Corpus-based acquisition of collocational prepositional phrases

Gosse Bouma and Begoña Villada

Alfa-Informatica
Rijksuniversiteit Groningen

Abstract

Collocational prepositional phrases like *ten koste van* (*at the expense of*), *met het oog op* (*with an eye on*), and *onder het mom van* (*under the pretext of*) are patterns of the form P-NP-P, which have a non-compositional semantics and which are syntactically rigid or idiosyncratic. We present a number of linguistic tests which set such items apart from regularly built prepositional phrases.

To find candidate strings which should be included in a computational lexicon as collocational prepositional phrases, we extract all instances of the relevant pattern from a corpus annotated with POS tags. Next, we introduce a number of statistical tests (mutual information, log-likelihood, and χ^2) to find those instances which behave like strong collocations.

The strongest collocations according to the statistical tests are compared with lists of such items presented elsewhere, and were evaluated by human judges.

1 Introduction

We are currently involved in the development of a wide-coverage lexicalist computational grammar for Dutch (Bouma, van Noord and Malouf 2001). Expressions with idiosyncratic syntax and semantics are problematic for a computational grammar because they often do not follow the rules of regular syntax. Furthermore, idiomatic expressions which are syntactically regular still need to be recognized as such, in order to account for the fact that their semantics is non-compositional. In this paper, we investigate the linguistic properties of one particular class of idiomatic expressions, collocational prepositional phrases, and explore whether corpus-based methods can be used to acquire such expressions. Ultimately, we hope to provide both a linguistic analysis of such expressions, compatible with the Alpino grammar framework, as well as a full listing of such expressions, to be included in the lexicon.

Dutch has a number of preposition-(determiner-)noun-preposition combinations, which are more or less fixed:

(1) *ten opzichte van* ('with respect to'), *in tegenstelling tot* ('as opposed to'), *in verband met* ('in connection with'), *in plaats van* ('instead of'), *op basis van* ('on the basis of'), *naar aanleiding van* ('in response to'), *ter gelegenheid van* ('on the occassion of'), *te midden van* ('amidst'), *in het kader van* ('in the framework of'), *aan de hand van* ('on the basis of')

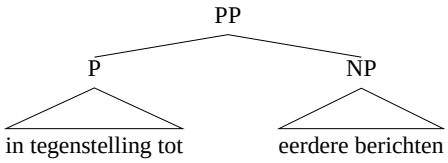
In Dutch linguistics such expressions are known as *voorzetsel-uitdrukkingen* (Paardekooper 1962). Here, we will refer to them as *collocational prepositional phrases* (CPPs).

One might analyze phrases introduced by these constructions, such as (2), as

prepositional multi-word units, as shown in (3).

(2) in tegenstelling tot eerdere berichten
as opposed to earlier reports

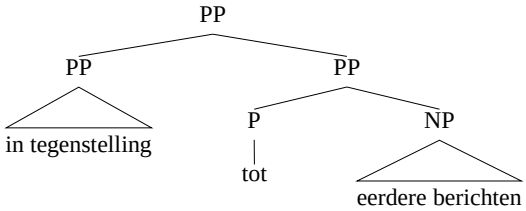
(3)



This is the analysis adopted by the annotation guidelines of the Corpus Gesproken Nederlands (Moortgat, Schuurman and van der Wouden 2000). It presupposes that these expressions behave as units syntactically, and can be listed in the lexicon.

Another option is to consider only the initial P-NP combination as a unit:

(4)



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 initial P-NP expression.

In the next section, we list a number of linguistic tests which suggest that collocational prepositional phrases need to be distinguished from regular P-NP-P constructions. However, the tests also indicate that for some CPPs the analysis in (3) is appropriate, whereas for others the analysis in (4) is more likely.

The analysis in {reftypeB requires that P-NP-P combinations are listed as complex prepositions in the lexicon, whereas the analysis in {reftypeC requires that P-NP combinations are listed as prepositional phrases selecting for a PP-complement headed by a specific preposition. Thus, in both cases the question arises whether a listing can be provided of such expressions. Paardekooper (1973) and the *Algemene Nederlandse Spraakkunst* (Geerts, Haeseryn, de Rooij and van den Toorn 1984) provide a list of CPPs, which is not claimed to be exhaustive, however. A (bilingual) dictionary like Van Dale, tends to mention CPPs in the lemma's for the corresponding noun.

In the second half of the paper, we will be concerned with the question to what extent corpus-based methods can be used to obtain a more complete listing of CPPs. In particular, we collected all occurrences of P-NP-P patterns from a corpus, and applied a number of statistical tests to the results to obtain ranked lists of potential CPPs. The results were evaluated by comparing these lists with a listing extracted from the Van Dale dictionary. Potential CPPs not included in this list were evaluated by human judges.

2 Linguistic Properties

In this section, we mention a number of linguistic properties which set collocational prepositional phrases apart from syntactically and semantically regular PPs. Most of these properties were already observed by Paardekooper (1962) and Paardekooper (1973).

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 also occur in fixed, archaic, expressions only. Thus, the inflected forms of the nouns following *ten* in the examples below, do not occur outside fixed expressions:

(5) *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:

(6) *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:

(7) Dit vormt *(de) basis van haar betoog
this forms the base of her story

(8) 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

(9) 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:

(10) *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 phrasal verbs (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 (10) have collocational status.

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

(11) *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:

(12) 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

(13) 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

(14) *In de verschrikte ogen van zijn tegenstanders was het voorstel een
In the frightened eyes of his opponents was the proposal a
ramp
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 *metaphoric* meaning disappears, are often considered to be tests for identifying idiomatic expressions and collocations.

Restricted functionality as complement. Some verbs select for a PP-complement introduced by a specific preposition. In such cases, complex prepositional phrases are excluded:

(15) Kim gelooft in de toekomst
Kim believes in the future

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

(17) Kim twijfelt aan zijn bedoelingen
Kim has doubts about his intentions

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

(19) Kim rekent op een overwinning
Kim counts on a victory

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

This suggests that CPPs must be syntactically 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:

(21) 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 het neighbor

(22) *Kim heeft het plan *in tegenstelling* ondersteund *tot* haar buurman

(23) 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

(24) *dat ik geen beslissingen op basis neem van geruchten

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

(25) 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

(26) 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 everbody 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, whereas for others this is almost[1] impossible.

(27) *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')

(28) **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')

[1]One reviewer points out that *ten koste hiervan* may in fact be acceptable. We entered all phrases in (27) and (28), both with *hier* and *daar* as adverb, as search terms in Google (`www.google.com`) and found between 910 and 19,300 occurrences of the phrases in (27) and virtually no occurrences of the phrases in (28) (*ten koste hiervan (1), ten koste daarvan (15), bij wijze hiervan (1), onder het mom daarvan (2)*, all other phrases were not found).

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:

(29) Na afloop (van de voorstelling) klonk applaus
After end of the show sounded applause
Applause sounded at the end (of the show)

(30) 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)

In other cases, omission of complements is impossible.

(31) Kim speelt in plaats *(van de geblesseerde keeper)
Kim plays in place of the injured keeper
Kim plays instead of the injured keeper

Discussion. The properties listed above provide evidence for the fact 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 seems to imply 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.

3 Extracting CPPs from a corpus

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) contains a list of 54 items, which is included in the list of 83 items given in the ANS. This list is not claimed to be complete, however.

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. A number of statistical tests can be ap-

plied to select patterns with strong collocational properties (as opposed to patterns which just consist of frequent words) from such a list. Below, we describe how we collected the initial data. In the next section, the results of applying various statistical tests are presented.

We used a corpus consisting of 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 tags, using the WOTAN tagset (Berghmans 1994). The tagset is briefly described in van Halteren, Zavrel and Daelemans (2001). Tagging was performed automatically, using a Brill-tagger for Dutch (Drenth 1997). The accuracy of the tagger is estimated to be around 95%.[2]

We used Gsearch (Corley, Corley, Keller, Crocker and Trewin 2001) to extract syntactic patterns from the corpus. Gsearch allows one to search for substrings matching expressions defined by a context-free grammar. Terminals may refer to (regular expressions matching) POS-tags. For instance, we used the following definition to identify *base noun phrases* (i.e the initial (non-recursive) part of a noun phrase up to and including the nominal head) :

(32)
```
bnp --> det ap* noun          base (non-recursive) NP
bnp --> ap* noun
det --> <Art.*>               determiner
det --> <Pron\(.*attr\)>      possessive pronoun
det --> <Num\(.*attr.*>       numeral
adj --> <Adj\(attr.*>         prenominal adjective
noun --> <N\(.*>              common noun
```

Potential CPP strings can now be found 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:

(33)

1,253	in plaats van	579	ten opzichte van
816	op basis van	549	in tegenstelling tot
710	onder leiding van	541	op grond van
659	op het gebied van	520	na afloop van
609	aan het eind van	511	aan de hand van

We removed from the results all strings in which the BNP contained a capital letter or a number (*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')), as these involve names, acronyms, dates, numbers, etc. which we do not consider to be part of potential CPPs. 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') instantiates the search pattern, but is in fact an instance of a PP containing a circumposition

[2] Drenth (1997) reports 95.1% accuracy on the Eindhoven corpus, using 80% for training and 20% for testing, and using only word class information.

(*op .. na*). 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* ('from day to day') contains the matching substring *in dag uit*.

4 Statistical collocation tests

The simplest statistical test for finding collocations is mere co-occurrence frequency. High co-occurrence frequency is often claimed to be a feature of collocations. 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* occurs very frequently in the corpus, but this could be due just 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. In this section, we apply a number of statistical tests to the data extracted with Gsearch. Evaluation proceeds by counting how many items of a predefined list of CPPs are among the N-best collocation candidates according to the test.

4.1 Three common collocation tests

The following tests are often used to determine whether two co-occurring words are potential collocations.

Mutual Information. Mutual information (Church and Hanks 1990, p.23) compares the probability of observing two words (w_1, w_2) *together* with the probabilities of observing the same words *independently* in a given corpus:

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

Log-likelihood score. In order to determine whether two words show a strong lexical association, the log-likelihood score explores which of the following two hypotheses is more likely:

- $\mathrm{H}_1 : P(w_2|w_1) = P(w_2|\neg w_1)$
- $\mathrm{H}_2 : P(w_2|w_1) \neq P(w_2|\neg w_1)$

H_1 assumes that the two words are independent, whereas H_2 states that the two words are dependent. If H_2 is more likely, the two words are potential collocations. Log-likelihood measures how much more likely H_2 is than H_1 (see Manning and Schütze (1999, p.173) and Dunning (1993)).

Pearson's χ^2 test. The χ^2 test computes for a bigram (w_1, w_2) how much the observed frequency of each of the bigrams (i, j), such that

$$(i, j) \in \{w_1 w_2, w_1 \neg w_2, \neg w_1 w_2, \neg w_1 \neg w_2\}$$

deviates from the expected frequency of these bigrams (Manning and Schütze 1999, p.169):

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

4.2 Applying bigram tests

The common tests for identifying collocations all assume that collocations are bigrams. However, we are interested in collocational patterns of the form P_1 BNP P_2. As BNP's 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 data-set, 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 (i.e. *aan_de_hand van*), whereas in the second case we obtain a bigram P_1 BNP_P_2 (i.e. *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. The final rank of a P_1 BNP 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 1. 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.

To evaluate how the statistical tests compare to using raw frequency, and to determine which of the tests works best, we compared the n highest ranked items found by a given test with a list of 88 CPPs extracted from the Van Dale dictionary (Geerts and Heestermans 1992). This list was constructed by checking for a number of nouns whether a CPP pattern was mentioned in the lexical entry for that noun. If this was the case, we took this as evidence for the collocational status of the pattern. The ANS contains a similar list.

Table 2 gives the results of applying mutual information (mi), log-likelihood (ll) and χ^2 to the extracted collocation candidates when treated as bigrams. We used 10 and 40 as frequency cut-offs (i.e. only patterns occurring at least 10 or 40 times are considered). The 100 and 300 best items found by the tests are compared with the list extracted from Van Dale, as well as the full list of items above the frequency threshold (all). The final row gives the score for raw frequency, i.e. the score for the 100 and 300 most frequent items, and for the full set of all extracted patterns. The latter is of interest mainly because it illustrates that some items (n=7)

[3] All test results were collected using Ted Pedersen's Bigram Statistics Package, `http://www.d.umn.edu/~tpederse/code.html`

1	in plaats van	2	1	11	met behulp van	15	14
2	onder leiding van	1	4	12	na afloop van	5	27
3	op basis van	4	3	13	aan de hand van	10	24
4	ten opzichte van	3	8	14	in verband met	24	10
5	op het gebied van	7	6	15	in opdracht van	20	15
6	aan het eind van	6	7	16	in het kader van	23	13
7	in tegenstelling tot	12	2	17	in ruil voor	26	11
8	op weg naar	14	5	18	op verzoek van	19	22
9	op grond van	11	9	19	in de loop van	18	23
10	naar aanleiding van	9	12	20	ten koste van	8	33

Table 1: The 20 highest ranked patterns using combined log-likelihood scores with a frequency cut-off of 10. The last column lists the rank using the P BNP_P and P_BNP P bigrams, respectively.

			nbest		
test	freq	n	100	300	all
mi	≥ 10	2,084	23	39	77
ll	≥ 10	2,084	53	67	77
χ^2	≥ 10	2,084	52	69	77
mi	≥ 40	317	47	67	67
ll	≥ 40	317	53	65	67
χ^2	≥ 40	317	55	65	67
raw freq		248,683	50	65	84

Table 2: Results of mutual information, log-likelihood, χ^2 obtained by combining the ranks of the two bigrams, and compared with raw frequency.

in the Van Dale list occur less than 10 times in the original datasets, and some do not occur at all (n=4).

Discussion. Mutual information, when used with a frequency threshold of 10, leads to a disproportional number of low frequency patterns among the highest scoring items, leading to poor results. It is well-known that the mutual information test performs poorly with sparse data even if large corpora are available and a frequency cut-off is used (Manning and Schütze 1999, p.182). Using a frequency threshold of 40 improves the results for mutual information considerably. As only 317 items occur at least 40 times, this effect can be observed best with N=100.

Log-likelihood and Pearson's χ^2 test perform almost equally well. Both perform well with low frequency data, and slightly outperform raw frequency.

4.3 Trigrams

An obvious alternative to using a combination of bigram scores, as in the previous section, is to use scores for trigrams (where the BNP, with varying length, is still treated as a unit). Both the mutual information and χ^2 test can be generalized to trigrams.

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:[4]

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

Pearson's χ^2. The χ^2 test computes for a trigram (w_1, w_2, w_3) how much the observed frequency of each of the trigrams (i, j, k), such that

$$(i,j,k) \in \{ \begin{array}{l} w_1w_2w_3, w_1w_2\neg w_3, w_1\neg w_2w_3, w_1\neg w_2\neg w_3, \\ \neg w_1w_2w_3, \neg w_1w_2\neg w_3, \neg w_1\neg w_2w_3, \neg w_1\neg w_2\neg w_3 \end{array} \}$$

deviates from the expected frequency:

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

Results and discussion. Table 3 shows the results of applying the mutual information and χ^2 test to trigrams.

			nbest		
test	freq	n	100	300	all
mi	≥ 10	2,084	23	45	77
χ^2	≥ 10	2,084	45	61	77
mi	≥ 40	317	46	67	67
χ^2	≥ 40	317	51	66	67
raw freq		248,683	50	65	84

Table 3: Results of mutual information and χ^2 applied to trigrams compared with raw frequency

As was the case for bigrams, mutual information performs poorly on low frequency data. The results are somewhat better than in the bigram experiments,

[4]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.

however. Overall, it seems that using trigrams does not lead to improved results. The best results are those obtained by using χ^2 on combined bigrams. Whereas in the bigram experiments it was possible to do slightly better than raw frequency, this is not the case in the trigram experiments.

It should be noted, however, that evaluation was carried out only on a list of CPPs extracted from a dictionary. The motivation for using a corpus-based method was that we expect that such lists are incomplete. One of the questions that remains is whether any of the tests provides us with CPPs not in the dictionary list.

5 Human evaluation

In Section 2 we proposed that CPPs behave as lexicalized multi-word units and therefore, should appear in a computational lexicon. As described in Section 3, there are 163,000 candidate types in the corpus that exhibit the same pattern as CPPs. Manual inspection of those 163,000 types would be time consuming and the results of that inspection would be highly dependent on the intuitions of a lexicographer or a linguist. To avoid this, we applied standard association measures that quickly extract strongly associated candidates from original datasets. Section 4 showed that the results of applying these tests were not particularly good.

However, previous work on lexical acquisition of collocations also report rather low coverage of the statistical models. For instance, Lin (1999) compares a list of over 200 extracted collocations to another list of 250 true collocations compiled from an idioms dictionary. To justify low precision (15.7%) and recall (13.7%) values, Lin (1999) attributes them to parser errors and gaps in the idioms dictionary. For comparison, Lin (1999) also uses a manually compiled dictionary of idioms which leads to better precision and recall values (39.4% and 20.9%, respectively). This difference in the results shows how different lexicographers may have different opinions about which non-compositional phrases should be in a dictionary (Lin 1999).

Such differences also make us think that our validation data ought to be expanded. In fact, manual inspection of higher ranked extracted candidates suggests that more candidates satisfy the properties of CPPs (e.g. *op kosten van*). Yet, those candidates are not present in the validation data. Therefore, they could not be accepted as true positives in the evaluation presented above.

An alternative approach would be to ask a lexicographer to manually examine all the extracted collocational candidates and to identify good CPPs along the lines described in Smadja (1993). Our slightly different approach borrows from both Lin (1999) and Smadja (1993).

The purpose of carrying out human evaluation is to determine whether collocations extracted by the statistical models and not included in the available validation data ought to be included in a list of CPPs. If we find out that some extracted candidates are true collocations then we achieve a double profit: (i) enlarge validation data for future research and (ii) establish a more accurate view of the coverage of the statistical models.

Three human judges[5] manually determined which of the extracted collocation candidates should be considered true CPPs. In the remainder, we describe how the evaluation data was prepared. Next, we report the results and their interpretation.

Preparing evaluation data. Since there exists little difference between the results of the χ^2 and the log-likelihood tests, we took the 200 higher ranked candidates result of applying the log-likelihood test to the bigrams setup, for two different frequency thresholds (10 and 40) and, also the 200 most frequent trigrams in the corpus.

To make the judges' task easier, the extracted collocations included in the Van Dale list (validation data) were removed except from 10 test items, 8 of which were true CPPs. Thus, the number of extracted collocations that needed to be examined by the judges was reduced. We assume that extracted candidates included in the validation data (thus, true CPPs) need not be manually evaluated. At the end, judges were given a list of 180 collocation candidates.

Instructions to the human judges Human judges were asked to identify those candidate expressions that fulfil the following five properties: (i) the noun inside the collocation candidate cannot be replaced by a synonym without changing the meaning; (ii) the collocation candidate is not followed by a specific noun; this ensures that the candidate is not part of a longer idiomatic expression; (iii) the second preposition is obligatory; (iv) the collocation candidate does not co-occur with one or two specific verbs and, (v) the noun within the NP does not admit modification.

Results Only 9.4% of the candidate expressions were identified as good CPPs by at least two judges. The list is given in (34).

(34) door gebrek aan, in antwoord op, in de aanloop naar, in plaats van, in reactie op, in tegenstelling tot, in termen van, met dank aan, naar aanleiding van, op advies van, op initiatief van, op kosten van, op uitnodiging van, te midden van, ten behoeve van, ter nagedachtenis aan, voor rekening van.

Among these, 12 (6.8%) expressions constitute new instances of CPPs. Note that this is the result of inspecting only 180 potential candidates, from which most of the known CPPs had already been removed. No significant difference can be observed between the true positives extracted by the log-likelihood score and the raw frequency test.

Only 5 of the 8 true CPPs that were included as test items were classified as good CPPs by at least two human judges. This illustrates the difficulty of the evaluation task.

[5]None of the authors was included.

Discussion The low coverage of the statistical tests can be attributed to two causes: first, the task of identifying CPPs proves to be rather difficult for both human judges and statistical models, and secondly, validation data used to measure the performance of the statistical tests is incomplete.

Human judges commented on the rigidity of the guidelines given to them and agreed that some candidate expressions may occur with a a few verbs but the complex prepositional phrase itself is fixed. Furthermore, the noun within some candidate expressions may be replaced by only a synonym and both expressions are instances of CPPs (e.g. *op zoek naar* and *op jacht naar*). Within some candidates the second preposition is optional but the bigram P_1 BNP is a lexicalized PP. Finally, some candidates look like CPPs but allow restricted modification (e.g. *in scherpe tegenstelling tot*). These comments reinforce our claims in Section 2 and emphasize the fact that CPPs do not constitute a uniform class. Instructions to judges should be more flexible and allow them to consider a dual or triple classification of CPPs.

Another source of difficulty for the statistical models is that many of the extracted collocations form a part of a larger fixed expression, saying or proverb. These expressions cannot be considered as CPPs. The statistical tests correctly identify such expressions as lexically associated word combinations, but they cannot infer that they are part of a larger expression.

6 Conclusions

For a wide-coverage computational grammar it is essential that it has some method for dealing with idiomatic and collocational expressions. At first sight, collocational prepositional phrases seem to behave syntactically as multi-word units. Consequently, they should be relatively easy to identify using corpus-based methods. The linguistic discussion in section 2 has shown, however, that while there are linguistic tests which set CPPs apart from regular PPs, CPPs also exhibit considerable syntactic variation, and furthermore, that not all CPPs behave alike.

It is not surprising, therefore, that the corpus-based, statistical, methods discussed in sections 3 and 4 have only limited succes in identifying known CPPs. The inherent difficulty of the task is confirmed by the results of the evaluation in section 5, which showed that human judges also show considerable disagreement when asked to classify potential P-NP-P patterns as CPPs. Our goal of providing a more exhaustive list of CPPs was reached to some extent, as a (small) number of new CPPs were identified.

We see a number of ways in which one could extend or improve on the results presented in this paper. First of all, one might consider using a larger and more balanced corpus than the 16 million word *Volkskrant 97* corpus. Second, more genuine CPPs can perhaps be found by simply asking human judges to take more potential candidates into consideration. Finally, more agreement between judges can probably be established by providing instructions which to some extent reflect the difficulty of the task and which also take into account the fact that some CPPs may be more rigid than others.

Acknowledgements

We would like to thank two anonymous reviewers for their remarks, as well as the human judges for their help with the evaluation. This research was carried out within the PIONIER Project *Algorithms for Linguistic Processing*, funded by NWO and the University of Groningen (see `www.let.rug.nl/~vannoord/alp`).

References

Berghmans, J.(1994), Wotan, een automatische grammatikale tagger voor het Nederlands. Masters Thesis, Dept. of Language and Speech, Katholieke Universiteit Nijmegen.

Bouma, G., van Noord, G. and Malouf, R.(2001), Alpino: Wide-coverage computational analysis of Dutch, *Computational Linguistics in The Netherlands 2000*, Rodopi, Amsterdam.

Church, K. W. and Hanks, P.(1990), Word association norms, mutual information & lexicography, *Computational Linguistics* **16**(1), 22—29.

Corley, S., Corley, M., Keller, F., Crocker, M. W. and Trewin, S.(2001), Finding syntactic structure in unparsed corpora: The Gsearch corpus query system, *Computers and the Humanities* **35**(2), 81–94.

Drenth, E.(1997), Using a hybrid approach towards Dutch part-of-speech tagging. Masters thesis, Computational Linguistics, Rijksuniversiteit Groningen.

Dunning, T.(1993), Accurate methods for the statistics of surprise and coincidence, *Computational linguistics* **19**(1), 61—74.

Geerts, G. and Heestermans, H. (eds)(1992), *Van Dale Groot woordenboek der Nederlandse Taal*, Van Dale Lexicografie, Utrecht-Antwerpen.

Geerts, G., Haeseryn, W., de Rooij, J. and van den Toorn, M.(1984), *Algemene Nederlandse Spraakkunst*, Wolters-Noordhoff, Groningen.

Lin, D.(1998), Extracting collocations from text corpora, *First workshop on computational terminology*, Montreal, Canada.

Lin, D.(1999), Automatic identification of non-compositional phrases, *Proceedings of ACL-99*, University of Maryland, pp. 317–324.

Manning, C. D. and Schütze, H.(1999), *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

Moortgat, M., Schuurman, I. and van der Wouden, T.(2000), CGN syntactische annotatie. Internal Project Report Corpus Gesproken Nederlands, see http://lands.let.kun.nl/cgn.

Paardekooper, P. C.(1962), Voorzetsel-uitdrukkingen, *Nieuwe Taalgids* **55**, 3–9.

Paardekooper, P. C.(1973), Grensproblemen bij v-z-uitdrukkingen, *Nieuwe Taalgids* **66**, 137–145.

Smadja, F.(1993), Retrieving collocations from text: Xtract, *Computational Linguistics* **19**(1), 143–177.

van Halteren, H., Zavrel, J. and Daelemans, W.(2001), Improving accuracy in word class tagging through the combination of machine learning systems, *Computational Linguistics* **27**(2), 199–230.