Data-driven identification of fixed expressions and their modifiability
Villada Moirón, María Begoña

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2005

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 1

Introduction

Data-driven identification of fixed expressions and their modifiability explores methodological and empirical issues that the computational linguist encounters when attempting to extend a computational grammar with fixed expressions. This thesis investigates two properties of fixed expressions: automatic identification and the establishment of their potential for internal modification (and other sorts of variation). These two issues are pertinent to building and updating computational lexica and also to augmenting grammars in the process of expanding the coverage of a computational parser. We have in mind wide-coverage parsers; thus, we investigate which data-driven methods are useful and how adequate they are in tackling the issues just mentioned. The experimental work applies to Dutch fixed expressions, in particular collocational prepositional phrases and support verb constructions.

1.1 Background

1.1.1 Chapter overview

This section (1.1) gives the theoretical and practical motivation for exploring automatic methods to expand a lexicon of fixed expressions; it also motivates the need to determine the potential for variation and modification of fixed expressions before a formalization of these expressions in a lexicalist grammar is attempted. Our formulation of the object of this study, specific aims and objectives are given in section 1.2. Section 1.3 introduces the theoretical framework that informed the design, development and evaluation of the data-driven models. Since the purpose of exploring and developing data-driven models is to build lexica of fixed expressions, we first review work on theoretical and descriptive linguistics trying to find out which expressions qualify as


Chapter 1. Introduction

fixed expressions and the characteristic features of these expressions. Next, an overview of data-driven models used in the past to acquire statistically significant word combinations follows. Section 1.4 describes the core part of the thesis. The section unfolds the approach we pursued to attempt to reach our objectives. Some applications of this research are described in section 1.5. Section 1.6 provides an overview summarizing the contents of all chapters.

1.1.2 What are fixed expressions?

The term fixed expression has been used to refer to a variety of linguistic expressions; among these are idioms, collocations, metaphors, support verb constructions, phrasal verbs, institutionalized phrases, sayings, proverbs and formulaic expressions. Examples of fixed expressions are the English expressions in (1):

(1) a. *spill the beans, shoot the breeze* (idiom)
b. *the eye of a needle, the evening falls* (metaphor)
c. *dogs bark, sound asleep* (collocation)
d. *make progress, take a bath* (support verb construction)
e. *Don’t count your chicks before they hatch, Curiosity killed the cat* (proverb, saying)
f. *The X-er, the Y-er* e.g. *The more, the merrier* (construction)  
g. *to cost an arm and a leg, listen like a police officer* (simile) 
h. *at school, by and large* (institutionalized phrase or set phrase)

Some of these expressions are by no means fixed sequences of words, thus researchers adopted other terms like multi-word units, multi-word lexemes, multi-word expressions (Sag et al., 2001), phrasal lexical items (Everaert and Kuiper, 1996) or phrasal lexical entries (Sailer, 2000).

A satisfactory definition of fixed expression is difficult due to the varied and idiosyncratic nature of the data. Aiming at a definition of fixed expression to be adopted by lexicographers, Everaert (1993) proposed the following:

'A combination of two or more words that must at least satisfy the (a) condition and perhaps, but not necessarily, condition (b) and/or (c):

(a) the word combination is fixed;
(b) the combination as a whole has a non-compositional or partially compositional meaning;
(c) the syntactic/morphological behavior of the fixed expression
1.1. Background

and/or its parts is not to be expected given the syntactic/morphologic behavior of the individual words or the combination as a whole (Everaert, 1993, pg. 18).\footnote{Throughout the thesis, each Dutch fixed expression used in the text is followed by its (approximate) translation in English. Appendix B provides a list of all these expressions with their word-to-word glosses and translations.}

Everaert’s definition emphasizes one necessary feature, that is, the lexical fixedness between the words inside the expression, allowing potential idiosyncrasies in the morpho-syntax and/or semantics. We interpret lexical fixedness as the requirement that (at least) two lexemes co-occur bearing a syntagmatic relationship. We use the term fixed expression in the sense portrayed in the above definition.

Fixed expressions constitute a non-negligible portion of the language in use, perhaps because of the various types of phenomena covered by the term. For illustration, the non-exhaustive Van Dale Idioom Woordenboek (Van Dale Dictionary of Idioms) (1999) consists of ca. 10,000 Dutch idioms and other fixed expressions. In addition, frequency figures reveal that many fixed expressions are too frequent to be left untreated in a computational grammar. Table 1.1 gives a few examples of Dutch support verb constructions found in the 300 million word Twente Nieuws Corpus (tWNC) (Ordelman, 2002). Next, we motivate why fixed expressions deserve special attention and the reasons why they require a different treatment in a computational grammar.\footnote{Throughout the thesis, each Dutch fixed expression used in the text is followed by its (approximate) translation in English. Appendix B provides a list of all these expressions with their word-to-word glosses and translations.}

<table>
<thead>
<tr>
<th>expression</th>
<th>translation</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>iets in de gaten houden</td>
<td>keep an eye on something</td>
<td>4,328</td>
</tr>
<tr>
<td>iemand voor de gek houden</td>
<td>to fool someone</td>
<td>421</td>
</tr>
<tr>
<td>iemand op de hoogte stellen</td>
<td>to inform someone</td>
<td>1,021</td>
</tr>
<tr>
<td>aan de slag gaan</td>
<td>get started</td>
<td>3,697</td>
</tr>
<tr>
<td>voet bij stuk houden</td>
<td>stick to one’s decision</td>
<td>380</td>
</tr>
<tr>
<td>om de tafel gaan zitten</td>
<td>to begin negotiations</td>
<td>403</td>
</tr>
</tbody>
</table>

Table 1.1: Some Dutch support verb constructions in the 300 million word Twente Nieuws Corpus and their frequency.

1.1.3 Regularity, semi-regularity and exceptions in the grammar and lexicon

Generally speaking, fixed expressions may exhibit idiosyncrasies in their morphosyntactic and/or semantic behavior. Consequently, it is not easy to
predict which lexemes within the expression undergo productive morphology, whether the expression undergoes passivization, topicalization, insertion of modification, ‘it-cleft’, extraction, etc. The former observation supports the argument that fixed expressions cannot always be built by regular grammar rules (Sailer, 2000; Riehemann, 2001). To illustrate this, consider an example of the fixed expression *iets in petto hebben* ‘to have something in store’:

(2) Ze hebben weer een nieuwe machinatie in petto
    they have again a new plot in store
    ‘Once again, they have a surprising new plot in store.’

Without further specifications in the lexicon, a robust computational parser would, probably, return the syntactic representation in (3) assuming an underlying regular structure:

(3) 

```
S               
/   \           
VP              
/     \         
NP              
/     \         
Ze
hebben         AdvP
weer           NP
een nieuwe machinatie
PP
\     /          
in petto
```

The verb *hebben* typically shows a transitive subcategorization frame \([\text{NP}_{\text{nom}}\ \text{NP}_{\text{acc}}]\), therefore, the parser considers *Ze* the grammatical subject agreeing with the tensed verb, *een nieuwe machinatie* the direct object and *in* the head of a PP functioning as an event modifier. In this context, the lexeme *hebben* is part of the lexicalized expression *iets in petto hebben*. To distinguish this peculiar phrase from the individual lexeme *hebben*, extended subcategorization requirements and lexical restrictions such as \([\text{NP}_{\text{nom}}\ \text{NP}_{\text{acc}}\ \text{PP}_{\text{in petto}}]\) are specified. However, these restrictions encounter a problem.

Parsing could fail due to missing lexical entries. With the exception of the expressions *in petto hebben* and *in petto houden* ‘to keep in reserve’, the Italian loanword *petto* ‘pocket, sack; (fig) heart’ does not occur in other expressions in the language. *Petto* needs to be included in the lexicon. It is unclear what word category should be assigned to *petto*. Since *petto* seems to be the complement of the preposition *in*, let us assume that *petto* is a noun. This ad hoc decision would make the wrong predictions about the behavior of *petto* and the phrase *in petto* in syntax. The insertion of a nominal lexeme *petto* would wrongly predict that the word is free to appear in slots where a noun is expected. *Petto* shows no modification, no plural morphemes nor determiners:
1.1. Background

Existing formalizations of fixed expressions propose instead to treat complement-like phrases selected by the verbal head as a fixed unit (Krenn and Erbach, 1994). To capture this, some language engineering proposals enter the fixed complement as a multi-word lexeme selected by the verb in the lexicon (Breidt et al., 1996). For the sake of the argument, we adopt this solution and label in petto as a special fixed phrase selected by the verb hebben. The updated subcategorization requirements of this use of the verb hebben are [NP_nom NP_acc fixed phrase_in petto]. The resulting syntactic analysis of (2) is given in (5):

The addition of a lexical category fixed phrase has also consequences in the grammar. Additional grammar rules that only apply to fixed phrase units are needed. Such grammar rules license a fixed expression provided a fixed phrase chunk combines with a selecting verb and all the verb’s subcategorization requirements are met. At the same time, the component lexemes within fixed phrases do not participate in regular morphological and syntactic operations.

Fixed arguments show strong preferences for a particular location in certain clause types. In petto may occur at the end of a matrix clause (6-a). In subordinate clauses (in Dutch, verb final is required), in petto occurs in preverbal position (6-b):

(6) a. De carrière van Faneyte had nog veel moois in petto.  
the career of Faneyte had yet much nice aspects in store  
‘Faneyte’s career still promised many future highlights.’

b. (...) omdat niemand weet wat voor maatregelen Mahathir  
... because nobody knows what for measures Mahathir  
nog in petto heeft  
yet in store has
Chapter 1. Introduction

‘(...) because nobody still knows what sort of measures Mahathir has in store.’

Whereas complement PPs may allow extraposition in Dutch (7), no instances of an extraposed in petto were found in the TwNC corpus. In petto behaves differently from regular complements in that it appears before the verbal cluster in subordinate clauses, allowing only auxiliaries or modals before its selecting verb. To control this, additional rules need to state where a fixed phrase may occur within the clause. Irregularity thus has to be introduced in the lexicon and in the grammar itself.

(7) Vanaf 1 januari moet de luchthaven zich houden aan de regels.
    From 1 January must the airport itself hold on to the rules
    ‘From January 1st, the airport must adhere to the rules.’

Whereas the fixed phrase approach may be accurate for totally fixed arguments, other arguments that allow internal variation among specifiers or modifiers cannot be handled this way. In the expression in je maag zitten met ‘be troubled with’, the noun maag allows a definite (de) (8-a) or a possessive determiner (zijn, haar) that shares agreement features with the subject (8-b):

(8) a. ...zit duidelijk met de kwestie in de maag
    ...sits clearly with the issue in the stomach
    ‘(someone) is clearly troubled by the issue.’

b. Van Nieuwenhuyze, zit lelijk in haar, maag met de affaire.
    Van Nieuwenhuyze sits ugly in her stomach with the affair
    ‘Van Nieuwenhuyze is worried about the affair.’

In these expressions, arguments require internal syntactic structure; at the same time, lexical affinities between the verb (zitten), the preposition (in) and the head noun (maag) need to be specified. Provided the lexical restrictions are ensured, the structure in (9) could be suitable.

---

2The same domain preferences apply to other PP and NP arguments in many Dutch lexicalized expressions (p.c. Jack Hoeksema).
3Alternatively, the phrase in petto can be treated as a PP that declares a feature [fixed]. The feature [fixed] ensures that extraposition and extraction do not regularly apply to this lexicalized PP.
1.1. Background

A last piece of evidence suggests the need of further constraints. In the expression *het niet in je kouwe kleren gaan zitten* ‘sth. leaves its mark on you’, the specifier may be realized by a definite or a possessive determiner and, the noun *kleren* has to be modified by the adjective *kouwe* or *koude* ‘cold’ (within the PP). This shows that in addition to categorical constraints, fine-grained conditions on the realization of the internal modification may be needed.

Many fixed expressions also exhibit irregularities in their semantics. A fixed expression consists of two or more lexemes that may denote a different meaning from the meaning the individual lexeme has when it occurs on its own. In Dutch, *poetsen* ‘to polish’ no longer maintains the same meaning in the context of the expression *de plaat poetsen* (lit. the plate polish) ‘to depart unnoticed’. The figurative meaning is only present if all three lexemes *de plaat poetsen* co-occur. Due to the lexical affinities and the peculiar meaning of the combination, the expression *de plaat poetsen* is conceived as a lexical item and not as a phrase built up from independent lexemes.

If one tries to replace a lexeme by a (near-)synonym or insert an adjective modifier that is semantically plausible in a fixed expression, the semantic irregularity becomes apparent. *De plaat poetsen* neither allows synonym replacement (10-a) nor insertion of adjective modifiers (10-b):

(10) a. Bell *poestete/*maakte *de plaat* *schoon en La PrimaveraBell polished/*made the plate *clean and La Primavera
werd een hotel.
became a hotel
‘Bell left unnoticed and La Primavera became a hotel.’

b. Bell *poestete* de (*onopgemerkte) *plaat* en La PrimaveraBell cleaned the unnoticed plate and La Primavera
werd een hotel.
became a hotel
‘Suddenly, Bell left unnoticed and La Primavera became a hotel.’
However, the degree of irregularity varies across the expressions. Whereas synonym replacement is often impossible, insertion of modifiers is sometimes feasible. The \textit{NP} argument headed by \textit{gevolg} ‘consequence’ allows modification in \textit{gevolg geven aan} ‘to carry out’:

\begin{equation}
\text{Daarmee zou men een logisch gevolg geven aan de onzinnige bezuinigingen ... }
\end{equation}

\begin{equation}
\text{absurd cutbacks ...}
\end{equation}

‘With that, men will logically carry out the absurd cuts ...’

Although the idiosyncratic semantics of fixed expressions might not directly pose a problem for a parser, it may play a role in accounting for their \textit{a priori} unpredictable syntactic versatility (Nunberg et al., 1994); in particular, the underlying semantics might explain why some fixed expressions allow (limited) modification (11) and others do not (10-b).

The lexical status of component words, lexical and morphological restrictions, subcategorization requirements, location of fixed arguments in certain clauses, variation affecting specifiers and adjectival modification vary from expression to expression. In order to capture the idiosyncratic behavior, fixed expressions are treated as phrasal lexical entries i.e. described in the lexicon (Everaert and Kuiper, 1996; Sailer, 2000; Riehemann, 2001).

\section{1.2 Research questions}

A lexicalist constraint-based grammar describes well-formed phrases in a given language capturing regular and productive syntactic operations. Such phrases are made up of individual lexemes that may have undergone morphological operations by the time they enter syntax. The evidence discussed in section 1.1.3 supports the claim that fixed expressions cannot always be formalized with regular grammar rules (Sailer, 2000; Riehemann, 2001).

Fixed expressions belong to the lexicon-grammar interface. With regard to their semantics, meaning is assigned to the expression as a unit; this can be captured by treating the whole expression as a phrasal lexical item. Limited variation shows that word formation rules including morphological rules, do not regularly apply to the individual words; their restricted syntactic flexibility (phrase internal and sentential) hints at a certain frozen nature. However, some expressions seem to be partially lexicalized because they exhibit alternation of specifiers or insertion of modifiers.

Before one decides how to represent fixed expressions in the lexicon and
1.2. Research questions

how to encode them in the grammar, one first needs to find out which expressions qualify as fixed expressions and secondly, to characterize the internal structure and morphosyntactic restrictions shown by the lexemes or component phrases. Some of the problems we face are: how to identify combinations that behave as fixed expressions, which features or tests tell us that an expression requires special encoding in a grammar, how to establish which are the minimum required lexemes, what slots are fixed and which parts are variable within a fixed expression. Our aim is to develop methods for expanding the lexicon with fixed expressions. We also want to gather evidence of morphosyntactic variation and modification which is relevant for the linguistic description of these expressions in a lexicalist constraint-based grammar.\(^4\)

The research questions we will investigate here are the following:

- **Can data-driven methods identify which fixed expressions require special lexical mention?**
- **To what extent can we determine the potential for modification of fixed expressions using a corpus-based approach?**

To find an answer, this work focuses on two types of Dutch fixed expressions, namely *collocational prepositional phrases* and *support verb constructions*. However, we are interested in the general applicability of this research.

1.2.1 Aims and objectives

The main goals are the following. First, we investigate and develop a data-driven model to automatically extract fixed expressions. We concentrate on Dutch *collocational prepositional phrases* and *support verb constructions*. Secondly, we investigate a corpus-based method to infer the variation and modification potential of *support verb constructions*. As a derivative issue, we establish whether variation and modification potential pinpoint underlying semantic structure in *support verb constructions*.

The main objectives are a hybrid model to perform automatic identification of fixed expressions in corpora and evaluation of its performance. We also aim at a corpus-based method to extract evidence of variation and modification within *support verb constructions* with as little human intervention as possible. Evaluation of the corpus-based method provides positive input for future investigations. As by-products, we expect a characterization of two

\(^4\)The parser and the grammar being developed are described below in section 3.2.2.
Chapter 1. Introduction

types of fixed expressions, namely, *collocational prepositional phrases* and *support verb constructions* and a report on the variation and modification potential within *support verb constructions*.

Both, the characterization of two types of fixed expressions and the development of the data-driven models that identify such expressions in large corpora can contribute valuable insights towards a better understanding of these idiosyncratic expressions. Data-driven models can retrieve thousands of examples of the linguistic behavior of fixed expressions from open domain corpora. Knowledge concerning the linguistic behavior contributes the necessary empirical evidence to develop a theory of fixed expressions. Furthermore, existing grammatical theories have to face the challenge of describing and accounting for the idiosyncratic nature of fixed expressions.

1.2.2 Practical motivation

The Pionier program *Algorithms for Linguistic Processing* aims at developing a wide-coverage lexicalist grammar and parser for Dutch.\(^5\) Wide-coverage parsing ought to provide a syntactic analysis of any sentence in a language. Hence, the grammar needs to include appropriate linguistic descriptions of any syntactic construction considered to be well-formed. Furthermore, the lexicon (dictionary) should be updated and expanded as the vocabulary growth requires it. The ultimate goal of this work is to understand the grammar of Dutch fixed expressions in particular their potential for modification thus, contributing to an efficient processing of these expressions by a Dutch parser.

1.3 Linguistics, Corpora and Statistics

Past work on theoretical and descriptive linguistics reports detailed characterizations of idioms and other types of fixed expressions. A set of features that distinguish fixed expressions from regular expressions does not exist; idiosyncrasies vary across expressions. Large corpora provide a potentially rich source of fixed expressions and an invaluable collection of their linguistic behavior. Former corpus-based approaches are described to find out if they are suitable to pursue our goals.

\(^5\)A description of the research program is available at http://www.let.rug.nl/~vannoord/alp.
1.3.1 Previous linguistic descriptions

The theoretical and descriptive linguistics literature suggests that more and more types of linguistic expressions fall under the phenomenon of fixed expressions. Earlier work focuses on the archetype (Moon, 1998) of fixed expressions, that is, idioms. Recent corpus-based research investigates expressions such as English verb-particle combinations (Baldwin and Villavicencio, 2002), determinerless PPs (in English and Dutch) (Baldwin et al., to appear), varied expressions such as adjective-noun-verb combinations (Zinsmeister and Heid, 2003), cranberry words (Richter and Sailer, 2003) and complex prepositions (Trawinski, 2003) in German as well as Dutch adjective-verb, noun-verb, adverb-verb and preposition-noun-verb combinations (Hoeksema, 2004).

Let us review what features characterize the prototype of a fixed expression. Idioms do not exhibit a specific form or morphosyntactic structure (Fernando and Flavell, 1981; Everaert and Kuiper, 1996). Idioms have been described as showing defective morpho-syntax and irregular semantics. One line of research maintains that idioms are expressions with non-compositional meaning, that is, the meaning of the whole idiom cannot be regularly derived from the meaning of its parts (Fernando and Flavell, 1981; Schenk, 1994; Abeillé, 1995; Everaert and Kuiper, 1996). This view also maintains that internal adjectival modification if present, modifies the meaning of the whole idiom. Other instances of adjectival modification are ‘word play’ and considered as marginal. An alternative line of research, the one we follow, argued for the existence of both non-decomposable (non-compositional) and decomposable idioms (Nunberg et al., 1994; Moon, 1998). This view maintains that some (decomposable) idioms have internal semantic structure which is reflected in their syntactic flexibility (Nunberg et al., 1994). Furthermore, such idioms allow internal modification (Nicolas, 1995) that may affect only the meaning of an idiom’s constituent. Corpus-based studies corroborated Nunberg et al.’s (1994) views, not only with regards to the existence of two types of idioms differing in their semantic structure but also with regards to their internal flexibility (Moon, 1998; Riehemann, 2001; Sag et al., 2001). Riehemann’s findings further support Nunberg et al.’s claims (1994) that syntactic flexibility (modification, passivization, topicalization, etc.) is more likely to be found among semantically decomposable idioms.

---

6Below two different views on idioms are described. This work uses a more inclusive concept along the lines of Nunberg et al. (1994) and Moon (1998). Section 2.2.3 briefly discusses the use of the term idiom in the literature.

7Recently, international workshops have hosted numerous presentations on various aspects of collocations and fixed expressions cross-linguistically. To cite a few: the series of ACL workshops on multiword expressions held in 2003 and 2004, sessions on collocation and multi-word expression extraction at LREC’04, EURALEX’02 and ’04, etc.
Due to the difficulty of gathering a set of sufficient features that identify fixed expressions, a set of linguistic diagnostics that show distinctions between regular phrases and idiosyncratic ones can be helpful. Fernando and Flavell (1981) and Sailer (2000) propose some diagnostics to detect syntactic and semantic irregularities in English verb NP idiomatic expressions. Among others, the diagnostics comprise synonym replacement, insertion of adjectival modification, passivization, pronominalization and topicalization. Diagnostics that are specific to Dutch syntax and to the two types of expressions studied in this work (Dutch collocational prepositional phrases and support verb constructions) have been proposed by Paardekooper (1962, 1973) and Holl-ebrandse (1993). Such diagnostics are described in chapters 4 and 6.

To our knowledge, recent studies have not specifically investigated internal morphosyntactic variation and modification, or at least not systematically. Exceptions are Moon’s (1998) and Riehemann’s (2001) corpus-based investigations. These two authors explored the variation in a pre-compiled list of idioms, collocations and other fixed expressions found in English corpora. Riehemann (2001) extracted evidence that show changes involving specifiers (e.g. quantifiers), adjectival modifiers, number morpheme, passivization, extraction, relative clauses and of-prepositional phrases, among others. Before Riehemann’s study, researchers often accepted the idea that instances of variation and internal modification in idioms were rare. Riehemann’s study revealed that 25% of the occurrences of decomposable idioms and 3% of non-decomposable V NP randomly selected idioms exhibit some sort of variation (Riehemann, 2001, p. 131). These findings suggest that morphosyntactic variation is a non-negligible feature.

1.3.2 Corpus-based approaches

Corpus linguistics is a branch of descriptive linguistics that aims at extracting generalizations about language patterns by automatically exploring large collections of texts (corpora). This empiricist approach to linguistic analysis and exploration rests on the assumption that the knowledge speakers have of their language can be inferred from the linguistic evidence found in corpora. Large corpora have been found to be a valuable resource that can provide the descriptive linguist with all sorts of linguistic information: lexical word co-occurrence information, word frequency, distribution of linguistic structures and phenomena, etc.

A branch of corpus-based research aims at automatically identifying collocations in corpora. Most studies adopt the following definition of collocation proposed by Firth (1957):
`statements of the habitual or customary places of that word (...). The collocation of a word or a ‘piece’ is not to be regarded as mere juxtaposition, it is an order of mutual expectancy (Firth, 1957, p. 181).’

Firth’s view of collocation captures the fact that the component lexemes mutually expect each other. Although this criterion applies to interesting collocations s.a. dogs bark and sound asleep, it also applies to combinations such as eat apple, in the and doctor nurse which happen to cooccur in the same context very frequently. We use Firth’s statistical definition of collocation as an empirical notion to identify fixed expressions. Firth’s concept of collocation partially overlaps with fixed expression, but collocation refers to a broader phenomenon (cf. van der Wouden (1994); Sag et al. (2001)).

Early collocation extraction models were built on the assumption that the lexical affinities between the words in a collocation can be approximated by a strong statistical dependence between the words. These purely statistical models collect all possible patterns (word sequences) of various lengths from a textual corpus. Descriptive and inferential statistics measure the degree of dependence between the component lexemes. The simplest method sorts candidates according to their raw frequency. Other common statistical measures are: mutual information and log-likelihood. The models expect that word combinations showing a strong statistical dependence are good collocations (Damerau, 1993). If the collocations consist of words that occur non-adjacently, a purely statistical approach allows too much noise (Dias et al., 1999). Another drawback is that if one aims at collocations that involve a function word (e.g. prepositions) highly frequent combinations of a content word and a preposition or a determiner will show a strong statistical dependence, even though the combination is fully regular (e.g. beginning of, sing in, in the). Finally, manual evaluation of the retrieved lists is required due to the mixed types of the retrieved collocation candidates.

Later work on identification of collocations assumes a more restricted notion of collocation. This notion can be found, for instance, in Fontenelle’s work:

‘The term collocation refers to the idiosyncratic syntagmatic com-

---

8To the best of our knowledge, existing grammatical theories have not yet introduced a linguistic entity collocation. Recent HPSG-grammar proposals have postulated a ‘collocational module’. This module is an analytical device that allows a lexeme to declare the distributional context within which its collocates may occur. Such device has been shown to license idiomatic expressions (Sailer, 2000), certain collocations and negative polarity items (Sailer and Richter, 2002; Richter and Sailer, 2003).
bination of lexical items and is independent of word class or syntactic structure (Fontenelle, 1992, p. 222).”

This view of collocation introduces some fundamental features: (1) arbitrariness in the combination and (2) a lexico-syntactic relationship between the component lexemes. In addition, this view sets no restrictions on the component words nor on the form of the combination. Let us now explain how more recent approaches model this notion of collocation.

With improvements in part-of-speech tagging, phrase chunking and full parsing, data which is automatically annotated with linguistic information has become available; these developments in natural language processing have enabled identification models that extract only collocation candidates that satisfy certain morphosyntactic requirements (verb–object, adj–noun, verb–pp, etc.). Known as hybrid models, they make use of linguistic information and statistics and they are the preferred models nowadays (Pearce, 2002; Kermes and Heid, 2003; Spranger, 2004). Decisions concerning the annotation of the extraction data, extraction of candidate datasets, choice of statistical measure and evaluation metrics, all contribute to differing views on the success of these models. Richly annotated data is preferred to part-of-speech tagged data, since it facilitates collecting candidates whose constituents are non-adjacent and bear a syntagmatic relation. Since candidates show similar morphosyntactic structure, evaluation can be fully automatic provided one has access to machine readable dictionaries or a lexical database.

Two problems faced by all models are: the identification of collocations among low-frequency data and the size of the candidate expressions. Unless one has a large amount of data, statistical measures will produce unreliable scores for low-frequency data. A measure that produces reasonable scores even among lower-frequency data is log-likelihood. Most collocation statistics can only be applied on bigrams (two word candidates), thus they pose an important restriction. When candidate expressions have a size of $n \geq 3$, statistical measures are not easy to adjust. More complex models, among them a loglinear model implemented by Blaheta and Johnson (2001), by-pass the size problem, allowing candidates of any length.

In order to compile a computational lexicon of fixed expressions, one could collect the data from existing monolingual (or bilingual) dictionaries. We decided against this approach for several reasons. Existing dictionaries and glossaries are static resources, often covering a limited proportion of the existing data. Due to the lack of a sound definition of fixed expression, these expressions are not consistently annotated in dictionaries (Everaert, 1993). Further, compiling a list from existing dictionaries is not easy because some fixed expressions are listed as special phrases of a word entry and other
1.4. This thesis

expressions are cited in examples. Finally, morphosyntactic constraints are not yet available in most dictionaries. Therefore, we decided to investigate data-driven methods for automatically identifying fixed expressions.

Besides providing the raw data for collocation identification models, corpus evidence shows preferences and restrictions, and it shows that ‘variation (within fixed expressions) is much more common than some models suggest (Moon, 1998)’. Concordances and UNIX tools (grep, sort, etc.) have been proven useful to collect evidence of how much variation and morphosyntactic flexibility is found among a pre-compiled list of fixed expressions (Moon, 1998; Riehemann, 2001). Although the results are satisfactory, these methods require intensive manual work. We aim at a corpus-based method that extracts relevant evidence with as little manual inspection as possible.

1.4 This thesis

The accuracy of the representation and linguistic description of fixed expressions in a lexicalist constraint-based grammar is crucial for improving parsing coverage and accuracy. The work described in this thesis aims at facilitating the expansion of a lexicalist constraint-based grammar for Dutch. Fixed expressions exhibit idiosyncrasies at different linguistic levels; and these idiosyncrasies are not always found in existing Dutch dictionaries. Therefore, we explore their behavior in large corpora and from such evidence infer the required morphosyntactic constraints. First, we attempt to build lexica of fixed expressions found in large corpora and second, we seek to extract evidence of variation and modification in such expressions from a corpus. On the basis of the extracted evidence, a grammar writer decides the lexical representation of a fixed expression, especially the necessary lexical and morphosyntactic constraints. In addition to investigating data-driven methods for acquiring a lexicon of fixed expressions, we also seek to establish the variation and modification potential of required arguments within fixed expressions.

1.4.1 Characteristics of fixed expressions

A general characterization of the idiosyncratic lexical, morphological, syntactic, and semantic features of fixed expressions is necessary for two reasons: (1) the need to identify which properties of fixed expressions introduce requirements to be captured by the identification models and (2) the importance of assessing the quality of the output of the identification models.

Among the properties involved, there are lexical (e.g. nonce words, lexical restrictions), morphological (e.g. archaic forms, compounding and deriva-
Chapter 1. Introduction

Properties of fixed expressions show strong lexical affinities among them; these lexemes may be adjacent to each other or they may allow intervening elements. Regular morphological and syntactic rules may apply to some constituents in fixed expressions. Consequently, the degree of lexicalization varies, showing a continuum from expressions that are totally fixed (ad hoc) to rather flexible expressions that allow adjectival modification (spill the royal beans).

General properties of fixed expressions and more specific features of Dutch collocational prepositional phrases and support verb constructions impose certain constraints on identification models. Fixed expressions that consist of unknown words, archaic words and non-productive morphological variants (case marking remnants), ill-formed structures or verbal lexemes with unexpected syntactic constituency, could be ignored by hybrid identification models. We explain how this can happen.

Identification models extract candidate expressions from automatically annotated data. If the knowledge resources (lexicon and grammar) lack a description of fixed expressions, the annotation tools (parser) may, for instance, fail to label a dependency between a verb and one of its dependents containing an unknown word. The consequence is that actual fixed expressions could be missed during extraction of candidate datasets even though they occur in the extraction corpus. A robust parser, like the Alpino parser (section 3.2.2), that uses heuristics to guess the part-of-speech of unknown words partially solves the problem. In the task of identifying support verb constructions, our solution was to not rely fully on the dependencies proposed by the parser; instead we make use of lexical and phrasal chunks information during the extraction of candidate expressions. The more of these properties that are neglected by the annotation tools, the less linguistic information is available.

Identification models output a ranked list of candidate expressions. Next, the researcher assesses which expressions are valid fixed expressions. Alternatively, human judges (native speakers) are given guidelines and asked to identify those candidates in the extracted lists that satisfy one or more distinguishing features of fixed expressions. Besides factors that affect fixed expressions in general, specific characteristics of Dutch collocational prepositional phrases and support verb constructions are taken into account during the qualitative evaluation of the output. In this thesis, evaluation of the identification models involved assessment by human judges; our informants commented on the difficulty of making a decision sometimes; we attribute this difficulty to the lack of a good definition of the phenomena.
1.4.2 Identification of fixed expressions

Fixed expressions differ in their external form, internal structure and their behavior. Different linguistic behavior of the data we seek could make data-driven extraction methods unsuitable.\(^9\) We investigate data-driven methods with two different types of expressions to see the problems the methods encounter. We start with expressions that we consider rather fixed – *collocational prepositional phrases* – and move on to more complex expressions. Based on the limitations of the first model, we design other models and test them on the task of identifying *support verb constructions* in corpora. The identification task proceeds in various steps: pre-processing extraction data, building candidate datasets, ranking candidates with statistical scores and evaluation.

We pursue identification models that provide a list of fixed expressions in addition to relevant evidence for a linguistic description. During pre-processing, we used part-of-speech tagged data and automatically parsed data. Part-of-speech tagged data can be useful for models that aim at fixed expressions whose component lexemes are adjacent to each other. Syntactic patterns defined over part-of-speech tags can extract interesting expressions from a part-of-speech tagged corpus. Chapter 4 describes identification of Dutch collocational PPs in a part-of-speech tagged corpus. Fully parsed data offer multiple possibilities e.g. searching for patterns whose component lexemes bear a dependency relation and occur separately from each other. For example, if we want to extract collocations like *een vergadering sluiten* ‘to adjourn a meeting’, we collect all instances of the pattern ‘verb object NP’ found in the corpus. Using treebank query tools, syntactic patterns with varied external form and internal structure can be retrieved. Extraction of support verb constructions used fully parsed data. To avoid potential parser mistakes, we used phrasal chunks collected from a full parse during the extraction of candidate datasets.

Variation of the internal structure (due to e.g. the insertion of modification) has an effect on the length of the candidate expression. The length of the candidate can be a problem for applying certain statistical measures. Such measures are typically applied on bigrams (candidates that consist of two words). Target fixed expressions often involve three or more words (e.g. *in de gaten houden*). In our experiments, pre-nominal specifiers and adjectives are ignored, thus, a candidate expression is representend by a triple (verb, preposition, noun). Chapters 4 and 5 report how we applied the bigram statistics on triples and a loglinear model that sets no restrictions on

\(^9\) A long term goal is to design a model that identifies any type of fixed expression in corpora.
Some statistical tests are highly sensitive to low frequency data. A frequency cutoff is a common technique to avoid unreliable scores. Although many fixed expressions are very frequent in corpora, some others are not. A large number of fixed expressions show low frequency, which means that applying a frequency cutoff would discard them. We decided against using a cutoff. In the absence of a large gold standard list, we prefer to ignore the impact of a frequency cutoff. To avoid unreliable scores, two association measures that are less sensitive to low-frequent counts and therefore less dependent on a frequency cutoff were identified — log-likelihood and salience. The salience measure was the most successful in the support verb construction experiments. Given that our validation data was rather limited, we do not use standard precision and recall; instead, the various statistics are evaluated with accuracy graphs and a non-standard measure: uninterpolated average precision (also used in information retrieval). The advantage of this measure is that it indirectly measures the recall.

Among the candidates retrieved by the best statistical tests, noise surfaces among the top scores. Systematic errors made by the support verb construction identification model involve adjuncts and regular complements of verbs. Chapter 6 presents a theory-informed and data-driven identification method that exploits the syntactic idiosyncrasies of the expressions. Hollebrandse (1993) proposed linguistic diagnostics that distinguish Dutch support verb constructions from regular verb phrases. Tests such as extraposition, scrambling, nominalization patterns and coordination draw a distinction between regular verb phrases and support verb constructions. On the basis of these diagnostics, we designed an automatic method that discards a substantial percentage of the errors. The diagnostics are applied as a post-identification filter. To the best of our knowledge, this type of approach to error reduction has not been tried before.

1.4.3 Exploring variation and modifiability

Eventually fixed expressions need to be described in a lexicalist constraint-based grammar. Thus, it seemed appropriate to attempt a characterization of fixed expressions that studies observable linguistic knowledge, namely, lexical, morphological and syntactic features. Our second goal is to establish the modification potential in support verb constructions (SVCs). We aim at extracting relevant morphosyntactic features from corpora and based on these features, we predict the potential of modification of each expression.

Chapter 7 describes the results of some experiments on establishing modification potential. Alternation of specifiers, morphological productivity and
adjectival modification in support verb constructions show a continuum; at one end, we find totally fixed expressions and at the other end, expressions that allow specifier variation, limited morphological productivity and adjectival modification. Nevertheless, these findings need to be interpreted carefully; the fact that no evidence of variation or modification has been found in our corpus does not mean that none is possible. Inspired by Fellbaum (1993), we study the correlation between determiner changes and adjectival modification in the evidence extracted from the corpus. In a sample of more than 100 SVCs, we established that if noun phrases within required arguments in SVCs show alternation between either zero, definite and possessive determiner or, either indefinite or demonstrative or a quantifier and one or more other determiner types, then the noun phrase also shows adjectival modification. Therefore, we expect that arguments within SVCs for which we found such determiner alternations are likely to allow adjectival modification. The underlying intuition is that fixed expressions that allow variation and internal modification are likely to be semantically decomposable also. This introduces an often controversial property of fixed expressions: whether fixed expressions have semantic structure, that is, whether the meaning of the whole expression may be split among its constituents. Nunberg et al. (1994) argued that the fact that certain English idioms are semantically decomposable shows up in their syntactic flexibility. Internal modification is a manifestation of syntactic flexibility. We take Nunberg et al’s observation and assess its validity concerning Dutch support verb constructions. Chapter 7 reports that no clear-cut correlation exists between morphosyntactic flexibility and internal semantic structure in support verb constructions.

1.5 Applications

Lexicon update, grammar development and parsing An immediate application is lexicon updating and grammar development foreseeing a positive effect on parsing accuracy and coverage. Lexical acquisition methods provide a quick means of compiling a collection of expressions that satisfy pre-specified (morpho-)syntactic requirements. The lexical acquisition model used to extract support verb constructions can be easily modified for extracting any syntactic pattern shown by fixed expressions. Such patterns may be collocations, support verb constructions, idiomatic expressions, verb-particle combinations, etc. With this model, the existing lexicon can be expanded with new types of fixed expressions.

After the compilation of a computational lexicon, one can explore the syntactic distribution and behavior of the lexical items in corpora. Such
findings motivate certain decisions in grammar development and implement-
ation. The corpus-based method described in chapter 7 extracts evidence of
alternations that involve determiners, quantifiers, etc., morphological pro-
ductivity and adjectival modification. From this evidence, one can infer
morphosyntactic restrictions of a given fixed expression that the lexicon and
the grammar ought to capture.

Improvements in the lexicon and in the grammar ought to be reflected in
parsing accuracy. Access to a lexicon of prepositional support verb construc-
tions and idiomatic expressions has been shown to improve PP–attachment
disambiguation decisions made by a parser (Volk, 2002).

**Linguistic exploration**  The methods described extract valuable empirical
evidence of actual language use. This evidence can be very helpful to test,
revise and improve existing linguistic theories of idiomaticity.

**Computational Lexicography**  In recent years, a growing interest has led
researchers to compile databases of idioms, collocations and other figurative
expressions to be integrated in natural language processing systems as well
as for linguistic exploration. Examples are: a German database of verbal
idioms called *Phraseo-Lex* (Dormeyer et al., 1998), a collection of German
collocations by Krenn (2000b), *SAID* – a Syntactically Annotated Idiom
Dataset of English idioms produced by the Linguistic Data Consortium (LDC)
and a database of Dutch fixed expressions involving prepositions by Loonen
(2003), to name a few. In addition, large dictionaries of fixed expressions and
sayings exist in many languages: the *Cambridge International Dictionary of
idioms, *Van Dale Idiom Woordenboek* (1999) with circa 10,000 Dutch idioms,
etc. Recently, machine readable dictionaries and/or printed dictionaries have
started to include morphosyntactic preferences of the lexemes and phrases
inside fixed expressions.

Gaps in existing dictionaries (printed and electronic ones) are well-known
to lexicographers; lexicographic projects could profit from a hybrid identifi-
cation model that automatically extracts thousands of patterns from large
corpora to update and extend existing dictionaries. Hybrid models have
been successfully applied in English and German lexicographic projects.
After applying filtering techniques (informed by linguistic diagnostics), a
lexicographer would need to manually inspect the extracted lists. Further,
morphosyntactic constraints that affect the internal phrasal structure in fixed
expressions can be helpful to language learners, natural language generation
systems, parsing, etc. Automatic extraction of these constraints facilitates a
lexicographer’s work.

Information Retrieval and Machine Translation In cross-lingual information retrieval, a glossary or index of fixed expressions (or domain specific multi-word terms) is needed in source and possibly target languages because concepts are often expressed as multi-word terms (Thurmair, 2004). A word by word translation of the original multi-word term is likely to destroy the original meaning of the source term. In open domains, one would also expect that a list of fixed expressions can be helpful in both applications.

Similar collocation identification methods have been applied to acquire multi-word lexemes from specialized corpora. Attempts to extract NOUN CASE MARKER VERB multi-word lexemes from a Hungarian corpus have produced reasonable results (Kis et al., 2004). Extracted lists are used in machine translation.

1.6 Chapter summaries

Chapter 2 provides a general characterization of fixed expressions from a descriptive linguistics perspective. Lexical, morphological, semantic and syntactic properties and irregularities of fixed expressions are discussed. Properties of fixed expressions are taken into account during the design of the corpus-based identification models. An awareness of these properties is crucial for assessing the results of the identification models as well as the limitations encountered while exploring the modification potential.

Chapter 3 describes data-driven methods previously applied in the identification of collocations. We describe hybrid approaches that combine linguistic information with probabilistic techniques and contrast them with purely statistical approaches. The chapter describes lexical resources used, methods adopted and evaluation methodology. Furthermore, we also comment on the success of hybrid approaches.

Chapter 4 reports on the performance of a hybrid approach in identifying Dutch collocational prepositional phrases in a corpus. Collocational prepositional phrases are rather fixed PREPOSITION NOUN PREPOSITION combinations. For this case study, we used a part-of-speech tagged corpus. Various statistical association measures are compared.

Chapter 5 aims at establishing how useful a hybrid approach is in identifying support verb constructions in corpora. The required constituents in a support verb construction may be separate from each other. Therefore, we use automatically parsed data. We compare the accuracy of standard association measures such as log-likelihood and $\chi^2$ to a log-linear model.
Chapter 6 investigates whether linguistic diagnostics are helpful for the task of discarding noise from automatically extracted lists of support verb constructions. Various linguistic properties introduce distinctions between regular phrases and fixed expressions. We rely on such distinctions to improve the output of identification models.

Chapter 7 describes a corpus-based method to extract relevant information from corpora and investigate whether the method is useful to determine modification and variation potential within one class of Dutch fixed expressions, that is, support verb constructions that require a PP argument.

Chapter 8 enumerates our conclusions and briefly establishes future directions.