Chapter 6

Additional Linguistic Diagnostics

Abridged versions of this chapter were originally published as Villada Moirón (2004a,b).

6.1 Introduction

Association measures capture lexical affinities between words but fail to distinguish productive verb phrases from idiosyncratic lexicalized ones. After extracting a list of potential support verb constructions from a corpus, we apply additional linguistic diagnostics to discard the noise.

This chapter aims at filling in an important gap in the validation of the nbest lists proposed by statistical measures used in automatic acquisition of Dutch support verb constructions. In particular, we pursue a method of distinguishing regular prepositional complements and adjuncts of full verbs from fixed arguments of support verbs. If we can eliminate the noise from the automatically acquired lists in a systematic way, the identification methods will produce more reliable lexica.

We investigate a filtering mechanism that applies additional linguistic diagnostics to discard noise from the automatically extracted nbest list. The method uses some of the linguistic diagnostics proposed by Hollebrandse (1993) to distinguish regular verb phrases from light verb constructions. In the remainder of this section, we present the sort of expressions that add noise and an overview of the chapter.
6.1.1 Sources of noise in nbest list

In our experiments we focused on the automatic extraction of expressions consisting of the verb *houden* ‘to hold’ and a prepositional phrase. Among the higher ranked expressions, some show the [verb preposition] combination *Houden aan* that could appear in examples like:

(1) Ik houd me 
    aan die afspraak.
    I hold myself to this agreement
    ‘I adhere to this agreement.’

(2) Die vroeg de journalist om 
    de man aan de praat te 
    houden.
    hold
    ‘He asked the journalist to keep the man talking.’

As the translations indicate, *houden aan* in (1) means ‘to adhere to’, something different from *houden aan* in (2) ‘to keep someone hanging on’; *Houden aan praat* constitutes part of a support verb construction when it appears in examples like (2) above. In this case, *houden* behaves like a typical support verb because the verb itself does not contribute the main semantic relation denoted by the predicate. It is the combination of *houden* and the PP (e.g. *aan de praat*) that supplies the core meaning of the verbal predicate. *Houden* contributes tense, aspect (progressive action), aktionsart (continuation) and causation. On the contrary, when *houden aan* means ‘to adhere to’, the verbal lexeme denotes meaning on its own. Furthermore, the object NP slot is free, that is, there are many nouns that can fill in the slot provided their semantic content is compatible with the verb’s meaning.

The examples in (1) and (2) illustrate two types of expressions: a support verb construction (e.g. *iemand aan de praat houden* ‘to keep s.o. hanging on’ in (2)) and combinations of a verb with a prepositional complement (e.g. *(zich) aan de afspraak houden* ‘to adhere to the agreement’ in (1)). In this second case, we do not have a fully lexicalized support verb construction but a *syntactic colligation* (Sinclair, 1966) or a *grammatical collocation* (Everaert, 1993).

PP complements of a full verb are considered as noise because they can be treated by regular syntactic rules. Other types of noise in the *nbest* list are:

- locative PPs (*houd onder kraan* ‘hold under the tap’), temporal PPs (*houd op zaterdag* ‘hold on Saturday’) and directional PPs (*houd naar kapel* ‘(lit.) hold towards the chapel’);
• PPs whose head preposition introduces a required complement inside a nominal or adjectival support verb construction, for example, *houd met wensen* ‘(lit.) hold with wishes’ and *houd van politiek* ‘(lit.) love politics’. These PPs may occur in expressions such as *rekening houden met* ‘take something into account’ and *zich ver houden van* ‘stay away from’, respectively;

• other adjunct PPs that are not syntactic dependents of a specific verb (*houd onder auspiciën* ‘hold under the auspices’). Some of them show idiosyncratic morphosyntax (*houd tot taak* ‘hold as task/aim’, *houd in dienst* ‘hold in duty’, etc.).

### 6.1.2 Overview

We describe a filtering mechanism that builds on the assumption that the syntactico-semantic irregularities of *svces* can be identified by some linguistic diagnostics. The method proposed automatically applies various linguistic diagnostics that distinguish productive and regular verb phrases from idiosyncratic constructions. Section 6.2 summarizes and illustrates some of the linguistic diagnostics proposed by Hollebrandse (1993). We describe a method of applying the diagnostics semi-automatically in section 6.3. A discussion about the efficiency of the linguistic diagnostics and the proposed method is the focus in section 6.5. Section 6.6 summarizes our conclusions.

### 6.2 Diagnostics and evidence

Hollebrandse (1993) motivates a distinction between Dutch full verbs and light verbs drawing on tests that check morpho-syntactic and semantic features of the expressions. Among the diagnostics proposed by Hollebrandse (1993), we selected NP pronominalization, scrambling, PP over verb, nominalization and coordination. Sections 6.2.1–6.2.4 provide relevant examples of the diagnostics in application. Such examples are taken from the Twente Nieuws corpus (TWNC) (Ordelman, 2002).

#### 6.2.1 NP pronominalization

If the noun phrase (NP) object inside the prepositional complement can be realized as a ‘clitic’ (namely ‘r, ‘t, ‘m) or the referential *er* pronoun, then the
combination of verb + PP is a regular verb phrase.\(^1\)

NP pronominalization is possible with some expressions like \textit{zich aan de wet houden} ‘obey the law’:

\begin{quote}
(3) Hoewel niet alle rechters gelukkig zijn met \textit{deze wet}, houden de meesten zich er toch aan ‘Although not all judges are lucky with this law, most of them still obey it.’
\end{quote}

We found no examples of pronominalized NPs where the antecedent would be the object noun inside a light verb construction. Failure to find evidence of NP pronominalization inside PPs thought to be arguments of an SVC does not mean pronominalization is ruled out. However, grammaticality judgments of native speakers suggest that NP pronominalization inside an SVC like (\textit{iets of iemand}) \textit{in de gaten houden} is not possible. Failure to allow NP pronominalization indicates that the NP inside the fixed argument has no meaning on its own and it is not referential.\(^2\) This is evidence that the verb and PP word combination is fully lexicalized, a common feature in non-compositional \textit{fixed expressions}.

### 6.2.2 Scrambling

Prepositional complements may appear in three different locations in the Dutch sentence: initial position, middle position and following the verb cluster. PPs that constitute part of an \textit{uitdrukking} (i.e. fixed expression) cannot follow the verb cluster. If they do, the verb phrase loses its idiomatic interpretation (Haeseryn et al., 1997)[p.1375].

Concerning prepositional complements of a full verb, the PP tends to \textit{immediately} precede the verbal head in the verb cluster. However, tense auxiliary verbs and modals may separate the PP complement from its verbal head as in:

\begin{quote}
(4) Haar zus zou zich niet aan die afspraak hebben gehouden.
Her sister might self not on the agreement have held ‘Her sister might not adhere to this agreement.’
\end{quote}

---

\(^1\)The clitics ‘r, ‘t, ‘m correspond to the accusative feminine, neuter and masculine pronouns (English, \textit{her}, \textit{it} and \textit{him}).

\(^2\)Kuhn (1994) gives examples of PPs in German support verb constructions with a referential object noun: \textit{jdn. auf eine Idee bringen} ‘give s.o. an idea’. We do not rule out the possibility that referential nouns surface in transparent support verb constructions.
Scrambling arises when an adjunct occurs between a complement and its verbal head in verb final contexts. We hardly find examples where a PP complement of a verb (houden) is scrambled. As an illustration, an intervening locative adjunct in (5).

(5) Als je je niet aan de regels hier én in andere landen wilt houden, moet je daar de consequenties van dragen. ‘Here and in other countries, if you don’t adhere to the rules, you’ll have to face the consequences.’

If the prepositional phrase is scrambled, then the PP is not a fixed argument of a light verb. No examples of scrambling with supposedly complements of the light verb were found.

6.2.3 PP over verb and nominalization

PP over verb In Dutch verbal clusters, the verb tends to occupy the final position. PP complements may easily be located after their verbal head in a non-finite clause or in a finite subordinate clause; this configuration is known as PP over verb or PP extraposition. (6) offers an example where the PP complement aan de regels follows houden, its verbal head:

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

According to Haeseryn et al. (1997), those PPs that constitute part of a fixed expression cannot follow the verb cluster; otherwise, the verb phrase loses its idiomatic interpretation. This explains why only prepositional complements of full (non support) verbs exhibit PP over verb.

The expression een vinger aan de pols houden ‘to keep up with the latest developments about sth’ is a fixed expression. In (7), the PP argument aan de pols also follows its lexical head within the infinitive VP.

---

3The locative phrase may be interpreted as modifying the preceding noun (regels) or as an event modifier, thus, functioning as an intervening adjunct. Only in the second interpretation is this a valid example of scrambling.
Chapter 6. Additional Linguistic Diagnostics

(7) Laten we blij zijn dat de overheid wil weten wat er onder de burgers leeft, een vinger wil houden aan de pols van de samenleving on the pulse of the society
‘Let’s be happy that the government wants to keep up with the mind of the civilians and the latest developments in society.’

Various informants find (7) a surprising example and question its well-formedness. The PP extraposition seems to be due to heaviness, an effect of the additional material that modifies pols (the PP van de samenleving).\(^4\)

Nominalization A Dutch verb in its infinitive form can be nominalized through the addition of the definite determiner het. Nevertheless, a determiner (definite, possessive, etc.) is not compulsory. Eten ‘to eat’ is nominalized into het eten ‘the eating’. When the verb has a PP complement, one encounters (at least) two types of nominalization: (i) PP complement precedes the nominal infinitive or (ii) a PP follows the nominal infinitive. We refer to these nominalization patterns as PP verb and verb PP, respectively. The verb PP pattern can be considered a sub-case of PP over verb because the complement PP follows its lexical head, the nominal infinitive.

Hollebrandse (1993) adds that examples of nominalization of prepositional light verb constructions where the PP follows the nominal infinitive are difficult to find. He only provides an ungrammatical example of a light verb construction headed by nemen ‘to take’: *Het nemen in arrest van de criminelen (lit. ‘The taking under arrest of the criminals’).

Example (8) illustrates the nominalization of the predicate zich aan de regels houden ‘to adhere to the rules’.

(8) Je niet houden aan de regels van het dualisme is de grootste mogelijke zonde.
‘The biggest possible transgression is to not adhere to the rules of dualism.’

Two corpus examples showing the nominalization pattern PP verb involve support verb constructions:

\(^4\)This was suggested by Jack Hoeksema (p.c.).
6.2. Diagnostics and evidence

(9) De leden van dit team houden zich alleen bezig met het in de gaten houden van ‘verdachte personen’.

The members of this team hold themselves only busy with ‘the in the holes hold of suspected people’.

‘The members of this team keep themselves busy by keeping an eye on ‘suspects’.’

(10) Want op enige financiële steun van de politiek voor ‘het van de straat houden van jongeren’ hoeven je als voetbalclub niet te rekenen.

Because on only financial support of the politics for ‘the of the street hold of youngsters’ need you as football club not to count

‘As a football club, you should not count on any financial support from politics for keeping the youngsters off the street.’

6.2.4 Coordination

If a PP dependent is coordinated with a regular PP complement of the same verb, then the verb is a full verb. According to Hollebrandse (1993) fixed arguments of light verbs may coordinate with other fixed arguments but not with regular complements of a full verb. The intuition behind is that if coordination of a regular complement and a fixed argument of a light verb were possible, then we would have two lexemes in one: a full verb and a light verb. The two instances of the verb are different lexemes, which denote different predicates and (may) exhibit different valence requirements.\(^5\) Example (11) shows the coordination of two regular complements and (12) shows the coordination of two fixed PP complements of a light verb:

\(^5\) The structure [VERB FIXED ARGUMENT COORDINATION REGULAR COMPLEMENT] (as in (i)) would be an instance of zeugma, that is, a figure of speech that gives a punning effect. Zeugma shows in a language as the coordination of items that play different syntactic or semantic roles.

(i) ?Ze houden van bloemen en elkaar aan het lijntje.

they hold of flowers and each other on the little line

‘?They love flowers and keep each other dangling.’

The semantic relations of \textit{van bloemen} and \textit{aan het lijntje to houden} are different. The meaning of \textit{van bloemen} is fully compositional but not the meaning of \textit{aan het lijntje}. Native speakers assess example (i) as bad, which suggests that instances of zeugma might be rare and considered a source of ill-formedness in the language.
(11) Ook de beste zakenpartners zijn die zich aan de wet en aan internationale afspraken houden.
Also the best business partners are they who selves on the law and on international agreements hold
‘The best business partners are those who comply with the law and international agreements.’

(12) Ze houden elkaar aan de gang en in bedwang, als de radertjes van een monsterlijk, zichzelf opwindend horloge.
They hold each other on the go and in control, as the cogwheels of a monstrous, self-up-winding clock
‘They keep each other in motion and in control, as the cogwheels of a monstrous and fascinating clock.’

6.2.5 Some additional remarks
Hollebrandse (1993) adds that NP ellipsis, wh-movement, heavy-NP shift and binding phenomena are possible in regular verb phrases but not in light verb constructions. Furthermore, adjectival modification, pluralization and the use of diminutive are rather restricted inside the complements of support verb constructions. All diagnostics are important to determine whether a PP is part of an SVC or of a regular VP; however, we concentrate on a few diagnostics that can be checked semi-automatically. In Chapter 7, we investigate the variation and modification observed in support verb constructions.

6.3 Applying diagnostics semi-automatically
According to the theoretical expectations the diagnostics enumerated above should show a distinction between regular prepositional phrases and arguments in light verb constructions. Thus, in our preliminary experiments, we apply a semi-automatic method to explore if the linguistic diagnostics successfully capture the distinction and to determine which diagnostics help to discard the existing noise.

6.3.1 Data resources and tools
As inputs, we are given a list of expressions ([VERB PREP NOUN] combinations) found among the top scores in the nbest list. The list is an excerpt of the output of a statistical identification model.

The Twente Nieuws Corpus (TWNC), made up of newspaper text and some television news reports (Ordelman, 2002) was already tokenized and
6.3. Applying diagnostics semi-automatically

prepared for further processing. Further, the T2NC was processed by an information retrieval tool called mg (Witten et al., 1999) to extract sentences that contain certain words or word combinations.

The corpus data was annotated with the Alpino parser earlier described in section 3.2.2. The parser assigns each corpus sentence a DEPENDENCY STRUCTURE, which is a relatively theory independent annotation format. Alpino tries to follow the annotation guidelines and standards used in the project Corpus Gesproken Nederlands (Corpus of Spoken Dutch).\(^6\) Dependency structures represent the grammatical relations that hold in and between constituents. In addition, the syntactic category, lexical entry and string position are added to each leaf. Figures 6.1 and 6.2 show a dependency structure tree and its corresponding XML encoding, respectively.

Figure 6.1: Dependency structure assigned to the sentence *Ik moet het wel goed in de gaten houden*. ‘I have to keep a good eye on it’.

Further, to explore in which contexts the expressions occur, we used \texttt{dt\_search}. This is a Perl script built around XPATH, a language that operates on the abstract, logical structure of an XML document and allows navigating through the hierarchical structure of the document. XPath models an XML document as a tree of nodes.\(^7\) Applied to parsed data, an XML document (fig. 6.2) depicts a tree representation of a sentence (fig. 6.1).

\(^6\)http://lands.let.kun.nl/cgn.
\(^7\)More specific details available at http://www.w3.org/TR/xpath.
Figure 6.2: XML dependency tree depicting the parsed sentence *Ik moet het wel goed in de gaten houden* ‘I have to keep a good eye on it’.
Bouma and KloosterMAN (2002) developed \texttt{dt\_search} to support treebank queries involving constituent dependency and word order constraints over dependency trees encoded in XML.

### 6.3.2 Method

To determine what diagnostics are satisfied by the expressions in the \textit{nbest} list, we first extract sentences that include the three lemmas inside the expression from the TwNC corpus. These sentences are collected in a subcorpus that is parsed automatically. The parsed trees are represented as XML dependency trees. Next, \texttt{dt\_search} queries are used to determine what expressions exhibit a syntactic structure with scrambling, PP over verb, etc.

**Data annotation** All subcorpora were parsed with the \texttt{Alpino} parser. No error correction was performed on the parsed data and only the best parse is returned by the parser. As illustration, the sentence in (13) is represented by the XML encoded dependency tree in Figure 6.2.

\begin{equation}
(13) \text{Ik moet het wel goed \textit{in de gaten houden}.}
\end{equation}

\begin{itemize}
  \item I
  \item must
  \item it
  \item well
  \item good
  \item in
  \item the
  \item holes
  \item hold
  \item ‘I must keep a good eye on it.’
\end{itemize}

This shows a support verb construction that is already annotated in the \texttt{Alpino} lexicon \textit{(iets in de gaten houden)}. In Figure 6.2, each node in the dependency structure is labelled with a dependency relation and the syntactic category (phrasal or clausal) of the constituent subsumed by the node; furthermore, the node specifies other attributes such as \texttt{pos} (part of speech) information of the terminal lexeme, beginning (\texttt{begin}) and end (\texttt{end}) position and, some nodes also include the lexeme’s lemma and surface form (\texttt{root} and \texttt{word}). For example, the dependency tree includes a subject (\texttt{rel="su"}), an object (\texttt{rel="obj1"}), two modifiers (\texttt{rel="mod"}) and a lexicalized phrase (\texttt{rel="svp"}) occurring within the local context of the verbal head (\texttt{rel="hd"}) \textit{houden}. The local context of \textit{houden} is subsumed by the node with a verb cluster label (\texttt{rel="vc"}), framed in an infinitival verb phrase (\texttt{cat="inf"}). Lexicalized expressions introduce a ‘fixed’ argument (\texttt{pos="fixed"}) that does not project any phrase. In the tree, \textit{in de gaten} is a lexicalized string.

**Applying diagnostics** \texttt{dt\_search} searches through the XML-encoded dependency trees and extracts those sentences that satisfy the conditions expressed in the query. \texttt{dt\_search} allows us to specify head-complement dependencies, lexical restrictions, word order constraints and, clause type re-
strictions in the query. Finally, a human assesses the retrieved sentences to verify whether the evidence is valid.

As an example we show how one of the diagnostics, scrambling, can be encoded in a `dt_search` query. The example concerns the expression *aan afspraak houden* (zieh aan de afspraak houden ‘stick to an appointment’). We refer to the PP inside the expression as the *target PP* from here onwards.

To find out if an adjunct occurs between the target PP and *houden*, the query below searches for a node in a dependency tree with category value PP (@cat="pp"); this PP node should have two daughter nodes, one whose terminal has the preposition *aan* as value of the attribute word, and another daughter – the object of the preposition *aan*; further, inside the object, the head NOUN specifies its lemma *afspraak* (@root="afspraak"). An additional constraint requires that an adverbial phrase (@cat="advp") exists between the target PP’s end position and the beginning (begin) of the node headed by *houden*. Word order constraints are stated by the node attributes @begin and @end.

```
dt_search -s '//node[@cat="pp" and ./node[@word="aan"] and ./node[@rel="obj1" and ./node[@rel="hd" and @root="afspraak"]]
  and (@end = //node[@cat="advp"]/@begin)
  and (@end < ../node[@rel="hd" and @root="houd"]/@begin)],
  houd.aanafspraak.sub/*.xml
```

Adjuncts can also be realized by other phrase types, not only adverbial. A more general query states no constraints on the phrasal category. The following query extracts all sentences where a constituent functioning as a modifier (@rel="mod") occurs between the PP and its verbal head *houden*:

```
dt_search -s '//node[@cat="pp" and ./node[@word="aan"] and ./node[@rel="obj1" and ./node[@rel="hd" and @root="afspraak"]]
  and (@end = //node[@rel="mod"]/@begin)
  and (@end < ../node[@rel="hd" and @root="houd"]/@begin)],
  houd.aanafspraak.sub/*.xml
```

**Manual analysis of extracted evidence** To assess whether the chosen diagnostics efficiently identify an SVC, we focused on expressions with *houden* ‘hold’ in our preliminary experiments. To this end, the sentences extracted with `dt_search` queries were manually checked. Two native speakers determined whether (i) the retrieved sentence effectively illustrates the diagnostic being tested and, (ii) if (i) is affirmative, whether the expression has a figurative (opaque) interpretation or a literal interpretation.
6.3.3 Preliminary results

Pronominalization, PP over verb and the nominalization pattern point at differences between an SVC (e.g. *iemand in de gaten houden* ‘keep an eye on s.o.’) and a regular verb phrase (e.g. *zich aan de regels houden* ‘adhere to the rules’). Scrambling is useful to distinguish optional adjuncts from complements, but it does not always show a distinction between regular prepositional complements and fixed arguments in an SVC. Finally, coordination is a weak test because before judging what type of coordination an expression exhibits, one needs to know whether the PP is part of a fixed expression or not. As an illustration, Table 6.1 shows which diagnostics are satisfied by the expressions on the left column; these were ranked among the highest scores in the *nbest* list.

<table>
<thead>
<tr>
<th>Nbest candidate expression</th>
<th>pro</th>
<th>scr</th>
<th>PP over V</th>
<th>coord PP</th>
<th>SVP</th>
<th>nom PP</th>
<th>V PP</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>houd aan praat</em> ‘keep s.o. hanging on’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd in bedwang</em> ‘keep s.o. in control’</td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd in gaten</em> ‘keep an eye on’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd in stand</em> ‘keep in existence’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd voor gek</em> ‘make a fool of’</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd oogje in zeil</em> ‘keep a good eye on’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd aan afspraak</em> ‘adhere to an agreement’</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>houd aan regels</em> ‘adhere to the rules’</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| *houd met wensen* lit. ‘keep with wishes’ | | | | | | *
| *houd onder auspiciën* lit. ‘hold under auspices’ | * | * | * | |
| *houd van sport* ‘love sport’ | * | * | | | |

Table 6.1: Diagnostics evidence. pro stands for pronominalization, scr for scrambling, coord checks coordination pattern (PP or an SVP(fixed argument)), nom states nominalization pattern.
6.4 Evaluation

To assess whether the diagnostics help to reduce the noise in automatically extracted nbest lists, we selected 7 Dutch support verbs. From the nbest list, we extracted the 100 highest ranked expressions for each of 7 verbs. Next, we randomly collected 10% of the expressions related to each verb. Thus, we had a list of 70 expressions that were ranked among the higher scores by the salience statistic.

During automatic extraction of datasets clause boundary information was ignored. For this reason, the nbest list contains expressions where the verb and the PP never or almost never co-occur within the same minimal clause. Applying the diagnostics to such expressions is meaningless, which led us to remove 6 items in the test data.

6.4.1 Methodology

The list of 64 expressions was given to three human judges that are Dutch native speakers. They were asked to assign a ‘1’ if they considered the expression (part of or) a lexicalized verb phrase (svc), a ‘0’ if they could not think of a related lexicalized phrase and a ‘?’ if they knew a lexicalized phrase headed by a different (support) verb but with the same PP. We allowed the third judgement because some PPs co-occur with more than one support verb denoting different aktionsart (e.g. op bezoek krijgen/hebben ‘get/have a visit’). Our gold standard list consists of those expressions assigned a ‘1’ by at least two judges or expressions assigned a ‘1’ and two ‘?’. According to the statistic all the 64 expressions are svc’s. However, according to the human judgements, 54.7% of the expressions in test data are false positives (our baseline).

We took the test data (N=64) and applied all diagnostics except coordination. This time, the evidence retrieved was not attested by a human, thus we relied completely on a fully automatic method. Expressions that allow pronominalization, scrambling, PP over verb or show the nominalization pattern V PP are false positives. Expressions that satisfy no diagnostics or only show the nominalization pattern PP V are considered true positives (svc’s).

6.4.2 Results

Using the human judgements as reference, the diagnostics correctly classify 44 items (27 as true positives and 17 as false positives). This also means that the diagnostics correctly assess an item among the automatically extracted
expressions as a true SVC or as noise in 70% of the cases, which is a positive outcome.

Diagnostics and human judges disagree on: (i) expressions consisting of a predicative PP (*in beroering* ‘in movement’), (ii) a homonymous expression whose PP may be part of an SVC (*iemand van zijn stuk brengen* ‘to surprise s.o.’) or a modifier with only literal interpretation, (iii) one expression misparsed by the parser that the human judges recognized as a true SVC (*het niet in je kouwe kleren gaan zitten* ‘to have an effect on; sth. leaves a mark on you’) and (iv) two directional PPs evaluated as SVCs by the diagnostics (*naar bed gaan* ‘go to bed’). For some expressions, no evidence was found of any of the diagnostics. This can be interpreted in two ways: either the expression satisfies none of the diagnostics or our corpus is not representative of the phenomena.

6.5 Discussion

The method described is built on a number of assumptions; some of these assumptions encountered a few limitations that make the process of distinguishing support verb constructions from regular phrases less robust.

6.5.1 Efficacy of the linguistic diagnostics

SVCs exhibit a number of syntactico-semantic irregularities that distinguish them from regular verb phrases. Such irregularities are indicated by a few of the linguistic diagnostics proposed by Hollebrandse (1993).

Evidence of PP over verb groups regular prepositional complements of full verbs, PP complements inside a nominal or an adjectival support verb construction (for example, *rekening houden met* ‘take into account’ where *met* introduces a prepositional complement) and adjuncts. These are all expressions that we consider noise. (14) shows an example of the PP *onder auspiciën* ‘under auspices’ that introduces an adjunct.

(14) Dat blijkt uit een internationale enquête onder slachtoffers,
That appears out an international survey under victims,
die werd gehouden onder auspiciën van de Verenigde Naties.
that was held under auspices of the United Nations
‘That comes out of an international survey about victims, that was conducted under the auspices of the United Nations.’

Neither prepositional complements of a full verb nor fixed arguments of a light verb seem to allow scrambling (at least) in written text. During
our preliminary experiments, we found rare instances of scrambling of a full verb’s prepositional complement. In contrast, scrambling distinguishes adjunct phrases from other complements. Note that adjuncts are a frequent source of noise in automatically acquired lists, what makes scrambling an effective test.

Regarding coordination, the attested evidence (in preliminary experiments) corroborates Hollebrandse’s (1993) observation that a fixed argument is typically coordinated with another fixed argument, whereas regular complements coordinate with similar complements. Let us give some examples.

Houd van straat allows coordination with an apparently regular PP (15) but it shows the nominalization pattern of a non-regular phrase (10). At first sight, (15) looks like a counterexample; however, the second PP conjunct (van het slechte pad) requires metaphorical interpretation too; therefore, both PPs are somehow figurative.8

(15) De jongerenleiders van de Milli Güors houden de jongeren van de straat en van het slechte pad.

The young-leaders of the Milli Guors hold the youngsters off the street and off the bad path.

‘The young leaders of the Milli Guors keep the youngsters off the street and the wrong path.’

The distinct nature of the coordinating conjuncts would allow a mechanism to distinguish regular phrases from svcs, however, there is a problem. A fully automatic method cannot infer a priori the type of conjuncts (a fixed argument or a regular complement) shown by a candidate expression. Therefore, we excluded coordination in the evaluation described in section 6.4.

From a theoretical perspective, any complement of a verb may appear to the left or to the right of the corresponding nominalized infinitive (Reuland, 1988; Haeseryn et al., 1997). Consequently, we expect that the nominalization of a verb taking a prepositional complement shows both patterns: PP over verb and verb PP. Among the expressions examined, all instances found in our corpus that exhibit the nominalization pattern VERB PP involve regular complements. In contrast, expressions that exhibit the PP VERB pattern are support verb constructions. Furthermore, the first group of expressions allow PP over verb, whereas the second does not.9 These findings suggest that Hollebrandse’s observation is valid (see section 6.2.3) and they unveil an im-

8Other lexicographers also consider iemand van de straat houden ‘keep s.o. off the streets’ a lexicalized phrase (cf. Van Dale (1986), a bilingual dictionary).

9We only found one example of an SVC that shows the PP argument post-verbally; the example given earlier in (7) was characterized as ‘questionable’.


important difference between PP arguments of a *light verb* and PP complements of a full verb.

Pronominalization, *PP over verb* and nominalization group PPs in the nbest list in two classes: fixed PP arguments of a light verb in one class and regular PP complements and adjuncts in a second class. Because PP complements of both full and light verbs show a preference to *immediately* precede the verbal head inside the verb cluster, scrambling does not show a clear distinction between the two types of argument. However, scrambling distinguishes adjuncts from complements, thus, identifying a frequent type of noise. Coordination is a weak test because before judging the type of coordination shown by an expression, the model needs to know if a PP is a complement or a fixed argument, what makes the test less helpful.\(^\text{10}\)

To sum up, expressions that exhibit PP over verb, scrambling, the VERB PP nominalization pattern or pronominalization are regular verb phrases. Expressions that show the PP VERB nominalization pattern and none of the other diagnostics are ‘possibly’ SVCs. Note that predicative and directional PPs in our test data exhibit similar behavior as fixed arguments in SVCs; this is why the diagnostics classify such PPs as correct SVCs.

### 6.5.2 Annotated data, search queries and evidence

We started from the assumption that linguistic diagnostics can be effectively translated into *dt_search* queries and, these queries retrieve relevant evidence from syntactically annotated corpora. Automatically retrieved evidence is sufficiently reliable and on the basis of this evidence, one can accurately distinguish SVCs from regular verb phrases. In the following, we report limitations encountered by our assumptions.

We investigated whether with the available resources and tools, we can automatically discard noise from the nbest list.

**Annotated data**

The success of the method is highly dependent on parsing accuracy (as well as on the correctness of the search queries). If a sentence was erroneously parsed, the retrieved parsed evidence is likely to be wrong.

PP–attachment disambiguation is a well-known difficult task. The Alpino parser has trouble deciding whether to attach a PP to a preceding noun or to a verbal head. Typically, the parser favors noun attachment. A PP part

\(^{10}\)One could use the coordination test in conjunction with the results of the other tests. Once it is known that a $PP_X$ is a semi-fixed expression and it coordinates with $PP_Y$, then $PP_Y$ is also semi-fixed.
Chapter 6. Additional Linguistic Diagnostics

of a support verb construction is sometimes wrongly analyzed as a nominal post-modifier or as a PP dependent in a nearby clause. These misparses bring up two problems:

- queries stating a head complement dependency between the verb and the target PP will not retrieve a PP embedded under a noun local node (where the PP is a postmodifier of the head noun).

- queries that need to select specific clause types (subordinate, infinitival) could be more precise if we state the clause type explicitly. If a PP was misparsed as dependent of an unrelated verb, \texttt{dt\_search} may fail to locate the PP in the specific clause we are interested.

To avoid both previous weaknesses, the search queries are stated more generally trying to anticipate the parser’s decisions.

Search queries

Any linguistic diagnostic can be encoded using the expressive power of \textsc{xpath} and the tool \texttt{dt\_search}. The challenge is to anticipate the possible analyses proposed by the parser in the search queries. More precisely, queries ought to take into account (a) the possibility of incorrect PP–attachment decisions (discussed above), (b) the uncertainty of how the parser treated those lexicalized expressions that are already described in the lexicon and (c), homonymous expressions that have one or more \texttt{svc} uses and a literal use.

Some fixed expressions are already annotated in the Alpino lexicon; the Alpino grammar can license fixed expressions in some frequent syntactic configurations (main and subordinate clauses, passives, questions, etc.) but not in others such as nominalization or coordination. If the parser recognizes a verb phrase as a fixed expression, the fixed PP argument is assigned the syntactic category SVP, otherwise the parser assigns a regular structure to the expression. Search queries need to allow both possibilities, else queries fail to retrieve valuable corpus evidence.

An expression in the \texttt{nbest} list is represented as a pattern of a \texttt{[VERB(lemma) PREPOSITION NOUN(surface form)]}. To guarantee wider coverage of our queries (i.e. to access as much corpus data as possible), queries seek instances of the pattern \texttt{[VERB(lemma) PREPOSITION NOUN(lemma)]} in the required contexts. The general pattern \texttt{[VERB(lemma) PREPOSITION NOUN(lemma)]} clusters several linguistic uses of the expression. E.g. \texttt{[HOU\\texttt{D}(lemma) IN HAND\texttt{]}(lemma)\texttt{]} applies to the fixed expressions \textit{zijn duim in de hand houden} and \textit{het heft in handen houden} ‘be/stay in charge’, as well as to the literal use
6.5. Discussion

*iets in de hand houden* ‘hold something in the hand’. Whereas a more general query retrieves more corpus data, it has the disadvantage of supplying mixed information about the various uses of homonymous expressions.

**Retrieved evidence**

The search queries retrieve a substantial amount of good evidence, but erroneous evidence is also found. We characterize the erroneous evidence retrieved by the queries corresponding to each individual linguistic diagnostic first and erroneous evidence common to all queries afterwards. Coordination is excluded for reasons discussed above in 6.2.4.

**Scrambling** PP constituents with a modifier dependency relation that are located between a target PP and the verbal head often function as complements or alternatively, as modifiers of the head noun within the target PP. Such a PP is not an adjunct, thus the evidence does not exhibit *scrambling*. Queries that search for adverbial phrases, adjective phrases or NPs are more successful.

**PP over verb** Legitimate evidence of *pp over verb* involves verb-final contexts (subordinate or non-finite clauses). Erroneous evidence includes sentences where the verb is finite but the parser failed. In main clauses, the PP complement (or fixed argument) typically follows the verb.

**Nominalization** The parser cannot handle nominalization. Queries do not always anticipate the parser’s analysis. Occasionally, erroneous evidence of the nominalization pattern PP VERB shows a non-finite verb phrase made up of an object NP headed by *het* (indefinite pronoun), the target PP and *houden*. An example is the svc *het voor gezien houden* (lit. it for seen hold) ‘pack it in; approved’ where *het* is an object NP complement required by *houden* and not the definite determiner *het* that sometimes introduces a nominalized VP.

**Pronominalization** Anaphora resolution is not performed by the Alpino parser. Instances of pronominalization within target PPs are sought as follows: we extract sentences where an NP is headed by the head noun inside the target PP (e.g. *afspraak* ‘appointment’); a second requirement is that the verb (e.g. *houden*) heads a clause with a pronominalized PP, whose head is a specific preposition (*aan*) and whose object is an R-pronoun (*er, daar, hier*). Good evidence is difficult to extract.
Retrieved evidence shows different uses of an expression, thus potentially leading to an incorrect classification. The source of this drawback are homonymous expressions. To increase the precision, search queries should use the noun’s surface form only (see section 6.5.2 above).

A few expressions do not satisfy any of the diagnostics; there are two possible reasons for this. Either, the expression does not allow scrambling, pronominalization, *pp over verb* and, in any case, no instances of coordination or nominalization are found in our corpus. Among fixed expressions, idiosyncrasies vary from expression to expression, thus not all expressions necessarily exhibit the same irregularities. Alternatively, no evidence is found because the corpus size is too small and therefore, not representative of the phenomena. In addition to searching for lexicalized expressions, the diagnostics study the expressions in very specific syntactic configurations; an open question is how big a corpus is needed to guarantee enough data to validate the selected diagnostics?

Possible improvements

Keller and Lapata (2003) argue in favor of using the World Wide Web to avoid data sparseness in building probabilistic language models. We could ameliorate the limited corpus size problem by searching for more data in the World Wide Web. If we can easily retrieve relevant sentences from the Web, then we could apply the same methodology on the extended subcorpora.

To see whether this would lead to some changes, we performed a search on Google to find out whether the pattern *boven hoofd houd* corresponding to the lexicalized phrase *de hand boven het hoofd houden* ‘protect’ allows nominalization and which pattern exhibits. Two valid hits were found, one of them given in example (16). The expression *de hand boven het hoofd houden* shows the PP VERB nominalization pattern thus, showing the trends of other svcs; this suggests that web data can compensate for the lack of corpus data.

(16) Neemt u de heer Rosenmoller serieus met het te lang
Take you the sir Rosenmoller serious with the too long
*de hand boven het hoofd houden* van Varma Singh?
The hand over the head hold of Varma Singh?
‘Do you take Mr. Rosenmoller seriously about protecting Varma Singh?’

The Alpino grammar and the lexicon are constantly being expanded and improved. The coverage of idiomatic phenomena and fixed expressions needs
to be improved. It is precisely for this reason that we investigate methods for automatic lexicon and grammar development. Effective lexicon and grammar development is a circular process. The better quality lexicon we compile, the better the coverage of the grammar and the better the parser decisions. However, to acquire lexica we use data annotated with the grammar and the lexicon we are developing. Ultimately, an improvement in the two resources will contribute an improvement in parsing accuracy.

The selected diagnostics work well for discarding adjuncts and regular prepositional complements but there are still other sources of noise left. Predicative PPs and directional PPs pattern with SVCs. Future research needs to find out other diagnostics that distinguish these PPs from fixed arguments.

6.6 Conclusions

Linguistic theory provides us with some linguistic diagnostics that distinguish full verbs from light verbs. The relevant diagnostics that show a distinction between fixed arguments and prepositional complements were identified, namely, *pp over verb*, pronominalization and nominalization. *PP over verb* and nominalization check the syntactic distribution of the candidate expression. Pronominalization checks the semantic status of the nominal complement within a target PP. With well-refined queries, one may improve the accuracy of the retrieved evidence but for 100% precision, a human needs to assess the interpretation of the retrieved evidence. To conclude, our method can be useful for providing evidence of the syntactic and semantic irregularities that need to be specified or encoded in the lexicon.