Chapter 4

Identification of Collocational Prepositional Phrases

In this chapter, we describe an identification model that uses part-of-speech tagged extraction data. We evaluate the model on the task of extracting Dutch collocational prepositional phrases (CPPs) from a corpus. We describe joint work with Gosse Bouma. Part of the material in this chapter has been published as Bouma and Villada (2002) and Villada Moirón and Bouma (2002). Evaluation and results comprise original work.

4.1 Introduction

In this case study we explore whether a hybrid model that requires a part-of-speech tagged corpus and various statistics is feasible to identify collocational prepositional phrases.

Dutch has a number of preposition-(determiner-)noun-preposition (P-NP-P) combinations, which are more or less fixed (1). In Dutch linguistics such expressions are known as voorzetsel-uitdrukkingen (Paardekooper, 1962). Here, we will refer to them as collocational prepositional phrases (CPPs).


In a lexicalist grammar, one might analyze phrases introduced by these constructions, such as (2), as prepositional multi-word units, as shown in (3).
This is the analysis adopted by the annotation guidelines of the Corpus Gesproken Nederlands (Moortgat et al., 2001). It presupposes that these expressions behave as units syntactically, and can be listed in the lexicon.

(2) \textit{in tegenstelling tot eerdere berichten} \\
as opposed to earlier reports

(3) \[
P \quad PP \\
en tegenstelling tot \quad eerdere berichten
\]

In certain contexts, intervening adverbs (4) or other material (5) separate the P-NP from the second preposition.

(4) a. De teneur is dat Camus ‘niets aan actualiteit heeft’ the purport is that Camus ‘nothing on current affairs has ingeboet’, \textit{in tegenstelling misschien tot andere} filled up’, in contrast perhaps with other ‘existentialisten’ in het naoorlogse Parijs . . . ‘existentialists’ in the post-war Paris . . . ‘The purport is that Camus has not contributed anything to current affairs, as opposed to other existentialists in post-war Paris’

b. Economisch rekenen kunnen de ambtenaren al 47 jaar, economic figures can the officials already 47 years, \textit{in tegenstelling dus tot de dames en heren} . . . in contrast thus with the ladies and gentleman . . . ‘(lit.) . . .economic figures can the officials already 47 years, in contrast thus with the ladies and gentleman . . .’

c. \textit{met dank natuurlijk aan Shakespeare} . . . with thanks obviously to Shakespeare . . . ‘. . .thanks to Shakespeare, obviously’

(5) Het orkest zal \textit{onder leiding} staan van een Duitse dirigent. The orchestra shall under guidance stand of a German director ‘The orchestra will be directed by a German director.’

This evidence rules out an analysis like (3). An alternative is to consider only the initial P-NP combination as a lexicalized unit (6) that selects another phrase headed by the second preposition. This analysis still requires that such elements are listed in the lexicon, and that the P-NP combination is rigid, but it does allow the second PP to behave as a complement of the
4.2. Linguistic properties

initial P-NP expression.

(6)

\begin{center}
\begin{tikzpicture}
  \node (p) {PP} ;
  \node (pp) [below of = p, yshift = 0.7cm] {PP} ;
  \node (in) [left of = pp, xshift = -0.5cm] {in tegenstelling} ;
  \node (p1) [below of = p, yshift = 0.5cm] {P} ;
  \node (tot) [below of = p1, yshift = 0.5cm] {tot} ;
  \node (np) [right of = pp, xshift = 0.5cm] {NP} ;
  \node (eerdere berichten) [right of = np, xshift = 0.5cm] {eerdere berichten} ;
  \draw (p) -- (pp) ;
  \draw (p1) -- (tot) ;
  \draw (np) -- (eerdere berichten) ;
\end{tikzpicture}
\end{center}

A number of linguistic tests (described below in section 4.2) suggest that collocational prepositional phrases need to be distinguished from regular P NP P combinations. As we will explain, the tests also indicate that for some CPPs the analysis in (3) is appropriate, whereas for others the analysis in (6) is more likely. The analysis in (3) requires that P-NP-P combinations are listed as complex prepositions in the lexicon, whereas the analysis in (6) requires that P-NP combinations are listed as prepositional phrases selecting for a PP-complement headed by a specific preposition. In both cases the question arises whether it is in fact possible to list such expressions.

An exhaustive listing of CPPs does not exist and, given the amount of variation within the class of CPPs, it may not be easy to decide on a definite listing. Paardekooper (1973) and the *Algemene Nederlandse Spraakkunst* (ANS – General Dutch Grammar) (Haeseryn et al., 1997) provide a list of CPPs. Paardekoper’s list is a subset of the list of 86 items given in the ANS. To obtain a more complete listing, we therefore considered whether a corpus could be used to identify potential candidates. In particular, it seems that frequent P-NP-P patterns are likely to contain CPPs.

In this chapter we address the question of to what extent corpus-based methods can be used to obtain a more complete listing of CPPs. Before providing a characterization of the model, we enumerate the linguistic properties of CPPs and motivate a distinction between different CPPs on the one hand, and regular prepositional phrases on the other hand. Idiosyncrasies in the syntactic distribution highlight the fact that CPPs are lexicalized phrases. The second half of the chapter describes an identification model that extracts candidate data from a tagged corpus and measures the lexical affinity between the component words using standard association measures. Next, we assess the limitations of such model in identifying CPPs.

### 4.2 Linguistic properties

A number of linguistic properties set collocational prepositional phrases apart from syntactically and semantically regular PPs. Most of these properties...
were already observed by Paardekooper (1962, 1973).

4.2.1 Idiosyncratic features

Idiosyncratic prepositions and nouns. Many P-NP-P patterns are introduced by the preposition te and its dative and genitive forms ten (from te+den) and ter (from te+der). This preposition has a restricted, formal, usage and otherwise occurs in fixed expressions only. Dative and genitive markings only occur in fixed archaic expressions. Thus, the inflected forms of the nouns following ten in (7) below, do not occur outside fixed expressions.

(7) ten opzichte van ‘in comparison with’, ten tijde van ‘at the time of’, ten koste van ‘at the expense of’, ten gunste van ‘to the benefit of’, ten gevolge van ‘as a consequence of’, ten nadele van ‘at the expense of’

There are also CPPs which contain a noun that is seen only rarely outside the context of this fixed expression:

(8) aan de vooravond van ‘at the eve of’, in navolging van ‘following’, met behulp van ‘with the use of’, bij monde van ‘according to’, onder het mom van ‘under the pretext of’, in samenspraak met ‘in negotiation with’, ten overstaan van ‘facing’, onder auspiciën van ‘sponsored by’, ter wille van ‘on behalf of’, onder de hoede van ‘under protection of’, in het bijzijn van ‘in the presence of’, op voorspraak van ‘recommended by’

The presence of idiosyncratic prepositions, case-marked nouns, and of idiosyncratic nouns is evidence for the collocational status of these expressions.

Absence of a determiner. Singular count nouns typically require a determiner. Leaving out the determiner de the sentences become ill-formed in the examples (9)–(11).

(9) Dit vormt *(de) basis van haar betoog
   this forms the base of her story
(10) De banken dringen aan op het ontslag van *(de) leiding
    the banks insist on at the firing of the management
    ‘The banks insist on firing the management.’
(11) Hij neemt *(de) plaats in van de geblesseerde aanvoerder
    he takes the place in of the injured captain
    ‘He replaces the injured captain.’
Yet, many of CPPs contain just singular count nouns without a determiner:

(12) *in plaats van* ‘instead of’, *op basis van* ‘based on’, *in tegenstelling tot* ‘as opposed to’, *in verband met* ‘in connection with’, *in ruil voor* ‘in exchange for’, *na afloop van* ‘at the end of’

Other constructions in which singular count nouns occur without a determiner are support verb constructions (i.e. *plaats maken voor* ‘to make way for’, *leiding geven aan* ‘to lead’, *verband houden met* ‘to be connected with’) and other more or less idiomatic expressions (*van huis* ‘away from home’, *naar school* ‘to school’). Again, this suggests that the elements in (12) have a special status.

**Restricted modification.** Many of the nouns found in P-NP-P patterns cannot be modified by an adjective:

(13) *met* (*directe*) *betrekking tot* ‘with (direct) relation to’, *naar* (*concrete*) *aanleiding van* ‘in (direct) reaction to’, *door* (*legitiem*) *middel van* ‘by (legitimate) use of’

In some cases, modification forces the collocation to take on a literal meaning:

(14) *In de ogen van* zijn tegenstanders was het voorstel een ramp.
    in the eyes of his opponents was the proposal a disaster
    ‘In the eyes of his opponents, the proposal was a disaster.’

(15) *In de verschrikte ogen van* zijn tegenstanders was paniek te zien.
    in the frightened eyes of his opponents was panic to see
    ‘In the frightened eyes of his opponents, one could see panic.’

(16) #*In de verschrikte ogen van* zijn tegenstanders was het voorstel een ramp
    a disaster
    ‘In the frightened eyes of his opponents, the proposal was a disaster.’

The fact that no modification is possible and the fact that, where modification is possible, the nonliteral meaning disappears, are often considered to be tests for identifying certain idiomatic expressions and collocations. Section 2.4.2 provided evidence of fixed expressions that allow various types of modification.

**Restricted functionality as complement.** Some verbs select for a PP-complement introduced by a specific preposition. Thus, *geloven* ‘to believe’ selects a complement introduced by *in, twijfelen* ‘to doubt’ by *aan*
and *rekenen* ‘to expect’ by *op*. In such cases, complex prepositional phrases cannot realize the complements of these verbs:

(17) Kim gelooft in de toekomst
    ‘Kim believes in the future.’

(18) *Kim gelooft in tegenstelling tot de toekomst
    Kim believes as opposed to the future

(19) Kim twijfelt aan zijn bedoelingen
    ‘Kim has doubts about his intentions.’

(20) *Kim twijfelt aan de hand van zijn opmerkingen
    Kim has doubts at the hand of his remarks

(21) Kim rekent op een overwinning
    ‘Kim counts on a victory.’

(22) *Kim rekent op grond van een overwinning
    Kim counts on ground of a victory

On the other hand, some verbs specifically select for a CPP:

(23) Het orkest staat onder leiding van een Duitse dirigent
    ‘The orchestra stands under guidance of a German director.’

(24) Kim houdt iedereen op de hoogte van de laatste ontwikkelingen
    ‘Kim keeps everyone at the height of the latest developments’

This suggests that CPPs must be syntactically and/or semantically distinct from regular PPs.

**Limited extraposition.** Dutch allows extraposition of PPs, both from within VPs and within NPs. Most CPPs resist extraposition, however:

(25) Kim heeft het plan in tegenstelling tot haar buurman ondersteund
    Kim has the plan as opposed to her neighbour supported
    ‘Kim has supported the plan, as opposed to her neighbor.’

(26) *Kim heeft het plan ondersteund tot haar buurman

(27) Kim heeft een beslissing op basis van geruchten genomen
    Kim has a decision on the basis of rumours taken
    ‘Kim has made a decision on the basis of rumours.’

(28) *dat ik geen beslissingen op basis neem van geruchten
4.2. Linguistic properties

Where extraposition is allowed, it seems to be restricted to certain verbs which select for a CPP:

(29) Het orkest zal onder leiding staan van een Duitse dirigent
The orchestra shall under guidance stand of a German director
‘The orchestra will be directed by a German director.’

(30) Kim moet iedereen op de hoogte houden van de laatste ontwikkelingen
Kim must everyone at the height keep of the latest developments
‘Kim must keep everybody informed about the latest developments.’

**Pronominal adverbs.** In Dutch, a preposition combining with a so-called R-pronoun (i.e. daar ‘there/that’, hier ‘here/this’) can be realized as a pronominal adverb (i.e. daarvan ‘of that’, hiervan ‘of this’). Some CPPs can be combined with an R-pronoun (31), whereas others cannot (32).

(31) in plaats daarvan ‘instead of that’, op basis daarvan ‘based on that’, naar aanleiding daarvan ‘in reaction to that’, in ruil hiervoor ‘in exchange for this’

(32) *ten koste hiervan ‘at the cost of this’, *bij wijze daarvan ‘by way of that’, *met ingang daarvan ‘starting on that’, *onder het mom hiervan ‘under pretext of this’

For the first type of CPP, an analysis which considers only the initial preposition and the NP as a unit seems appropriate. In such an analysis, the second P-NP combination is considered to be a regular PP, and thus, the possibility of realizing this PP by a pronominal adverb (which is syntactically equivalent to a PP) is predicted. The fact that the second type of CPP cannot combine with a pronominal adverb suggests that these are best analyzed as multi-word units consisting of a P-NP-P pattern.

**Optional complement.** The PP introduced by the second preposition is optional for some CPPs ((33),(34)). In other cases, omission of complements is impossible (35).

(33) Na aftloop (van de voorstelling) klonk applaus.
After end of the show sounded applause
‘Applause was heard at the end (of the show).’

(34) Het werk is uitgevoerd in opdracht (van de regering).
The work was carried-out on behalf of the government
‘The work was carried out at a request (of the government).’
Chapter 4. Identification of Collocational Prepositional Phrases

(35) Kim speelt in plaats *(van de geblesseerde keeper).
Kim plays in place of the injured keeper
‘Kim plays in replacement for the injured keeper.’

4.2.2 Discussion

The properties listed above suggest that CPPs should be distinguished from regular PPs. The fact that CPPs exhibit a number of idiosyncratic syntactic properties (archaic prepositional and nominal forms and inflection, absence of a determiner, restricted possibilities for modification, restricted functionality as complement) suggests that CPPs must at least to some extent be lexicalized.

The details of the lexical representation remain somewhat unclear, however, as there is considerable variation within the class of CPPs. The fact that some CPPs may combine with pronominal adverbs suggests that those CPPs actually consist of a P-NP phrase selecting for a regular PP. The fact that in some cases modification of the noun is possible suggests that the NP within a CPP cannot simply be represented by a (single) word or multi-word unit.

4.3 Extracting CPPs from a corpus

A number of statistical tests can be applied to select patterns with strong collocational properties (as opposed to patterns which just consist of frequent words) from a corpus. Below, we describe how the initial data was collected.

4.3.1 Resources

We used a corpus consisting of newspaper text from de Volkskrant op CD-ROM, 1997. The corpus consists of over 16 million words and over 1 million sentences. The text was tagged with part-of-speech (POS) tags, using the ‘simplified’ WOTAN tagset (Berghmans, 1994), that is, with only major categories. The tagset is briefly described in van Halteren et al. (2001). Tagging was performed automatically, using a Brill-tagger for Dutch (Drenth, 1997). The accuracy of the tagger is estimated to be around 95%.1

We used Gsearch (Corley et al., 2001) to extract syntactic patterns from the tagged corpus. Gsearch allows one to search for substrings matching expressions defined by a context-free grammar. Terminals may refer to (regular

---

1Drenth (1997) reports 95.1% accuracy on the Eindhoven corpus, using 80% for training and 20% for testing, and using only word class information.
4.3. Extracting CPPs from a corpus

expressions matching) POS-tags. For instance, we used the definition given in Table 4.1 to identify base noun phrases. A base noun phrase (BNP) spans the initial (non-recursive) part of a noun phrase up to and including the nominal head.

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>bnp → det ap* noun</code></td>
<td>base (non-recursive) NP</td>
</tr>
<tr>
<td><code>bnp → ap* noun</code></td>
<td></td>
</tr>
<tr>
<td><code>det → ( Art.*)</code></td>
<td>determiner</td>
</tr>
<tr>
<td><code>det → ( Pron.(*attr) )</code></td>
<td>possessive pronoun</td>
</tr>
<tr>
<td><code>det → ( Num.(*attr.*) )</code></td>
<td>numeral</td>
</tr>
<tr>
<td><code>adj → ( Adj(attr.*) )</code></td>
<td>prenominal adjective</td>
</tr>
<tr>
<td><code>noun → ( N(*) )</code></td>
<td>common noun</td>
</tr>
</tbody>
</table>

Table 4.1: Context-free grammar to license a base noun phrase.

4.3.2 Dataset extraction

Potential CPP strings were collected by searching for the pattern `p bnp p`. There were 285,000 matching strings in the corpus, instantiating 163,000 different strings (137,000 strings occur only once, 2,333 strings occur at least 10 times). The ten most frequent patterns are shown in Table 4.2.

<table>
<thead>
<tr>
<th>freq</th>
<th>candidate</th>
<th>freq</th>
<th>candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,253</td>
<td>in plaats van</td>
<td>579</td>
<td>ten opzichte van</td>
</tr>
<tr>
<td>816</td>
<td>op basis van</td>
<td>549</td>
<td>in tegenstelling tot</td>
</tr>
<tr>
<td>710</td>
<td>onder leiding van</td>
<td>541</td>
<td>op grond van</td>
</tr>
<tr>
<td>659</td>
<td>op het gebied van</td>
<td>520</td>
<td>na afloop van</td>
</tr>
<tr>
<td>609</td>
<td>aan het eind van</td>
<td>511</td>
<td>aan de hand van</td>
</tr>
</tbody>
</table>

Table 4.2: Ten most frequent patterns in the dataset.

From the extracted patterns, we removed all strings in which the BNP contained proper names, acronyms, dates, numbers, etc. which we do not consider to be part of potential CPPs (e.g. `aan de Universiteit van` ‘at the University of’, `op het WK in` ‘at the World Championship in’, `op 3 januari in` ‘on January, 3 in’). About 40,000 strings (14%) were removed this way.

While most of the remaining strings are instances of the pattern we are interested in, some false hits occur as well. For instance, the string `op één na` ‘except for one’ looks like an instance of the search pattern, but `na` is a
post-position and the string is a productive pattern functioning as an adverb. Other sources of errors are larger idiomatic phrases which contain a substring matching P BNP P. For instance, the phrase van tijd tot tijd ‘from time to time’ contains a matching substring van tijd tot. Similarly, dag in dag uit ‘day in, day out’ contains the matching substring in dag uit.

### 4.4 Collocation statistical tests

High co-occurrence frequency is often claimed to be a defining feature of collocations (Firth, 1957). This means that two words that co-occur often enough in a given corpus could, in principle, be mutually associated. A problem with this approach is that combinations of frequent words can form frequent bigrams as well, even though they are not collocations. For example, the expression in het centrum van ‘in the centre of’ occurs very frequently in the corpus, but this could just be due to the fact that in and van are highly frequent prepositions, and het centrum is a reasonably frequent NP. More sophisticated statistical tests measure whether a sequence of words occurs more often than would be expected on the basis of the frequencies of the words involved and thus, do not suffer from this problem. We applied three statistical tests to the data extracted with Gsearch, namely mutual information, Pearson's $\chi^2$ and the log-likelihood ratio. (The statistical tests were described in section 3.3.2.) The baseline we used is raw frequency. We experimented with two different setups: a bigram and a trigram model.

#### 4.4.1 Bigram model

As discussed in Section 3.3, common tests for identifying collocations assume that candidates are bigrams. However, we are interested in collocational patterns of the form $P_1$ BNP $P_2$ (preposition base noun phrase preposition). As BNPs can consist of multiple words, this means that we are dealing with strings of length 3 or more. In order to apply the bigram tests to our dataset, we assumed that either $P_1$ BNP forms a unit or that BNP $P_2$ forms a unit. In the first case, we obtain a bigram $P_1$-BNP $P_2$ (e.g. aan de hand van), whereas in the second case we obtain a bigram $P_1$ BNP-$P_2$ (e.g. aan de hand van).

The statistical tests were applied to the set of ($P_1$-BNP, $P_2$) bigrams and to the set of ($P_1$, BNP-$P_2$) bigrams.\(^3\) This results in two ranked lists of bigrams.

---

\(^2\)This error could have been avoided with search queries that require a BNP to the right also.

\(^3\)All test results were collected using Ted Pedersen’s ngram statistics package available at [http://sourceforge.net/projects/ngram](http://sourceforge.net/projects/ngram).
4.4. Collocation statistical tests

Table 4.3: The 20 highest ranked patterns using combined log-likelihood scores with a frequency cut-off of 10. The last column lists the rank assigned to the $P_1 \big< \text{BNP} \big< P_2$ and $P_1 \big< \text{BNP} \big< P_2$ bigrams, respectively.

<table>
<thead>
<tr>
<th>Candidate bigram</th>
<th>Rank</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>in tegenstelling tot</td>
<td>2</td>
<td>6213.12</td>
</tr>
<tr>
<td>in tegenstelling tot</td>
<td>12</td>
<td>4126.01</td>
</tr>
<tr>
<td>ten koste van</td>
<td>33</td>
<td>1648.41</td>
</tr>
<tr>
<td>ten koste van</td>
<td>8</td>
<td>4842.68</td>
</tr>
</tbody>
</table>

Table 4.4: Log-likelihood scores and ranks assigned to two different bigrams representing the triples in tegenstelling tot and ten koste van.

The final rank of a $P_1 \big< \text{BNP} \big< P_2$ pattern was determined on the basis of the sum of the ranks assigned in the two bigram-sets. An example is shown in table 4.3. Note that the rank of a pattern can differ strongly, depending on the method that was used to form the bigram. The pattern in tegenstelling tot is assigned ranks 12 and 2, respectively. The difference can be explained by observing that in is a highly frequent preposition and tot a relatively infrequent preposition.

The probability distribution of an outcome $P_1 \big< \text{BNP} \big< P_2$ differs from the probability distribution of an outcome $P_1 \big< \text{BNP} \big< P_2$. We believe that the association score assigned to events of types $P_1 \big< \text{BNP} \big< P_2$ and $P_1 \big< \text{BNP} \big< P_2$ cannot be directly compared because the magnitude of the scores can be very different as the two examples show in table 4.4.

The magnitude of the score assigned to an outcome $P_1 \big< \text{BNP} \big< P_2$ might be very high and the magnitude of the score assigned to its associated outcome $P_1 \big< \text{BNP} \big< P_2$ relatively low. Since the outcomes belong to two different prob-
ability distributions, it is not obvious that the scores could simply be added up. For this reason, in order to determine the association score of the outcome $P_1 \text{BNP} P_2$, instead of adding up the scores assigned to each outcome ($P_1 \text{BNP} P_2$ and $P_1 \text{BNP} P_2$) we added up the outcomes’ corresponding ranks. Ranks also reflect the degree of association between the component words inside an outcome, therefore we believe that results should not degrade. Recall that a high rank means that the statistics inferred a strong dependence between the candidate’s component words.

4.4.2 Trigram model

An obvious alternative to using a combination of bigram scores is to use scores for trigrams (where the BNP, with varying length, is still treated as a unit). We attempted a method to compute the mutual information and $\chi^2$ test of trigrams.\(^4\)

**Mutual Information.** The mutual information score of a trigram is the result of dividing the joint probability of the words inside a particular trigram by the product of the independent probabilities of each word in that trigram:\(^5\)

$$I(w_1, w_2, w_3) = \log_2 \frac{P(w_1, w_2, w_3)}{P(w_1)P(w_2)P(w_3)}$$

**Pearson’s $\chi^2$.** The $\chi^2$ test computes for a trigram $(w_1, w_2, w_3)$ how much the observed frequency $(O_{ijk})$ of each of the trigrams $(i, j, k)$, such that

$$(i, j, k) \in \{w_1 w_2 w_3, w_1 w_2 - w_3, w_1 - w_2 w_3, w_1 - w_2 - w_3, -w_1 w_2 w_3, -w_1 w_2 - w_3, -w_1 - w_2 w_3, -w_1 - w_2 - w_3\}$$

deviates from the expected frequency $(E_{ijk})$:

$$\chi^2 = \sum_{i,j,k} \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}}$$

\(^4\)In the trigram set-up, we do not apply the log-likelihood test given that the necessary term factorization is unknown to us.

\(^5\)Following an idea in Lin (1998), we also experimented with a formulation of mutual information where, in the denominator, the scores of $w_1$ and $w_3$ depend on $w_2$ (the idea being that the choice of the prepositions depends strongly on the noun), but in our evaluation, this gave the same results as the formula shown here.
4.5 Evaluation and results

4.5.1 Methodology

We considered two parameters for the evaluation of the models: a frequency threshold and varying nbest lists.

**Frequency threshold** Collocation density is typically lower among low-frequency data than among high-frequency data. We chose two different frequency cutoffs: \( f \geq 10 \) and \( f \geq 40 \). Once we applied the frequency thresholds, we are left with 2084 collocation candidates showing a frequency bigger than 10 and 317 candidates with a frequency bigger than 40. The assumption is that word combinations that occur less than 10 times in the corpus are not likely to yield reliable data. However, we keep the candidates with a frequency \( 10 \geq f \leq 40 \) to make the task a bit harder for the statistical model.

**Varying nbest lists** Nbest lists include the \( N \) highest ranked candidates resulting from the application of a statistical test. Three different nbest lists are chosen so that we evaluate those candidates with a frequency bigger than 10 among the top 100, the top 300 and all ranked candidates.

The accuracy of a test is measured by counting how many items of a pre-defined list of CPPs are among the nbest collocation candidates according to the test. We measure the accuracy of the tests with the (approximate) precision measure (see Section 3.5). To compare the behavior of the tests, we also calculate the Spearman’s rank correlation coefficient which we immediately describe.\(^6\)

The *Spearman’s rank correlation coefficient* (\( \rho_s \)) is a non-parametric procedure used to measure the degree of association between the values of two variables \( X \) and \( Y \) (Oakes, 1998). In this application of the Spearman’s coefficient, the \( X \) and \( Y \) values correspond to the association scores assigned to candidate CPPs and proposed by two statistics (e.g. mutual information and log-likelihood). The \( \rho_s \) value can be computed as follows:

\(^6\)Bouma and Villada (2002) and Villada Moirón and Bouma (2002) did not include this coefficient in the evaluation.
Chapter 4. Identification of Collocational Prepositional Phrases

\[ \rho_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \]

The association scores proposed by each statistic are ranked such that the highest score is given a rank of 1, the next highest value is given a rank of 2 and so on. For each candidate \( i \) in the dataset, two ranks are taken into account: one is assigned by statistic \( a \) and the other by statistic \( b \). The absolute value of the difference between these two ranks corresponds to \( d_i \) in the formula above. \( N \) is the number of common candidates in the two ranked lists. If the resulting coefficient value is greater than the critical value, we can be certain that there is an association between the two variables \( X \) and \( Y \); in this case, it means that the rankings proposed by the two statistics correlate significantly.

Validation data Until recently the only existing list of collocational prepositional phrases available was that proposed in Paardekooper (1973) and Haeseryn et al. (1997). The \( \text{ANS} \) list contains 86 CPPs, however, a few expressions are not used in current Dutch (e.g. `according to', `because of', `to the amount of').

Giving the limited validation data, Bouma and Villada (2002) manually compiled a list of CPPs from the monolingual Van Dale dictionary (Geerts and Heestermans, 1992). 200 candidate expressions proposed as true CPPs by the log-likelihood test were looked up in the abovementioned dictionary. In the event that a candidate expression (e.g. `opposed to') is listed under the lexical entry of the head noun (e.g. `tegenstelling') this special use of the noun is considered a true CPP. Bouma and Villada’s (2002) validation list consisted of 88 CPPs listed in Van Dale’s dictionary as ‘special phrases’. Table A.5 in appendix A provides this list.

Recently, Loonen (2003) published a systematic and thorough study of the morphology, syntax and semantics of Dutch prepositions. Loonen’s work provides two electronically available lexica: one of prepositions and another one with fixed expressions involving prepositions. The second lexicon consists of a large variety of expressions among which collocational PPs are included. From Loonen’s second lexicon, a list of 287 CPPs was compiled.

The evaluation data used in this chapter is the result of merging Bouma and Villada’s (2002) validation data and Loonen’s CPPs list. This list contains 315 CPPs.\(^7\) Table 4.5 provides the distribution of the true positives

\(^7\)Items that were originally absent in Loonen’s list are highlighted in bold face in Appendix A, Table A.5. 29 CPPs are specific to Bouma and Villada’s (2002) list.
present in our data. 25% of the CPPs in our validation data are not found in the dataset.

<table>
<thead>
<tr>
<th>Cutoff</th>
<th>Outcome types</th>
<th>True positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>138,857</td>
<td>232</td>
</tr>
<tr>
<td>$f \geq 10$</td>
<td>2,084</td>
<td>169</td>
</tr>
<tr>
<td>$f \geq 40$</td>
<td>317</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 4.5: True positives found in dataset

### 4.5.2 Bigrams results

Table 4.6 gives the results of applying mutual information (MI), log-likelihood (LL) and $\chi^2$.\(^8\)

<table>
<thead>
<tr>
<th>test</th>
<th>freq</th>
<th>n</th>
<th>nbest</th>
<th>freq</th>
<th>n</th>
<th>nbest</th>
<th>raw freq</th>
<th>freq</th>
<th>n</th>
<th>nbest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 10$</td>
<td>2,084</td>
<td>65</td>
<td>107</td>
<td>169</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 40$</td>
<td>317</td>
<td>65</td>
<td>96</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 10$</td>
<td>2,084</td>
<td>64</td>
<td>95</td>
<td>169</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\geq 40$</td>
<td>317</td>
<td>64</td>
<td>95</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Bigram model results of mutual information, log-likelihood, $\chi^2$ and raw frequency.

Mutual information, when used with a frequency cutoff of 10, leads to a disproportional number of low frequency patterns among the highest scoring items, leading to poor results. Using a higher frequency cutoff of 40, more true positives are found among the smaller nbest lists. Considering an nbest list larger than 300 does not improve results because there are only 317 items that occur more than 40 times in the dataset.

Comparing the results of log-likelihood and Pearson’s $\chi^2$ for the two different thresholds, it becomes apparent how these two tests are not as sensitive

\(^8\)Refer to section 3.3.2 for a description of these measures.
to low-frequency data as mutual information. In terms of precision, the performance of these tests is fairly similar in the bigram model, although $\chi^2$ is less accurate when nbest=100.

If we only consider the number of true positives found in the complete list (all column), raw frequency produces as good results as MI, LL and $\chi^2$; for smaller nbest lists raw frequency comes closer to the log-likelihood and $\chi^2$ coverage. This seems to suggest that selecting the most frequent instances of a sequence of adjacent syntactic categories in a part-of-speech tagged corpus may provide a rather accurate list of multi-word units. Nevertheless, the raw frequency ranking allows slightly more noise among the top scores (for nbest=50, raw frequency identifies 36 true positives and log-likelihood identifies 41). Perhaps, for some collocation patterns, raw frequency is a suitable test (cf. Evert and Krenn (2001)). In our experiments and given our settings, log-likelihood yields CPPs lists of a better quality.

Expressions included in the validation data and not found by any statistic either have a frequency value that is lower than the cutoff or, the expressions are not present in the extracted datasets (missing in the corpus). The use of a cutoff has a positive and a negative side effect. On the one hand, a cutoff of $f \geq 10$ brings about an improvement on the precision of the tests, but on the other hand it discards many true CPPs from the dataset. The figures in the last row in Table 4.6 show that 63 true CPPs are ignored due to the cutoff requirement.

### 4.5.3 Trigrams results

Table 4.7 shows the results of applying the mutual information and $\chi^2$ test to trigrams. As was the case for bigrams, mutual information performs poorly on low frequency data. However, the MI results are better than in the bigram model. Overall, it seems that using the proposed trigrams method does not lead to improved results.

### 4.6 Discussion

This section discusses to what extent the association measures are useful to identify CPPs in a part-of-speech tagged corpus. First, we compare the tests’ performance and select the statistics that produces the best list, both in terms of precision and coverage. Next, we discuss the errors made by the selected test. To conclude we comment on the positive effect of having more
4.6. Discussion

<table>
<thead>
<tr>
<th>test</th>
<th>freq</th>
<th>n</th>
<th>100</th>
<th>300</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>$\geq 10$</td>
<td>2,084</td>
<td>34</td>
<td>84</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td>MI</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>$\geq 40$</td>
<td>317</td>
<td>56</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>raw freq</td>
<td>$\geq 10$</td>
<td>2,084</td>
<td>64</td>
<td>95</td>
<td>169</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\geq 40$</td>
<td>317</td>
<td>64</td>
<td>95</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 4.7: Results of mutual information, $\chi^2$ and raw frequency applied on trigrams.

<table>
<thead>
<tr>
<th>tests</th>
<th>cutoff</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL - $\chi^2$</td>
<td>$\geq 10$</td>
<td>0.0421</td>
</tr>
<tr>
<td>MI - LL</td>
<td>$\geq 40$</td>
<td>0.0333</td>
</tr>
<tr>
<td>MI - $\chi^2$</td>
<td>$\geq 10$</td>
<td>-0.2044</td>
</tr>
<tr>
<td></td>
<td>$\geq 40$</td>
<td>-0.1715</td>
</tr>
</tbody>
</table>

Table 4.8: Spearman’s rank correlation coefficients of the rankings proposed by the association measures when applied on a bigram model.

validation data by comparing the results reported in Bouma and Villada (2002) and Villada Moirón and Bouma (2002) and the new results with the larger validation data.

4.6.1 Identifying the best association measure

Considering the similar accuracies achieved by the log-likelihood ratio and the $\chi^2$ (given in Table 4.6) it is interesting to compare the ranked lists of candidates produced by these statistics. We compared the rankings proposed by mutual information, $\chi^2$ and log-likelihood using the Spearman’s rank correlation coefficient. The resulting coefficients are given in Table 4.8. The critical value for the Spearman’s rank correlation coefficient is 0.165 ($\alpha = 0.05$) for a one-tailed test. A one-tailed test examines whether two variables X and Y are positively associated.

Among the 100 higher scores, the rankings proposed by the three association measures differ significantly. If low frequency data is included ($\geq$
10), the biggest difference is observed between MI and LL ($\rho = -0.2044$) and LL and $\chi^2$ ($\rho = 0.0421$). A similar trend is observed with a higher cutoff. When low frequency data is discarded, the rankings proposed by MI and $\chi^2$ correlate significantly ($\rho=0.5421$).

Two reasons explain this divergence: first, the two ranked lists (n=100) under comparison contain few common candidates and second, the rankings of the common candidates differ substantially. Among the 100 top scores proposed by MI and LL, with a frequency bigger than 10, there are only 14 candidates in common; among those proposed by $\chi^2$ and LL, there are 47 common expressions and, among those proposed by $\chi^2$ and MI there are 79 common candidates. Related to the second reason, Table 4.9 shows the rather different rankings of a few true positives proposed by the three association measures.

<table>
<thead>
<tr>
<th>expression</th>
<th>LL</th>
<th>$\chi^2$</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>in tegenstelling tot</td>
<td>1</td>
<td>112</td>
<td>131</td>
</tr>
<tr>
<td>op basis van</td>
<td>2</td>
<td>22</td>
<td>407</td>
</tr>
<tr>
<td>onder leiding van</td>
<td>3</td>
<td>3</td>
<td>105</td>
</tr>
<tr>
<td>ten opzichte van</td>
<td>7</td>
<td>166</td>
<td>61</td>
</tr>
<tr>
<td>op grond van</td>
<td>8</td>
<td>43</td>
<td>429</td>
</tr>
<tr>
<td>in handen van</td>
<td>31</td>
<td>114</td>
<td>1157</td>
</tr>
</tbody>
</table>

Table 4.9: Ranks proposed by three different association measures

Concerning the trigram model, the Spearman’s rank correlation coefficient values given in Table 4.10 show a significant correlation between the rankings proposed by mutual information and $\chi^2$; the correlation between $\chi^2$ and raw frequency is poor. Recall that the critical value of the Spearman’s rank correlation coefficient is $\rho = 0.165$ ($\alpha = 0.05$, one-tailed test). The level of correlation between the tests is greater when low frequency data is discarded. The values in Table 4.10 suggest that the proposed formulation of $\chi^2$ and mutual information produce comparable rankings.

Close inspection of the rankings reveals that the log-likelihood ratio ranks the true positives (actual CPPs) higher than the $\chi^2$ test does, thus log-likelihood allows less noise among the top scores. From this we may conclude that the strong associations inferred by the log-likelihood test are more accurate than those inferred by the $\chi^2$ test given our data. In addition, the performance of the log-likelihood test is the least sensitive to the frequency cutoff parameter. Based on this evidence, we select the log-likelihood as the best statistic for the task of identifying CPPs in a tagged corpus. In general
4.6. Discussion

<table>
<thead>
<tr>
<th>tests</th>
<th>cutoff</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>MI - (\chi^2)</td>
<td>(\geq 10)</td>
<td>0.3407</td>
</tr>
<tr>
<td>MI - rawfreq</td>
<td>(\geq 40)</td>
<td>0.7598</td>
</tr>
<tr>
<td>(\chi^2) - rawfreq</td>
<td>(\geq 10)</td>
<td>0.0943</td>
</tr>
<tr>
<td>(\chi^2) - rawfreq</td>
<td>(\geq 40)</td>
<td>0.0007</td>
</tr>
<tr>
<td>(\chi^2) - rawfreq</td>
<td>(\geq 10)</td>
<td>0.2374</td>
</tr>
<tr>
<td>(\chi^2) - rawfreq</td>
<td>(\geq 40)</td>
<td>0.1354</td>
</tr>
</tbody>
</table>

Table 4.10: Spearman’s rank correlation coefficients of the rankings proposed by mutual information, \(\chi^2\) and raw frequency when applied on a trigram model.

the low \(\rho\) scores are encouraging, since they indicate that we are comparing genuinely different tests.

### 4.6.2 Error analysis

All false positives found among the highest scores are shown in Table 4.11. These are known as ‘type I errors’ which means that the score assigned by the statistics suggests that one can confidently reject the hypothesis of independence. That is, the data that the statistics is based on provides more evidence of a dependence between the component words than of independence.

Table 4.11: Log-likelihood false positives among top ranks.

False positives ranked by log-likelihood among the 100 top scores exhibit
one or more of the following characteristics:

- The two prepositions in the candidate $P_1$ BNP $P_2$ are very frequent and the noun phrase is also rather frequent (e.g. `in de nacht van `in the night of`).

- BNP $P_2$ constitute a syntactic colligation; this combination is not restricted to occur in the context of $P_1$. Examples are `in interview met `in an interview with`, `in gesprek met `talking with`.

- A frequent $P_1$ BNP string cooccurring with a $P_2$ that introduces a post-nominal modifier with the preposition van (e.g. `bij de presentatie van `during the presentation of`, `in handen van `in the hands of`, `van het ministerie van `of the ministry of`).

- $P_1$ BNP phrases that often co-occur in the same context as $P_2$ due to an external cause which is often a verb selecting both, the phrase and a complement introduced by $P_2$. Examples are `in dienst van (stellen/nemen/zijn) `in service of (declare/take/be)`, `op smaak met (bren-gen) `(lit) bring on taste with`.

Close inspection of the type I errors reveals that this statistic favors frequent candidates over infrequent ones and candidates with frequent component words over infrequent ones. Thus, most errors consist of highly frequent prepositions (`in, van, met`), determinerless noun phrases (`zuiden, oosten, dienst, smaak, hulp`) or noun phrases that typically exhibit no internal modification but a postmodifier introduced by van `of`. Nevertheless, `in dienst van `in service of`, `in handen van `under the control of` and `op initiatief van `on the initiative of` could be considered true CPPs. These were not present in our validation data.

The previous frequency-based explanation would also explain why many true positives with low frequency ($f \leq 35$) are assigned a low rank (rank $\geq 300$). Some of the latter are shown in Table 4.12. Some of the true positives that are assigned a lower score have a low frequency due to the fact that they allow morpho-syntactic variation. For example, `uit het oogpunt van `from the standpoint of` (19 occurrences), `uit oogpunt van` (17) and `uit een oogpunt van` (12). Other cases include P NP P candidates whose P NP is observed with various prepositions, e.g. `van belang voor `of importance to` (52), `van belang(en) in` (13), `van belang vanwege` (2), `van belang tijdens` (2). Variation of the second preposition could be due to different uses of the phrase P NP. Each of these variants is treated as a separate outcome, thus the actual probability distribution of the outcome
4.6. Discussion

Table 4.12: Log-likelihood true positives among low ranks.

PREP NOUN PREP is split across all its variants. In our opinion, the fact that the model does not take into account that an NP may exhibit limited variation has a negative effect on the scoring process.

Possible improvements of this model could be (i) improve the search patterns such that only P NP P patterns that are followed by a baseNP are extracted and (ii) clustering the data by representing the noun phrase as its noun only.

4.6.3 Effects of having more validation data

Access to (almost) four times more validation data allows a more informed assessment of the performance of the statistics. Here we only report changes pertaining the bigram model, since the results are better than the trigram model results.

When we only consider the nbest list containing the 100 best scores, precision increase ranges from 11% (MI) to 15% (LL, $\chi^2$ and raw frequency) using the $f \geq 10$ cutoff. With a higher cutoff $f \geq 40$, MI and $\chi^2$ show a 10% precision increase and LL a 14% increase. This implies that more candidates among the 100 best scores were correctly identified by LL, $\chi^2$ and raw frequency than our previous results had shown. LL reached a 65% precision, $\chi^2$ 55%, raw frequency 64% and MI 32% (when nbest=100, f $\geq$ 10). Among the 300 best scores and using a low cutoff, the precision increase of LL and $\chi^2$ is more significant: 14%. MI and raw frequency show 12% and 10% precision increase, respectively. With a cutoff of $f \geq 40$ all tests’ precision increases 11%. Given these settings, LL and $\chi^2$ perform equally well (35.67% and 35.33% precision, respectively; $f \geq 10$, n=100). Overall, the test whose
performance is less sensitive to low frequency data is the log-likelihood test.

A conclusion to draw from these precision figures is that the substantial increase in validation data reveals that ranking the candidate patterns with the log-likelihood statistic still holds as the best technique. Furthermore, performance difference between the tests becomes larger.

4.7 Summary

Collocational prepositional phrases have a number of syntactic properties which suggest that they need to be distinguished from regular PPs. Although CPPs are collocational, they do not always act as fixed multi-word expressions. The background study revealed that some CPPs cannot be formalized as a multi-word-lexeme inserted as a fixed string in a lexicon. Variation needs to be allowed and therefore, internal structure is required.

We have described a corpus-based method for acquiring CPPs from corpora, in which potential CPPs are first extracted from the corpus on the basis of syntactic criteria, and next, a ranked list is constructed using statistical collocation tests. The statistical tests were evaluated against a list of CPPs extracted from a dictionary and from Loonen’s (2003) lexicon. The log-likelihood ratio test was identified as the one that provides the best trade-off between precision and coverage.