A corpus-based approach to the acquisition of collocational prepositional phrases

M. Begoña Villada Moirón and Gosse Bouma
Alfa-informatica
Rijksuniversiteit Groningen
Postbus 716
NL-9700 AS Groningen
The Netherlands
m.b.villada@let.rug.nl, gosse@let.rug.nl

Abstract

Collocational prepositional phrases in Dutch 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 dictionary as multi-word prepositional phrases, we extract all instances of the relevant pattern from a corpus. Next, we introduce a number of statistical tests 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

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

> *ten opzichte van* (`with respect to'), *in tegenstelling tot* (`as opposed to'), *in verband met* (`in connection with')

In Dutch linguistics such expressions are known as *voorzetsel-uitdrukkingen* [Paardekooper 62]. Here, we will refer to them as *collocational prepositional phrases* (CPPs). In section 2 we argue that some but not all CPPs can be analyzed as multiword units.

In section 3 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 a dictionary. Section 4 discusses the evaluation of potential CPPs not included in this list by human judges.

2 Linguistic Properties

We propose linguistic diagnostics that distinguish a fixed type of collocational PPs from a more flexible intermediate type of expressions. Most of these tests were already applied by [Paardekooper 62].

1. **Restricted functionality as complements**: Verbs that select for a prepositional complement whose preposition matches the initial preposition in the phrases at stake fail to admit collocational phrases as instantiations of their prepositional complement.
2. **Non-substitutability**: The noun inside the phrase cannot be replaced by a synonym.
3. **Idiosyncratic prepositions and nouns**: presence of inflected nouns (*opzichte*) or archaic prepositions (*te*) inside some phrases.
4. **Absence of a determiner**: NPs headed by a singular count noun fail to admit a determiner (*verband, tegenstelling*). However, some NPs allow a restricted set of determiners (*het kader, de hand*).
5. **Modification**: Once modification is added inside the NP, the special meaning disappears. A few cases admit certain adjectives (*in (scherpe) tegenstelling tot* 'in strong constrast with').
6. **Pronominal adverbs**: Combinations of a preposition and a pronoun are realized as an adverbial pronoun in Dutch. In some cases, the noun can be followed by such a pronoun (*in plaats daarvan*).
7. **Extraposition**: Dutch allows extraposition of PPs out of NPs and VPs. The PP introduced by the second preposition can be extraposed in some cases (*onder leiding staan van*) but not others.
8. **Optional complement**: The PP introduced by the second preposition can sometimes be removed without a change of meaning (*onder invloed*).

Non-substitutability, restricted modifiability and non-compositionality are often reported as properties exhibited by collocations [Manning & Schütze 1999]. Given the collocational properties of some of the phrases we propose to treat them as collocational prepositional phrases. Conditions 1 and 2 turn out to be the discriminating ones between compositional and collocational phrases. We analyse as totally fixed expressions those phrases that exhibit conditions 1, 3, and 4, and fail to satisfy condition 7. Expressions that satisfy these properties are formalized into a multi-word lexeme *prep NP prep* inserted in the lexicon. We favor a more flexible analysis for expressions satisfying conditions 6, 7, and 8. These expressions consist of a tuple *prep NP* inserted as a lexical unit in the dictionary.

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 [Haeseryn et al. 1997]. 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 applied 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. We used a corpus consisting of text from *de Volkskrant op CD-ROM*, 1997. The corpus consists of over 16 million words. The text was tagged with part-of-speech tags, using the WOTAN tagset [Drenth 1997].

We used Gsearch [Corley et al. 2001] to extract syntactic patterns from the corpus. Gsearch allows one to search for substrings matching expressions defined by a context-free grammar. Potential CPPs were defined as Prep BNP Prep patterns, where a BNP (*base NP*) consists of the initial (non-recursive) part of an NP up to and including the head. 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 listed in table 1.

1253	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

Table 1: Most frequent P-BNP-P patterns in the corpus.

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'), 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 idiomatic expression which functions as an adverb. Other sources of errors are larger idiomatic phrases which contain a substring matching P BNP P.

4 Statistical collocation tests

The simplest statistical test for finding collocations is mere co-occurrence frequency. 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 non-collocational bigrams. 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.

Three tests that are often used to determine whether two co-occurring words are potential collocations [Manning and Schütze 1999] are mutual information, the log-likelihood score and Pearson's χ^2 test.

The tests for identifying collocations all assume that collocations are bigrams. 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 data-set, we assumed that either P_1 BNP forms a unit or that BNP P_2 forms a unit.

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. (All test results were collected using Ted Pedersen's Bigram Statistics Package, http://www.d.umn.edu/~tpederse/code.html). This results in two ranked lists of bigrams. The final rank of a pattern was determined on the basis of the sum of the ranks assigned in the two bigram-sets.

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 [van Dale 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.

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 occur less than 10 times, and some do not occur at all.

Test	*Freq*	*N*	*N=100*	*N=300*	*All*
mi	≥ 10	2084	23	39	77
ll	≥ 10	2084	53	67	77
χ^2	≥ 10	2084	52	69	77
mi	≥ 40	317	47	67	67
ll	≥ 40	317	53	65	67
χ^2	≥ 40	317	55	65	67
raw freq		248683	50	65	84

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

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. The mutual information test performs poorly with sparse data even if large corpora are available and a frequency cut-off is used. Using a frequency threshold of 40 improves the results 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.

We also performed experiments with mutual information and χ^2 adjusted to trigrams. This allows us to compute results for the P_1 BNP P_2 trigrams directly. The results did not improve on the results presented above, however.

5 Human evaluation

Evaluation of the coverage of the statistical tests used in CPPs extraction is difficult. The validation data is rather scarse and furthermore, extraction of a complete list of CPPs from contemporary dictionaries is not straightforward. With the twofold purpose of enlarging the validation data and, of measuring the performance of the statistical tests we carried out a human evaluation experiment. Three human judges manually determined which of the extracted collocation candidates should be considered true CPPs.

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. In a previous evaluation experiment, we had elaborated a list of collocational PPs that were manually checked against the Van Dale dictionary. This list consisted of true CPPs, and prepositional phrases that either form part of a larger fixed expression (*van tijd tot*), or instantiate a fixed complement inside an idiom or support verb construction (*onder leiding van*). We will refer to this list as 'provisional Van Dale list'. To make the judges' task easier, the extracted candidates also included in the 'provisional Van Dale list' were removed except from 10 test items such that, 4 were true CPPs and 6 were PPs part of a support verb construction. 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 extracted collocation candidates.

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; (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.
The results are illustrative of how difficult the task turns out to be. Only 9.44% of the candidate expressions were identified as good CPPs by at least two judges. The list is given below:

> *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, only 12 (6.8%) expressions constitute new instances of CPPs. Judges disagreed over 5 of the test items; one judge claimed that they were not true CPPs. No significant difference can be observed between the true positives extracted by the log-likelihood score and the raw frequency test.

6 Conclusions

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. 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 Van Dale, with only slightly better results than using raw frequency. Finally, human evaluation of a list of potential CPPs shows that the task of identifying such items is very hard. There was little agreement between judges, even on test items included from the list extracted from the Van Dale dictionary.

References

[Corley et al. 2001] Corley, S., M. Corley, F. Keller, M. W. Crocker, & S. Trewin, 2001. Finding syntactic structure in unparsed corpora: The Gsearch corpus query system, in: *Computers and the humanities, 35(2).*

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

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

[Haeseryn et al. 1997] Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M. van den Toorn, 1997. *Algemene Nederlandse Spraakkunst*, Wolters-Noordhoff, Groningen.

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

[Paardekooper 1962]Paardekooper, P. C., 1962, Voorzetsel-uitdrukkingen, in: *Nieuwe Taalgids,* 55.

[Paardekooper 1973]Paardekooper, P. C., 1973. Grensproblemen bij v-z-uitdrukkingen, in: *Nieuwe Taalgids,* 66.