	expressies bestaande uit een werkwoord dat een subject neemt en een direct object dat bestaat uit een determinator en een enkelvoudig zelfstandig naamwoord

(6)

MWE patroon	**beschrijving**
MWEp2	expressies bestaande uit een werkwoord dat een subject neemt en een direct object dat bestaat uit een determinator en een meervoudig zelfstandig naamwoord

Het enige verschil tussen deze twee ECs is het verschil in getal van het zelfstandig naamwoord. Door ECs op deze manier te beschrijven is de kans groot dat er veel ECs zullen zijn met maar weinig MWE-beschrijvingen. Om het aantal ECs te verminderen en het aantal MWE-beschrijvingen per EC te vermeerderen, heeft Odijk (2004b) voorgesteld om kleine lokale verschillen tussen ECs te parameteriseren. Dit betekent dat in plaats van de twee ECs *MWEp1* en *MWEp2* er maar één EC nodig is met daarin zowel expressies met een enkelvoudig zelfstandig naamwoord als expressies met een meervoudig zelfstandig naamwoord, zie de beschrijving in (7), en waarbij parameters zoals [sg] (*singular* ('enkelvoud')) en [pl] (*plural* ('meervoud')) worden gebruikt om de juiste vorm van het zelfstandig naamwoord binnen de expressie aan te geven.

(7)

MWE patroon	**beschrijving**
MWEp3	expressies bestaande uit een werkwoord dat een subject neemt en een direct object dat bestaat uit een determinator en een zelfstandig naamwoord

In dit proefschrift is het parameteriseren van de ECM verder onderzocht en zijn er in totaal 26 parameters gedefinieerd voor het Nederlands. Uit berekeningen is gebleken dat in de oorspronkelijke ECM

1.308 ECs nodig zijn en in de geparameteriseerde ECM slechts 140, een vermindering van bijna 90%. De geparameteriseerde versie van de ECM is geïmplementeerd in DuELME.

De finale versie van DuELME bestaat uit twee delen, namelijk (1) een lijst met MWE-patroonbeschrijvingen en (2) een lijst met MWE-beschrijvingen. De ECs worden gevormd aan de hand van de patroonbeschrijvingen. In de oorspronkelijke ECM bestond een patroonbeschrijving uit een ID en een tekstuele beschrijving. De voornaamste uitbreiding in de huidige patroonbeschrijvingen is de toevoeging van een formele representatie van de syntactische patronen. De formele representatie is gebaseerd op dependentiestructuren zoals gebruikt in het CGN (Corpus Gesproken Nederlands, zie Hoekstra et al. (2003)). Deze structuren zijn uitgebreid met aspecten die noodzakelijk zijn voor correcte beschrijving van MWEs. Veel recente projecten maken gebruik van dit formaat en kan daarom worden gezien als een de facto standaard voor het Nederlands.

Het belangrijkste doel van het toevoegen van een formele representatie is het verder reduceren van het manuele werk met name voor systemen die gebruik kunnen maken van dependentiestructuren. Omdat het een uitbreiding betreft, ondervinden systemen die geen gebruik maken van dependentiestructuren geen nadeel van deze toevoeging.

Naast de patroonbeschrijvingen is er een standaard ontwikkeld voor MWE-beschrijvingen, die onderverdeeld is in een basis MWE-beschrijving en een additionele MWE-beschrijving. Het ontwerp en de implementatie van de standaard lexicale representatie is uitgebreid beschreven in hoofdstuk 4 van dit proefschrift.

De data voor DuELME zijn automatisch geëxtraheerd uit corpora. Uit een lijst met ruim 9.000 kandidaat-MWEs, dat wil zeggen woordcombinaties die potentieel een MWE of een deel van een MWE vormen, is handmatig bepaald welke combinaties echte MWEs zijn die moeten worden opgenomen in DuELME. De dataextractie en -selectie is beschreven in hoofdstuk 3.

Deel I eindigt met een discussie van de gevolgde methode en een beschrijving van een semi-automatische conversie van de standaardrepresentatie naar de representatie zoals gebruikt in Alpino, een dependentieparser voor het Nederlands. Uit een kleine test met Alpino is gebleken dat de parser daadwerkelijk beter presteert bij het parsen van zinnen met MWEs wanneer het Alpino-lexicon is uitgebreid met

de MWE-beschrijvingen uit DuELME.

Wat opvalt aan de geanalyseerde data is dat de meeste MWEs over het algemeen zeer frequent in één bepaalde vorm voorkomen, bijvoorbeeld de meest voorkomende vorm van de MWE *de boot missen* is *de boot missen*, dus met het lidwoord *de* en zonder modificatie. Hoewel dit de meest frequente vorm is, is niet geanalyseerd of dit ook daadwerkelijk de enige mogelijke vorm is.

Deel II van deze dissertatie onderzoekt de variatiemogelijkheden van MWEs in detail. Het onderzoek concentreert zich in het bijzonder op OBJ1-V idiomen, dat wil zeggen combinaties van een werkwoord en een direct object waarvan tenminste het zelfstandig naamwoord, het hoofd van het direct object, niet letterlijk gebruikt wordt. Voorbeelden van idiomen zijn *de boot missen*, *de benen nemen* en *de noodklok luiden*.

De theorie zoals gepresenteerd in dit onderzoek draait om het begrip *idiomatisch referent*: een idioomdeel heeft een idiomatisch referent als het refereert naar een element in de denotatie (of betekenis) van het idioom.

In hoofdstuk 6 voorspel ik dat het variatiepotentieel van een idioom afhangt van de aanwezigheid van een idiomatisch referent voor de idioomdelen:

1. Als een idioomdeel een idiomatisch referent heeft,
   dan is het variatiepotentieel van dit deel in principe ongelimiteerd.

2. Als een idioomdeel geen idiomatisch referent heeft,
   dan kan er geen variatie optreden waarvoor de aanwezigheid van een idiomatisch referent verplicht is.

Uit bovenstaande voorspelling valt af te leiden, dat er onderscheid wordt gemaakt tussen variatie waarvoor een idiomatisch referent aanwezig moet zijn en variatie waarvoor dit niet het geval hoeft te zijn. Zoals beargumenteerd in onder andere Odijk (1993) and Schenk (1994) kan bepaalde variatie alleen optreden bij elementen die betekenis hebben. Vertaald naar idiomen, betekent dit dat er een onderscheid gemaakt kan worden tussen variatie die alleen kan optreden als er een idiomatisch referent aanwezig is voor de idioomdelen en variatie waarvoor dit niet het geval hoeft te zijn. In dit onderzoek worden zeven typen

variatie systematisch onderzocht: variatie in de determinator, variatie in het getal van het zelfstandig naamwoord, diminutief gebruik van het zelfstandig naamwoord, topicalisatie, pronominalisatie, modificatie en passivizatie.

Hoofdstuk 7 test de voorspelling voor 25 OBJ1-V idiomen aan de hand van corpus data. Hoewel het gebruik van corpus data essentieel is en grote voordelen biedt, blijkt dat er in corpus data relatief weinig variatie voorkomt, zowel in verschillende typen als in aantallen per type. Daarom is besloten om naast corpus data beperkt gebruik te maken van geconstrueerde data. Een klein aantal voorbeelden is voorgelegd aan een panel van taalkundigen (met Nederlands als moedertaal).

Hoofdstuk 8 presenteert de belangrijkste conclusies van Deel II en beschrijft de consequenties van dit onderzoek voor de standaardrepresentatie van MWEs in DuELME. De hoofdconclusie is dat de data van 20 van de 25 geanalyseerde idiomen de voorspelling ondersteunen. Voor de overige vijf idiomen is meer onderzoek nodig om beter inzicht te krijgen in het variatiepotentieel van deze groep.

In hoofdstuk 9 tenslotte, worden beide delen kort samengevat en worden er een aantal suggesties gedaan voor verder onderzoek.

Curriculum Vitae

Nicole Grégoire was born in Delft on the 5th of March, 1979. After obtaining her Atheneum degree at *Het Westlandcollege*, she started an internship at Exact Software in Delft. She worked as a computer programmer for almost two years. In 1999 she started with Dutch Language and Culture at Utrecht University. In her third year she studied Linguistics at the University of Edinburgh, Scotland. In April 2004, Nicole obtained her master's degree (cum laude) in the direction of computational linguistics with a thesis entitled *Accentuation of Adpositions and Particles - Towards a set of rules for predicting accent locations on adpositions and particles for Dutch text-to-speech technology*. Nicole started her PhD project at the Utrecht Institute of Linguistics in 2005. The research journey is reflected in this book.