Chapter 5

Identification of Support Verb Constructions

The goal of this chapter is to investigate the feasibility of hybrid identification models to acquire fixed expressions of a verbal nature. Thus, the models are tested in the task of identification of support verb constructions (svc).

5.1 Introduction

Parsers that are informed by lexicalist grammars use the argument structure of the main predicate in a clause to build the skeleton of its syntactic representation. The participants in the argument structure are realized by the syntactic constituents required by the verb; these syntactic constituents are listed in the verb’s subcategorization frame. We may conclude that the accuracy of the subcategorization information supplied in a verbal lexical entry is crucial during parsing.

During operation, the parser needs to distinguish which constituents are to be analysed as required arguments of the verb and which as adjunct modifiers. This applies in contexts of PP-attachment site disambiguation. Under a lexicalist approach, the verb in a fixed expression selects its required constituents via the subcategorization frame. The syntactic dependents of a verb may be partially or fully lexicalized phrases; to provide a correct analysis of the lexical and grammatical relations between the verb and its potentially lexicalized arguments, this information needs to be captured in the lexicon.

Support verbs are highly frequent verbs that require the parser to make such decisions. As an illustration, houden ‘hold, keep’ is a rather frequent verb in Dutch (6938 instances in 16 million words); because of its multiple contexts of use, houden is highly polysemous. In the company of the PP voor
Chapter 5. Identification of Support Verb Constructions

de gek (lit. ‘before the crazy’), houden means ‘to fool someone’ as in (1).

(1) Ik heb steeds meer het gevoel dat ik voor de gek word gehouden.
   ‘I have more and more the feeling that I before the crazy was held.
   ‘I have more and more the impression that I am being fooled.’

To correctly analyze the that-clause in (1), a parser needs to consider the context around the verb and treat voor de gek as a required argument of houden, but not as an adjunct. This way, iemand voor de gek houden will be parsed as a lexicalized phrase and ik word voor de gek gehouden will be interpreted idiomatically. Were the PP analysed as adjunct then we would get a nonsense interpretation like (2):

(2) ‘#More and more, I have the impression that I was held before the crazy one.’

Given that support verbs are rather frequent in the language and also highly polysemous, for NLP applications that require parsing of text it is crucial to describe those contexts where support verbs necessitate more complex subcategorization frames in the form of lexical restrictions.

5.1.1 Support verb constructions

Support verb constructions (svcS) have been treated as idiomatic phenomena. svcS consist of a support verb and its required complementation. The relation between the verb and its nominal complement has been described as a mutual lexical selection (Kremn and Erbach, 1994; Kuhn, 1994). NLP implementations of these views capture the mutual lexical selection by fully specifying the lexical co-occurrence between support verbs and their complements in a lexicon. In addition to identifying verb complement combinations that show a strong lexical association, what needs to be established is the degree of lexicalization shown by the required complements.

The complement may be realised by a lexicalized prepositional phrase as in (1) above, a noun phrase as in rekening houden met ‘to take into account’ (3) or an adjectival phrase as in zich bezig houden met ‘be involved with’ (4):

(3) We moeten rekening houden met de wensen van bewoners
   we must account hold with the wishes of inhabitants
   ‘We must take into account the wishes of the tenants.’
5.1. Introduction

(4) Mijn grootste wens zou mijn me ooit alleen met de marketing mijn biggest wish shall be myself ever only with the marketing bezig te houden.
busy to hold
‘My biggest wish shall be to be entirely involved with marketing.’

Some support verbs may select for a lexicalized NP and a PP (het hoofd boven water houden ‘keep up with difficulties’):

(5) Zo houden we het hoofd boven water.
in this manner hold we the head above water
‘In this way, we keep up with difficulties.’

Among prepositional support verb constructions, some exhibit a fixed PP without a determiner ((5), (6)):

(6) in toom houden
in bridle hold
‘keep under control’

Others may allow various determiners and limited adjectival modification ((7), (8)) within the NP. Across the various SVCs, the internal structure inside the NP is not uniform:

(7) in het/een (Vaticaans) gareel houden
in the/a (Vatican’s) harness hold
‘keep under control/keep under the control by the Vatican’

(8) iets tegen het/een zo fel licht houden
something against the/a this thin light hold
‘investigate something’

The noun within the complement is often fixed. Singular/plural alternation and the use of diminutive are strongly restricted. Often, only one number morpheme (singular or plural) maintains the support verb construction interpretation. A change from plural to singular destroys the figurative meaning (9).

(9) Dat houdt de kosten binnen de perken/*perk.
that keeps the costs within the limits/limit
‘That limits the costs.’

Further evidence involves body parts; these are often found in fixed expressions. As an example, hand ‘hand’ occurs in various support verb con-
Chapter 5. Identification of Support Verb Constructions

Constructions headed by the verb *houden* ‘hold’. The singular noun occurs in the fixed expression *iets in één hand houden* ‘be the boss’ and the plural noun is found in *het heft in handen houden* ‘be/stay in charge’. Both, the singular and the plural can be found in the literal use *iets in de hand(en) houden* ‘hold something in the hand(s)’.

Another interesting feature of support verb constructions is that various PP verb combinations share a common PP (10) and their meaning is related. The light verb contributes tense, aspect and ‘Aktionsart’ information, denoting causation, change of state, result of an event, etc. The syntactic constituency of the expression varies according to the light verb.

(10) a. *aan de gang zijn/blijven*  
    *on the go*  
    ‘get started/keep going’

b. *iets aan de gang houden/brengen*  
    *something on the go*  
    ‘keep sth going/get sth started’

Morpho-syntactic information about what type of complements a support verb selects, the shape and internal structure of the complements, the minimum required lexemes and whether the complement may occur with other support verbs need to be specified in a lexicon. Such information is to be taken into account in the design of the identification models. Here, we focus on combinations of a light verb and a PP, and we do not investigate adjectival modification, lexemes variation, etc. The latter properties are investigated in Chapter 7.

5.1.2 Aims and overview

We seek to determine which data-driven methods applied to Dutch corpora serve to identify in what uses support verbs show an association with prepositional complements. These methods use statistical information to infer dependencies between the support verb and its complement(s). We consider an identification method effective when it extracts as many good SVCs as possible with few false hits in the extracted list.

We aim at identifying uses of support verbs selecting a specific prepositional complement. We treat support verb constructions as a type of collocation assuming that there is a mutual lexical selection between the support verb and a specific PP (following Krenn (2000b)).

This chapter describes experimental work carried out with different statistical models. Section 5.2 describes the experimental settings, pre-processing
5.2 Candidate extraction from corpora

and the extraction of datasets. Section 5.3 outlines the identification process itself and the evaluation methodology. Throughout the experiments, we assess the performance of various association measures, raw frequency and a loglinear model. Sections 5.4 and 5.5 present the results. Our results agree with previous findings reported in the literature and show evidence that the log-likelihood chi-square test works well for the identification of support verb constructions; nevertheless, the salience measure reaches better accuracy. We argue that a test which is not dependent on a frequency cutoff is preferable. Section 5.6 offers our conclusions.

5.2 Candidate extraction from corpora

In support verb constructions, the prepositional complement and the light verb often occur separated from each other due to intervening adjuncts, other complements and auxiliary or modal verbs. For this reason, it is convenient to use corpora with rich linguistic annotation that allows us to extract potentially discontinuous components of a support verb construction.

5.2.1 Preprocessing

Two different Dutch corpora were used: (i) De Volkskrant on CDROM - 1997 edition which contains ca. 16 million words, and (ii) the NRC 1994 section from the Twente Nieuws Corpus (TwNC 0.1) (Ordelman, 2002) which contains roughly the same amount of data. Both corpora contain newspaper text. All sentences no longer than 20 words from De Volkskrant and the NRC 1994 corpora were fully parsed with the Alpino parser (Bouma et al., 2001).

Once the best parse has been found, the parser reports for each word token the PoS-tag and lemma previously proposed by the PoS-tagger and lexical analyzer. In addition, the positional information is added. Given an input sentence like (11),

\[(11) \text{Groenink heeft in de maanden voorafgaande aan zijn benoeming}\]
\[\text{Groenink has in the months prior to his appointment}\]
\[\text{de strategie van de bank tegen het licht gehouden.}\]
\[\text{the strategy of the bank against the light held}\]
\[\text{`In the months prior to his appointment, Groenink has investigated}\]
\[\text{the strategy of the bank.'}\]

the output of the parser is the one shown in figure 5.1.
Figure 5.1: Sentence annotated with part-of-speech, positional information, lemmas and phrasal chunks. FRAME lines specify word information and CHUNK lines provide phrasal information.
5.2. Candidate extraction from corpora

Lines beginning with FRAME display information about each word in the input sentence. The two digits provide begin and end position. After the digits, the surface form of a word is followed by the word’s PoS-tag and the word’s lemma.

The parser returns a full parse for each input sentence. However, it is not guaranteed that the proposed analysis is the correct one. As a matter of fact, the PP tegen het licht ‘(lit.) against the light’ in (11) has been misparsed and treated as a modifier of the preceding noun bank. Figure 5.2 provides the output tree showing the syntactic structure and dependency relations proposed by the parser. The parser took the string de strategie van de bank tegen het licht as the NP object (obj1) of the verb houden failing to recognize that de strategie van de bank and tegen het licht are two separate syntactic dependents of the light verb houden in iets tegen het licht houden ‘to investigate something’. Clearly, it is expected that such errors occur due to the difficulty in resolving PP–attachment site ambiguity.

We try to avoid building on parser errors to the extent that is possible, therefore, we do not rely on the verb-complement dependencies or noun-complement dependencies proposed by the parser. Thus, from a fully parsed sentence, we use the phrasal boundaries proposed by the parser to ‘bracket’ and label noun phrases and prepositional phrases. As an illustration, the phrasal chunks shown in figure 5.3 were identified in sentence (11).

In figure 5.3, there are four PP CHUNKS. Among these, the PP tegen het licht is located between positions 14 and 17 and its head is located between positions 14 and 15. From the chunks above we know the location of the NP object (between positions 15 and 17) of the preposition tegen and the head noun in the NP.

Note that parsing is fully automatized and no error correction was done. This may be a source of noise to the extent that some verbs or PPs may be badly tagged or chunked. As a consequence, there may be candidates in the corpus that are missing in the dataset; alternatively, there could be spurious candidates in the dataset that were not verb PP constructions in the corpus but were wrongly analyzed by the parser. In any case, parser errors of this type are not very frequent; furthermore, we hope that the statistical measures are robust enough to penalize spurious candidates with a low score.

Once the extraction corpus is in the format shown in figure 5.1, candidate expressions can be collected to build the dataset.

5.2.2 Building datasets

Candidate expressions are the result of combining each verb with each co-occurring PP in the sentence. Candidates are represented as triples made up
Chapter 5. Identification of Support Verb Constructions

Figure 5.2: The parsed tree displays the syntactic structure and dependency relations within the verbal cluster in the sentence (11). The root *houd* ‘hold’, the head (hd) of the verbal cluster (vc), selects for an object (obj1). The PP constituent *tegen het licht* was misanalyzed as a post-modifier of the noun *bank*. 
5.2. Candidate extraction from corpora

Figure 5.3: NP and PP chunks found by the parser in sentence (11) above.

of the verb’s lemma, the preposition and the head noun within the NP. An abstract triple would be (verb lemma, preposition, noun). Figure 5.4 shows the triples extracted from the sentence in (11) above.

houd in maanden  
(hit. ‘hold in months’)

houd voorafgaande aan benoeming  
(hit. ‘hold prior to appointment’)

houd van bank  
(hit. ‘hold of bank’)

houd tegen licht  
(hit. ‘hold against light’)

Figure 5.4: Candidate triples extracted from the sentence in (11).

Whereas the verb’s surface forms have been reduced to the verb’s lemma, the noun maintains its surface form (without discarding inflectional information). The various surface forms of the verb correspond to different tenses; tense varies in support verb constructions, thus by lemmatizing the verb no crucial information is lost.\footnote{In proverbs and sayings, the verb typically exhibits one tense; in this case, tense information might be relevant in automatic identification tasks.} In contrast, if we reduce the noun to its lemma, different uses (figurative and literal) of the noun are clustered in one. By maintaining the surface form, the risk of clustering a figurative and a literal use is smaller. In preliminary experiments, the noun inside the PP was represented with its lemma. The results were worse.

Some verb occurrences are ignored while collecting candidates. Auxiliary verbs like zijn ‘to be’, hebben ‘to have’, worden ‘become’, etc. typically add tense, aspect, voice or modality to the predicate they head. If the parser decides that a verb is used as an auxiliary verb, we ignore it. The detailed
information contained in the PoS-tags allows us identify the auxiliary uses of a verb.\footnote{Occurrences of hebben ‘to have’ used as a light verb are not discarded.}

In these experiments, we treated any PP occurring within sentence boundaries as equally likely to realize a complement of any verb in the same sentence. Therefore, adjunct PPs may be tallied with totally unrelated verbs. We are aware of the noise we are introducing into datasets, but we hope that the statistical tests will be able to discard the noise.

In section 5.1 we mentioned that in some support verb constructions the noun within the PP may allow quantification and restricted modification. We decided to leave out prenominal determiners and modifiers for 2 reasons: to shorten the tuple’s length and to avoid the sparse data problem.

A total of 886,000 triples was collected from De Volkskrant corpus; among these, over 5,500 different types with the verb houden ‘to hold, keep’. More than 1.5 million triples were extracted from the two corpora. These include 9,300 triples involving houden. Hapax legomena (i.e. candidates with one occurrence in the dataset) account for 1.3 million triples in the bigger dataset. Dislegomena (candidates with two occurrences) add up to 112,000. The figures indicate that the collocation candidates are rather infrequent even in a corpus of ca. 32 million words. Table 5.1 reports the distribution of triples with the verb houden and with any verb (all data) in 3 cases: any frequency, dislegomena and hapax legomena.

<table>
<thead>
<tr>
<th>frequency</th>
<th>dataset</th>
<th>houden</th>
<th>all data</th>
</tr>
</thead>
<tbody>
<tr>
<td>any</td>
<td></td>
<td>5,569</td>
<td>886,000</td>
</tr>
<tr>
<td>2 occurrences</td>
<td>volks</td>
<td>453</td>
<td>55,900</td>
</tr>
<tr>
<td>1 occurrence</td>
<td></td>
<td>4,800</td>
<td>789,700</td>
</tr>
<tr>
<td>any</td>
<td></td>
<td>9,313</td>
<td>1,515,000</td>
</tr>
<tr>
<td>2 occurrences</td>
<td>volks nrc</td>
<td>874</td>
<td>112,340</td>
</tr>
<tr>
<td>1 occurrence</td>
<td></td>
<td>7,750</td>
<td>1,320,540</td>
</tr>
</tbody>
</table>

Table 5.1: Triple types distribution. Two datasets were extracted, one from the De Volkskrant corpus (volks) and one from the two corpora (volks nrc). The column houden enumerates the number of triples headed by the verb houden in the specific dataset. The column all data shows the number of triples headed by any verb. Rows provide the number of types with a given frequency in each dataset.
5.3 Identification process and evaluation

Once all candidate triples have been collected, identification proceeds by applying various statistical metrics. Two well-known problems in collocation identification using association measures receive special attention: the length of candidate expressions and the low-frequency data. To overcome the fact that candidates are triples (length 3), we treat them as bigrams. Whereas association measures are applied on bigrams, the loglinear model may be applied directly to triples (i.e. the original candidates in datasets). To control for the frequency bias, we investigate whether there exists a model or association measure that retrieves a reasonable list of collocations without the need to impose a frequency cutoff. Since our aim is to identify the best statistical model, we describe our criteria and evaluation methodology for selecting the best model.

5.3.1 Bigram model

The bigram model set-up is similar to that used in identifying collocational prepositional phrases described in section 4.4.1. When the expressions \( (w_1, w_2, w_3) \) are treated as bigrams, a candidate may be represented by any of the following bigrams:

- \( (w_1, w_2) \) e.g. (houd, in, gaten)
- \( (w_1, w_3) \) e.g. (houd, in, gaten)
- \( (w_2, w_3) \) e.g. (houd, gaten, in)

Krenn (2000b) and Evert and Krenn (2001) represent candidates as bigrams in the task of identifying PP VERB support verb constructions and figurative expressions in German. In their work, each candidate is represented as the bigram \( (w_1, w_2, w_3) \).

In preliminary experiments, only the bigram \( (w_1, w_2, w_3) \) had been used, however the results were worse. Our work departs from Evert and Krenn’s in that we use two bigrams \( ((w_1, w_2, w_3) \) and \( (w_1, w_2, w_3) \)) to estimate the association score of the component words in the candidate. Only the first two bigrams are taken into account assuming that a dependency ordering underlies the formation of triples. We also carried out experiments that included the third bigram; however, adding the third bigram did not help either.

Two partial associations are taken into account, the association between the verb and the PP and, the association between the tuple \( (verb, preposition) \)
Chapter 5. Identification of Support Verb Constructions

and the noun. If a strong lexical association exists between the lexemes in the triple the association score corresponding to the two bigrams should be high. On the other hand, if only partial lexical associations hold between the lexemes, the score of one of the bigrams stands out. Refer to section 4.4.1 for details on computing the triple association score.

5.3.2 Imposing a frequency cutoff

Low-frequency counts make the statistical scores unreliable. Because of this, some AMs like mutual information and $\chi^2$ perform badly with low-frequency data (Dunning, 1993; Evert and Krenn, 2001). The low-frequency bias has been tackled by applying either a cutoff that discards those candidates whose frequency is below a given threshold or alternatively, a cutoff that disqualifies those candidates whose association score is below an empirically established threshold. Both techniques pose problems because such thresholds may vary with the type and the size of the corpus, and also with the association measure (Dias et al., 1999; Kaalep and Muischnek, 2003); due to these problems, these thresholds are not reliable or portable across datasets and association measures.

More concretely, the use of a frequency cutoff can have a strong impact on the coverage of a statistical test. Applying a cutoff is likely to discard infrequent collocations and fixed expressions existing in the candidates data-set. Before we apply the statistical tests and select the cutoff, we ignore the impact the cutoff will have on the results. This makes the use of cutoffs or thresholds undesirable.

A look at table 5.1 shows that 94% of the triples in the dataset are hapaxes and dislegomena. What is really needed is a technique that can be successful with low-frequency data, not only from a practical point of view (automatic language acquisition) but also from a theoretical perspective (language acquisition). In the absence of such technique, we want to identify which test is less sensitive to low-frequency data and find out whether this test is good enough without the need to empirically determine an extra parameter. Section 5.4.2 further discusses the advantages and disadvantages of using a cutoff.

5.3.3 Evaluation methodology

Two decisions need to be made before we can identify the best statistic: (i) how to decide what a valid support verb construction is and, (ii) how to decide what the best model or metric is. To answer (i) a manually compiled
gold standard list is used. To answer (ii), we measure which statistic ranks more svc s (also in our gold standard list) among the top scores.

**Validation data** The lack of a well established definition of support verb construction and consequently the lack of an exhaustive list of collocations of the type we seek to extract make the evaluation of collocation identification models difficult, albeit interesting.

We collected a list of phrases with the verb *houden* ‘to hold’ from a monolingual dictionary (Geerts and Heestermans, 1992). Such phrases were found under the verb’s entry itself or under the head noun object of the preposition. The list was further enlarged with other uses of *houden* consisting of a pp complement which are frequently found inside the verb cluster in subordinate sentences. The resulting list was given to three linguists familiar with collocational phenomena. The three informants were asked to mark those expressions which should be considered lexicalized into a support verb construction and that should be entered in a dictionary. No further instructions were given.

Our three informants significantly disagree in their judgments. Only those expressions which were marked as good by at least two people were added into the validation list. From a list of 120 candidates, a total of 80 expressions were identified as support verb constructions with the verb *houden*. This non-exhaustive list constitutes our validation data. Appendix B displays this list.

Some validation candidates are missing in the candidate datasets; these valid expressions occur in the raw corpus but they appear in sentences longer than 20 words, therefore, the sentences were excluded during preprocessing. The number of true positives present in the original datasets is, 61 in the *volks* dataset and 63 in the *volks nrc* one. Roughly 25% of validation data is not found in datasets.

**Nbest lists** The expressions ranked according to each association measure, raw frequency and the loglinear model are compared to the validation data. Among the ranked expressions we select the top scores associated with $n$ expressions and build what is known as an $n$best list. The $n$best lists are compared to our validation data to determine which test performs best in identifying collocations. It is unclear how big the $n$best list should be. We chose $n$best = 500 since usually, few true positives (with the verb *houden*) are found after 500 expressions.

---

3This list was kindly provided by Jack Hoeksema.

4For the latest version of *Alpino*, such sentences are no longer problematic.
Evaluation measures  Since we lack an exhaustive list of the true positives included in the datasets, standard precision and recall cannot be directly measured. To be able to compare the different metrics we still need a figure that reflects how efficient a test is in identifying truly lexicalized candidates and penalizing false candidates. We compute the coverage and the uninterpolated average precision.

Coverage gives the number of true positives found in the nbest list. Examining the coverage of the complete output list (i.e. the list of candidates ranked according to each statistic) is uninteresting because all statistics eventually identify all the svc's common to the candidate dataset and our gold standard list. Instead, the coverage at varied $N$ levels throughout the nbest list allows us to assess which statistic identifies more svc's (Evert and Krenn, 2001). Coverage values can be read from the ‘accuracy graphs’ that we introduce below.

Uninterpolated average precision is often used in information retrieval, where evaluation measures determine which system manages to rank relevant documents before non-relevant ones (Manning and Schütze, 1999, p. 535). We adopt this metric for our task. Given an input list of candidate collocations $L$ such that $c_1, c_2, c_3, ..., c_i \in L$ where $L$ has size $N$, we want to determine for several values of $N$, $n_i$, $n_j$, $n_k$, ..., $n_z$ which statistic extracted more true positives overall. **Uninterpolated average precision** aggregates many precision numbers into one evaluation figure. At each point where a true positive in the retrieved list is found, the precision is computed and, all precision points are then averaged. Retrieved candidates that are false hits have an indirect effect on the **uninterpolated average precision**.

Table 5.2 displays two hypothetical nbest lists with 5 candidate expressions each. Within the nbest lists, 1 denotes a true positive and 0 a false hit. To compute the standard precision one just divides the number of true positives by the number of candidates, thus, $3/5$ gives a precision of 0.6% for both nbest lists. To compute **uninterpolated average precision** we add up all precision points and compute the average; thus, $(1/1 + 2/3 + 3/4)/3$ gives us 0.8 **uap** value for list 1 and $(1/2 + 2/4 + 3/5)/3$ gives 0.53 **uap** value for list 2. The **uap** value suggests that the list 1 is more accurate than the second one; in fact, list 1 allows fewer false hits among the top scores, thus providing a more interesting list. Whereas the standard precision suggests that both rankings are equally accurate, the **uninterpolated average precision** clearly suggests that list 1 is more interesting.

As far as it is known to us, there is no unique criterion that says how large the nbest list should be. Using **uninterpolated average precision**, we circumvent the problem of where to set the nbest value. In computing the **uninterpolated average precision**, all true positives down the extracted list have to be taken
5.4. Quantitative results

<table>
<thead>
<tr>
<th></th>
<th>list 1</th>
<th>list 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>uap</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 5.2: Computing standard precision and uninterpolated average precision (uap) for a hypothetical ranked list with 5 candidates.

into account. Indirectly, uap measures an ‘approximated’ recall also. Hence, we rely on this measure to identify the best statistic.

**Accuracy graphs** Such graphs plot the trend line that displays the number of true positives (y axis) found while traversing the nbest list (x axis). All graphs include the maximum accuracy line. Maximum accuracy is reached if all higher ranked expressions in the nbest list are valid collocations (present in our validation data). If a test reached maximum accuracy, its corresponding accuracy line should overlap with the grey line (labelled ‘maximum’).

5.4 Quantitative results

This section characterizes the results of the different tests paying attention to the accuracy, coverage and uninterpolated average precision. The association measures, raw frequency (our baseline) and the loglinear model are applied to the datasets described in section 5.2, where one of them (volks) is a subset of the other (volks nrc). First, we report the results of the metrics without any additional parameter. Next, we explore the advantages and disadvantages of adding a frequency cutoff. This issue is followed by some remarks on the performance of the loglinear model. We also assess if the results of the models in identifying support verb constructions with one verb resemble the results for other light verbs. Finally, we add some remarks related to evaluation.
Chapter 5. Identification of Support Verb Constructions

5.4.1 Bare models

The different tests applied on the smaller dataset (volks) display the accuracy lines depicted in figure 5.5.

Figure 5.5: Raw frequency and various AMs applied to the dataset volks. The y-axis denotes the number of support verb constructions correctly found among the first 500 candidates (x-axis).

Raw frequency sets the baseline. Three tests perform poorly: mutual information, left Fisher and loglinear. Loglinear performs the worst because the model retrieves very few expressions with houden. We leave out the loglinear accuracy line. Mutual information and left Fisher rank hapax legomena among the top scores and these turn out to be mostly false positives. 5% of the top scores ranked by $\chi^2$ are true positives; accuracy improves very little throughout the rest of the retrieved nbest list. Salience and log likelihood ($G^2$ from here onwards) reach higher accuracy (than the baseline) through
the top 30% of the nbest list. Between the interval 30–70% (approx.) of the nbest list, raw frequency reaches better accuracy. Towards the end salience finds a few more true positives than raw frequency.

Table 5.3 gives the number of true positives (tps) retrieved by each statistic (measured at nbest=500) and the uninterpolated average precision (uap) values. Raw frequency and salience retrieve a similar number of true positives. $G^2$ and $\chi^2$ reach lower coverage. The corresponding uninterpolated average precision values in the right column show that salience reaches the best uap value closely followed by $G^2$.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Statistic</th>
<th>tps</th>
<th>uap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BASELINE</td>
<td>48</td>
<td>0.52</td>
</tr>
<tr>
<td>volks</td>
<td>mi</td>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>36</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>$G^2$</td>
<td>42</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>leftFisher</td>
<td>21</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>salience</td>
<td>49</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>loglinear</td>
<td>1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 5.3: Number of true positives (tps) and uninterpolated average precision value for raw frequency (BASELINE), the various AMS and the loglinear model.

In terms of accuracy and coverage, salience and raw frequency are the best tests (based on the accuracy lines in figure 5.5). The uninterpolated average precision values shown in Table 5.3 suggest that either salience or $G^2$ are good tests.

### 5.4.2 Adding a frequency cutoff: pros and cons

Since our goal is to identify the most parsimonious test that retrieves the best list of collocations, we explore whether a frequency cutoff is desirable. We focus on tests from here onwards: salience and loglinear (the best and the worst tests in our previous experiments). We set to determine what cutoff is required by each test to reach best accuracy, coverage and uap. Figures 5.6 and 5.7 illustrate the effect of varying the cutoff from 1 through 6.

Observed in figure 5.6, setting the cutoff to 2 makes no difference in the accuracy line of salience (see overlapping dotted and dotdashed lines). A cutoff of 3 improves the accuracy line (line marked with an inverted triangle) but the coverage decreases slightly. In fact, a cutoff of 3 causes a shortening of the nbest list, since only 300 candidates (with houden) reach that cutoff.
Figure 5.6: Effect of increasing frequency cutoff: salience. Each line corresponds to a frequency cutoff $n$ (e.g. $f_1$ applies a cutoff of 1). Maximum displays the maximum accuracy possible.
5.4. Quantitative results

Figure 5.7: Effect of increasing frequency cutoff: loglinear model. Each line corresponds to a frequency cutoff $n$. Maximum displays the maximum accuracy possible.
Thus, a decision to leave out hapaxes and dislegomena discards expressions which are indeed true positives. Raising the cutoff causes substantial improvements in the loglinear model also. With a cutoff of 3, loglinear reaches the best coverage (figure 5.7); with a cutoff of 4, one achieves the best compromise between accuracy and coverage.

Figures 5.6 and 5.7 illustrate that a frequency cutoff is a parameter that needs to be chosen specifically for each statistic or model. The issue is in fact even more complicated, since the cutoff needs to be empirically selected on the basis of the dataset size. Evidence of this is given in figures 5.8 and 5.9.

![Figure 5.8: Increasing dataset size: readjustment of cutoff for salience.](image)

When a larger dataset is available, salience reaches the best accuracy when a cutoff of 4 is applied; however note that the accuracy line with a cutoff of 3 reaches higher coverage (figure 5.8). A cutoff set to 5 ensures the best accuracy and coverage for the loglinear model (figure 5.9).
Figure 5.9: Increasing dataset size: readjustment of cutoff for loglinear model.
In collocation extraction, most collocations are spread among low-frequency ranges, exactly the frequency ranges most likely to be discarded by the cutoffs. Leaving out low-frequency data makes the statistical scores more reliable. Thus, one expects that the accuracy of the identification models should improve. However, given that we lack an exhaustive list of svc's, by applying a cutoff we would rule out the possibility of finding new svc's (absent in our validation list). Further, the cutoff needs to be specifically tuned for the chosen statistic and it is also dependent on dataset size (and indirectly on extraction corpus size). We conclude that the use of a cutoff is undesirable in identification tasks that aim at expanding an existing list of (any sort of) collocations.

One question that emerges is: Is there a test that retrieves a reasonable and accurate list of collocations and, that does not require a cutoff?

Let us return to the original dataset (volks) and select the cutoff that allows each test to reach the best accuracy line and uap value. We compare the most parsimonious test (raw frequency) with the best and the worst two tests. First, the ‘optimal’ cutoffs are computed: salience (3), $G^2$ (3), mutual information (4) and loglinear (3). Table 5.4 provides the uninterpolated average precision values. The $G^2$ and salience are almost undistinguishable but slightly better than the $G^2$ and salience and significantly better than that of the loglinear model.

<table>
<thead>
<tr>
<th></th>
<th>mi</th>
<th>$G^2$</th>
<th>salience</th>
<th>loglinear</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>cutoff</td>
<td>0.805</td>
<td>0.809</td>
<td>0.816</td>
<td>0.56</td>
<td>0.52</td>
</tr>
<tr>
<td>without cutoff</td>
<td>0.01</td>
<td>0.60</td>
<td>0.63</td>
<td>0.05</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 5.4: Uninterpolated average precision values.

If a test reaches good accuracy, coverage and uap value without the need to set a cutoff, this test has a clear advantage over other tests that clearly require a cutoff. This is the case with the salience test. Figure 5.10 displays the accuracy lines of mutual information and the loglinear model (with a cutoff) and the accuracy lines of salience, $G^2$ and raw frequency (without a cutoff). The results are appealing. $G^2$ reaches poorer coverage than the other tests, but salience reveals a good accuracy line, higher coverage than the other tests and a reasonable uap value (refer to second row table 5.4).

With the evidence given, we conclude that a test that is not dependent on the adjustment of a frequency cutoff is more interesting and useful than any other tests, since it is exploits the available data more extensively.
5.4. Quantitative results

Figure 5.10: Comparison between raw frequency, salience, $G^2$, mutual information and loglinear. No cutoff applied to salience nor $G^2$. 
5.4.3 A few remarks on the loglinear model’s performance

The main advantage of Blaheta and Johnson’s loglinear model (described in section 3.4) is that it generalizes to ntuples, that is, candidate expressions of any length. A second advantage is that it (supposedly) handles the low-frequency bias. To realize these advantages, two parameters need to be set: a frequency cutoff and a significance level. Although this model requires the tuning of these two parameters, we still considered it an attractive alternative because our candidate ntuples did not need special treatment.

Suppose the cutoff 3 gives us best accuracy and coverage. Would accuracy and coverage further improve if we now tune the significance level? If we isolate the accuracy lines for two different cutoffs and two different significance levels we get the graph in figure 5.11. Compare the lines labelled “f3 sf=0.1” and “f3 sf=0.9”. A significance level of 0.9 is the most permissive because it admits all association scores that chance would account for 90% of the time. A change in significance level slightly diminishes accuracy but it has no apparent effect on coverage. Now compare the lines labelled “f4 sf=0.1” and “f4 sf=0.9”, using a cutoff of 4. The effect of the significance level is positive, although the difference between the accuracy lines is not as noticeable as with cutoff 3. This suggests that after an optimal cutoff is found, the effect of the significance level is less manifest. In addition, the improvement in accuracy is accompanied by a decrease in coverage but, this is mainly caused by the cutoff increase.

The graph supplies evidence that a cutoff of 4 and significance level of 0.1 produce the best compromise between accuracy and coverage (see dotdashed line). This best accuracy line, the same as is displayed in figure 5.10, is still far from giving results comparable to those of some association measures, namely salience and $G^2$.

5.4.4 Experiment’s validation: other support verbs

Salience and $G^2$ (log-likelihood) are the two tests that reach best $uap$ in the extraction of support verb constructions with *houden*. An identification model to expand a lexicon of e.g. support verb constructions, ought to be efficient for any type of collocation, independently of the lexemes involved. In order to assess the applicability of our findings, we expand the task to identify support verb constructions headed by one of the following verbs: *hebben* ‘have’, *maken* ‘make’ and *doen* ‘do’.
Figure 5.11: Tuning significance level for loglinear model. Small dataset.
**Ntuples distribution**  In the extraction corpora we used, there are more different candidate expressions headed by *hebben*, *maken* and *doen* than there are with *houden*. Table 5.5 shows the number of triple types headed by each of the four verbs. The number of candidates with *hebben* is 4 times the number of expressions with *houden*. A similar ratio 1:4 is observed for dislegomena and hapax legomena. Candidates with *doen* and *maken* double the number of candidates with *houden*.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frequency</th>
<th>houden</th>
<th>doen</th>
<th>maken</th>
<th>hebben</th>
</tr>
</thead>
<tbody>
<tr>
<td>any</td>
<td></td>
<td>5569</td>
<td>10057</td>
<td>12660</td>
<td>22140</td>
</tr>
<tr>
<td>volks</td>
<td>dislegomena</td>
<td>453</td>
<td>893</td>
<td>1093</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>hapax legomena</td>
<td>4800</td>
<td>8535</td>
<td>10675</td>
<td>17945</td>
</tr>
</tbody>
</table>

Table 5.5: Distribution of all triple types consisting of the support verb *houden*, *maken*, *doen* or *hebben* in the *Volkskrant* corpus.

**Validation data**  We followed the procedure described in section 5.3.3 to compile the validation data. In our validation data, there are very few prepositional support verb constructions with *doen* and *maken* because these verbs collocate mostly with NPs but rarely with PPs. *Hebben* seems to collocate with PPs very often. Tables A.3 to A.2 in appendix A provide the lists of validation data with *houden*, *hebben*, *maken* and *doen*. We stress that these validation lists are not exhaustive. Table 5.6 displays the number of expressions in our validation data found in the candidate dataset *volks*.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>True positives with</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>houden</td>
</tr>
<tr>
<td>volks</td>
<td>61</td>
</tr>
</tbody>
</table>

Table 5.6: Number of true positives found in the dataset *volks*.

**Results**  First we report how the tests perform in identifying other support verbs. Next, we focus on *hebben* and compare the accuracy and coverage of the different statistics. Without a cutoff, mutual information, left Fisher and the loglinear model perform poorly in the identification of *houden* and the same is seen in the identification of the other verbs. We limit the following description to raw frequency, salience and $G^2$. 
5.4. Quantitative results

As in the previous experiments with *houden*, raw frequency sets the baseline. Table 5.7 provides the coverage and *uninterpolated average precision* values of the tests.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Statistic</th>
<th>houden</th>
<th>hebben</th>
<th>maken</th>
<th>doen</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BASELINE</td>
<td>48 (0.52)</td>
<td>42 (0.13)</td>
<td>6 (0.07)</td>
<td>10 (0.13)</td>
</tr>
<tr>
<td>volks</td>
<td>$G^2$</td>
<td>42 (0.60)</td>
<td>46 (0.27)</td>
<td>4 (0.19)</td>
<td>6 (0.39)</td>
</tr>
<tr>
<td>salience</td>
<td></td>
<td>49 (0.63)</td>
<td>52 (0.29)</td>
<td>6 (0.20)</td>
<td>9 (0.40)</td>
</tr>
</tbody>
</table>

Table 5.7: Identification of 4 different support verbs: true positives and *uninterpolated average precision* values (in parentheses) corresponding to raw frequency (BASELINE) and $G^2$, salience. The number of true positives corresponds to nbest = 500.

There are more true positives with *hebben* than with *houden* in the dataset. Among the top 500 expressions, raw frequency identifies expressions with *houden* better than expressions with *hebben*. $G^2$ and salience differ; these two tests find more true positives with *hebben* than with *houden*. This shows that $G^2$ and salience discard noisy candidates that happen to be very frequent, whereas raw frequency promotes them to the top scores.

Concerning the other two verbs, there are very few SVCs common to the validation data and the dataset (8 with *maken* and 10 with *doen*). This makes the identification task harder for all three tests. The tests identify almost all true positives with *maken* and *doen* among the top 500 scores.

The $uap$ values show that it is easier for all tests to identify collocations with *houden* than collocations with *hebben*, *doen* or *maken*. On the basis of a comparison of the $uap$ values of all three tests, we have further evidence to claim that salience and $G^2$ are more suitable than raw frequency (and the other tests) for extraction of support verb constructions.

5.4.5 Discussion

This section provides a few remarks about the tests’ performance followed by some comments on the evaluation methodology.

**On tests’ performance**

Raw frequency, salience and $G^2$ better identify support verb constructions with *houden* than with the more frequent support verbs *hebben*, *doen* and *maken* in our experiments. This may be due to the facts that (i) *houden* occurs less frequently than the other verbs, (ii) *houden* is less polysemous
than the other verbs and (iii), *maken* and *doen* more often collocate with NPs, therefore, the few verb PP collocations stand out among the candidates in the dataset.

If we were to use only raw frequency to identify support verb constructions that ought to be added in a computational lexicon or dictionary, many false positives would have to be manually discarded. In contrast, salience or $G^2$ produce nbest lists with higher precision; this implies that a computational lexicographer would have to spend less time in checking nbest lists to retrieve the same amount of support verb constructions.

Neither $\chi^2$ nor $G^2$ is as sensitive to low-frequency data as mutual information is. In order to accomplish a reliable classification when we apply $\chi^2$ and $G^2$, a large sample size is desirable, that is, the expected frequency of each cell should be equal or greater than 5 (Agresti, 2002). In our data, many bigrams exhibit an expected frequency smaller than 5; this hurts $\chi^2$ more than $G^2$.

**On evaluation methodology**

For evaluation, nbest lists of candidates ranked according to the different tests are selected. The size of the nbest list is arbitrary but it is crucial. To gather reliable results, coverage and accuracy should be provided for several nbest sizes, preferably large lists. Attempting extraction of “adjective noun” collocations in German, Evert and Krenn (2001) report results about two nbest lists, size 100 and 500. In this research we chose nbest=500, which we consider large enough for examining candidates with a unique verb.

Evaluation requires a boolean decision: either an expression in the nbest list is a valid collocation (true positive, thus present in our validation data) or it is not (false positive). Given our limited validation data, it may happen that some expression in the nbest list is regarded as a false positive because it is not in our validation data, even though the expression is a perfectly legitimate collocation. To carry out a reliable assessment of the accuracy and coverage of the tests, we also perform qualitative evaluation of top scores inside the nbest lists. This is described in Section 5.5.

We argue that in order to measure how successful a statistic or model is, given our limited validation data, the *uninterpolated average precision* is an appropriate measure. If we only compare standard precision achieved by the tests, the difference between the statistics would not be so obvious (refer to precision values in Table 5.3). Note that several tests identify a similar number of true positives, therefore their precision values are identical. *uap* is a more reliable measure than precision because it tells us which test is more accurate in identifying good collocations and indirectly measures *recall*. A
high \textit{uap} value is obtained if the top scores in the nbest list include mostly true positives and few false positives occur in between them. Nevertheless, caution in interpreting the \textit{uap} is needed. Due to the use of a cutoff or a strict significance level, a test may return a small nbest list that shows high accuracy. If this nbest list is compared to a larger nbest list with a reasonable accuracy and better coverage, the \textit{uap} value for the smaller nbest list will be higher leading us to inaccurate conclusions. Thus, we recommend that one use the \textit{uap} values together with the accuracy graphs.

There is a final issue that needs discussion since it may influence the results we obtained. On one hand, when the association measures are applied to a dataset, the resulting list displays the same set of expressions originally present in the dataset but re-ordered on the basis of the AM score they were assigned. The resulting list includes exactly the same number of expressions as there were in the candidate dataset. From this resulting list an nbest list of 500 expressions with a specific verb is extracted for evaluation. On the other hand, when the loglinear model is applied to the same candidate dataset, only those expressions for which the loglinear model estimates a significant association score are retrieved. Even if we set the most permisive significance level and a low frequency cutoff value, the loglinear model still retrieves shorter output lists than the AMs. Aware of this difference, one could conclude that the comparison between the AMs and the loglinear model is unfair. To establish similar evaluation settings we need to apply a test of significance to the scores proposed by the AMs. Significance is assessed by looking up the p-value assigned to the statistic score in a table. As far as we know, this is possible for tests like $\chi^2$ or $G^2$ for which a $p$-value specifies whether the score is significant enough to establish that the candidate is a collocation. Thus, one should apply a $p$-value to the scores proposed by the AMs ($\chi^2$, $G^2$) and discard those expressions whose scores have a $p$-value above a certain threshold.

\section{Qualitative results}

In this section we inspect what sort of expressions cause a divergence between the accuracy line of salience and that of $G^2$. We also examine why not all true positives existing in the dataset are ranked among the 500 best scored expressions. Finally, we check if all expressions classified as false positives are indeed false positives.
5.5.1 Qualitative divergence between best tests

To answer the first question, we compared the 100 highest ranked expressions (with *houden*) retrieved by the two tests to determine whether a qualitative difference emerges. We observe that a total of 7 true positives included in the salience nbest list are not included in the $G^2_{\text{nbest}}$ list. Refer to Table 5.8.

<table>
<thead>
<tr>
<th>Expression</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>iemand aan zijn woord houden</td>
<td>‘to fulfil a promise’</td>
</tr>
<tr>
<td>iets in eigen huis houden</td>
<td>‘to host an event’</td>
</tr>
<tr>
<td>iets in leven houden</td>
<td>‘keep sth going/alive’</td>
</tr>
<tr>
<td>in het midden houden</td>
<td>‘leave something unclear’</td>
</tr>
<tr>
<td>in pand houden</td>
<td>‘keep in custody’</td>
</tr>
<tr>
<td>op de rails houden</td>
<td>‘keep sth going’</td>
</tr>
<tr>
<td>op zak houden</td>
<td>‘keep sth as a possession; save’</td>
</tr>
<tr>
<td>iets voor ogen houden</td>
<td>‘not to forget about sth’</td>
</tr>
</tbody>
</table>

Table 5.8: True positives among 100 highest ranked expressions only retrieved by salience.

$G^2$ assigns the seven expressions in this table a lower score; consequently these 7 expressions are not even included among the 500 nbest expressions. Low-frequency candidates consisting of rather frequent prepositions and nouns tend to be ranked low by the $G^2$ (e.g. *voor ogen houden*, *in leven houden*).

Overall, the salience and the $G^2$ tests retrieved exactly 49 and 42 true positives, respectively (nbest=500). The remaining 11 true positives present in the dataset but not classified among the 500 best scores by salience are given in Table 5.9.

These expressions together with those in Table 5.8 are not among the 500 highest scores proposed by $G^2$. The tests assign a low rank to these expressions, therefore they fall outside the chosen nbest list.

5.5.2 Error analysis

Evaluation of the false positives reveals that many expressions are combinations of the verb *houden* with a regular prepositional complement. Examples are *(zich) aan NP houden* ‘to adhere to NP’ and *van NP houden* ‘to love NP’.

Among the prepositions that introduce this type of error are *aan* ‘on, to’, *van* ‘of’ and *met* ‘with’. Errors that show a prepositional complement with *met* ‘with’ are most likely instances of the support verb construction *rekening houden met* ‘take into account’. The lexicalized material is the NP *rekening*.
5.5. Qualitative results

<table>
<thead>
<tr>
<th>Expression</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>in zijn bezit houden</td>
<td>‘to own’</td>
</tr>
<tr>
<td>in zijn gedachten houden</td>
<td>‘keep one’s thoughts on sth.’</td>
</tr>
<tr>
<td>het heft in handen houden</td>
<td>‘stay in control’</td>
</tr>
<tr>
<td>in petto houden</td>
<td>‘keet sth up one’s sleeve’</td>
</tr>
<tr>
<td>de bal in de ploeg houden</td>
<td>‘keep s.o. in your own group (metaph.)’</td>
</tr>
<tr>
<td>onder de pet houden</td>
<td>‘keep sth back’</td>
</tr>
<tr>
<td>zich op de achtergrond houden</td>
<td>‘keep a low profile’</td>
</tr>
<tr>
<td>op poten houden</td>
<td>‘keep alive’</td>
</tr>
<tr>
<td>handen op zijn rug houden</td>
<td>‘do not interfere’</td>
</tr>
<tr>
<td>op schoot houden</td>
<td>‘keep in your lap’</td>
</tr>
<tr>
<td>iemand voor de zot houden</td>
<td>‘consider s.o a fool’</td>
</tr>
</tbody>
</table>

Table 5.9: Items in the validation data assigned a very low rank by salience and $G^2$.

but not the PP. In the above cases, the statistical tests correctly infer an association between the prepositions aan, van or met and the verb houden, however, no lexical restriction applies on the type of noun that realizes the object NP.

Among other false positives, the tests include some expressions with houden where the PP itself is lexicalized and used as a predicative PP. An example is uit zicht ‘out of sight’. Borderline cases surface among the false positives. For instance, in bewaring houden ‘keep under custody’, in voorarrest houden ‘take into custody’, (iemand) in spanning houden ‘keep s.o. intrigued’. It’s not clear to us whether they are fully lexicalized constructions or a lexicalized PP that combines with a restricted set of verbs. Other sources of frequent errors are locative, temporal, directional PPs.

Nevertheless the associations proposed by the statistical tests are not always inappropriate. Among expressions labelled as false positives, three expressions are actual support verb constructions: aan zichzelf houden, op knip houden and van domme houden. These expressions form part of longer support verb constructions (of the same type as example (5)), namely de eer aan zichzelf houden ‘take the honourable way out’, de hand op de knip houden ‘keep a tight hand on one’s purse’ and zich van de domme houden ‘play ignorant/the innocent’. Expressions with similar structure were already included in our validation list, for example boven water houden which is part of the longer expression het hoofd boven water houden ‘keep one’s head above water’. Two other expressions retrieved by salience and $G^2$ that are true collocations but were not in our validation data are bij de les houden ‘pay full attention’ and in wurggreep houden ‘have a stranglehold on s.o.’.
in wurgreep houden is also a collocation, a variant of in greep houden ‘keep firmly in one’s grasp’ but with a stronger negative connotation.

5.5.3 Updating gold-standards

After completing the qualitative evaluation of $G^2$ and salience, we also examined more carefully the quality of our original validation list. Some expressions in the validation list that were not retrieved by any of the statistics are questionable. Although some of the judges proposed them as support verb constructions, either: (a) part of the expression is a predicative PP that may occur with other verbs as well (op de achtergrond houden, op poten houden, op de rails houden, voor de geest houden) (e.g. op de achtergrond stellen ‘put sth in the background’, op poten zetten ‘start up sth’, voor de geest halen ‘bring sth to mind’, or (b) the expression is not found in contemporary Dutch (te gang houden), therefore they cannot be in our corpora.

5.5.4 Discussion

Qualitative assessment of the results is very important because (1) relevant differences between statistical tests may emerge, and (2) this assessment gives insight into the accuracy of the validation data. We compared the results of two tests that seemed to behave very similarly. We showed that the $G^2$ test manages to find clear collocations among the hapax legomena; true positives among hapax legomena will never end up among the top scores in a salience nbest list. Because of this, we believe that if the candidates datasets are small, salience will fail to find collocations among hapaxes; thus, it is better to apply both tests and combine the retrieved lists, to ensure that the number of identified collocations is as large as possible.

Secondly, error analysis and assessment of missing true positives revealed two important issues; on one hand, the validation data needs to be recompiled since there are some candidates in that list that are not likely to be found in contemporary Dutch newspaper text. On the other hand, further refinements to the chosen statistics are needed to prevent predicative PPs from being ranked among the top scores, as well as instances of the verb houden and a preposition that introduces a typical complement of the verb.

5.6 Conclusions

Identifying support verb constructions in a language for which a large treebank is not available can be approached by performing the extraction of can-
didate datasets from corpora that has been automatically annotated with a robust parser; in our case, this parser was informed by a wide-coverage lexicalist grammar.

In our experiments, we lacked an exhaustive list of support verb constructions. Thus, we argue that a test that does not require setting a cutoff is preferred to a more complex model that depends on tuning extra parameters like a cutoff and significance level. If we do not know how many true positives exist in low-frequency strata, it is risky to use a cutoff because we may discard many collocations. The experiments done identified salience and log-likelihood chi square ratio as the tests that reach the best compromise between accuracy and coverage, in the given order.

Identification of collocational uses of highly frequent verbs like houden, hebben, doen, etc. cannot accurately be done with a test like raw frequency. Light verbs tend to be highly frequent verbs; a mere listing of a light verb with each co-occurring PP ranked on the basis of its frequency in the corpus finds many accidental collocations. The obvious gain of applying a more ‘sophisticated’ test like $G^2$ or salience is an increase in accuracy.
Chapter 5. Identification of Support Verb Constructions