

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

M. Begoña Villada Moirón and Gosse Bouma

Alfa-Informatica

Rijksuniversiteit Groningen

The Netherlands

`{M.B.Villada,G.Bouma}@let.rug.nl`

NLP tasks such as parsing and language generation require that linguistic expressions with irregular syntax and semantics be treated as multi-word lexical units inserted in the lexicon. An appropriate description of the syntactic and semantic information of idiosyncratic phrasal lexemes may improve performance in those systems involved in NLP applications (machine translation (MT), information retrieval (IR), dialogue systems, etc.).

Dutch provides us with many expressions that exhibit the pattern [Prep NP Prep]. In syntax, the middle NP may ban modification, quantification or exhibit idiosyncratic morphology. In semantics, it turns out difficult to replace the noun in the NP by a synonym. These and other properties suggest that these expressions share properties with collocational phenomena.

In this paper we aim at identifying an accurate way to analyze Dutch collocational phrases and second, at testing and establishing previously proposed data-driven methods to identify them in corpora.

## 1 Introduction

Most recent syntactic theories ‘project’ syntactic structure from the lexicon (Briscoe and Carroll, 1997). Parsing and language generation systems are informed by large constraint-based grammars whose rules license potential syntactic combinations of lexical entries. The accuracy of the lexical information propagates into accuracy and efficiency of the recognizer or generator. Furthermore, in a larger domain, that accuracy will influence the performance of systems employing the grammar. Undoubtedly, the lexicon plays a crucial role in current NLP work.

Expressions with idiosyncratic syntax and semantics should be treated as complex lexical units inserted in the lexicon. Among others, Breidt, Segond, and Valetto (1996) proposed idioms, phrasal verbs, separated particle verbs, lexical and grammatical collocations and compounds. Thus, we argue that treating the irregularities of collocational PPs in lexica (as opposed to in the grammar) facilitates and improves the processing in NLP recognition and generation tasks.

We concentrate on a particular type of Dutch expression of the form [Prep NP Prep] that exhibits rigid syntax and/or a non-totally compositional semantics. Examples are given in (1).

- (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 occasion of'), *te midden van* ('amidst'), *in het kader van* ('in the framework of'), *aan de hand van* ('on the basis of')

In NLP generation tasks the grammar should license these phrases in those constellations that preserve their *metaphorical* meaning. To avoid the split of a collocational phrase it has often been proposed that such phrases should be entered as multi-word lexemes in lexica. However, the syntactic behavior of Dutch [Prep NP Prep] phrases is far from being uniform. The decision to be made is how to best treat this type of expressions: as a multi-word lexeme [Prep NP Prep] (fixed collocation), an intermediate type of phrase [[Prep NP ] Prep] or a totally compositional phrase [Prep [ NP [ N [PP [Prep ... ]]]]].

In section 2 we argue for two different types of collocational prepositional phrases that motivate the insertion of two different multi-word lexemes in the computational lexicon. In section 3 we describe and evaluate the extraction method and statistical tests applied to acquire collocational prepositional phrases. Finally, section 4 reports the conclusions of our investigation and future work.

## 2 Two types of collocational PPs

Totally compositional prepositional phrases pose no problems since they are made up out of individual lexical entries combined following regular grammar rules. 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 (1962).

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 contrast 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, op bezoek gaan bij*) but not others (*\*dat ik geen beslissingen op basis neem van geruchten*).
8. **Optional complement:** The PP introduced by the second preposition can sometimes be removed without a change of meaning (*onder invloed*) but not systematically (*\*in plaats, \*in tegenstelling*).

Non-substitutability, restricted modifiability and non-compositionality are often reported as properties exhibited by collocations (Manning and Schütze, 1999, p.184). 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 noun Prep ] inserted in the lexicon. On the other hand, we favor a more flexible analysis for those expressions satisfying conditions 6, 7 and 8. These expressions consist of a tuple [ Prep NP ] inserted as a lexical unit in the dictionary.

To summarize, two types of expressions were proposed. Rigid syntactic distribution of the fixed phrases can be accounted for given that the three adjacent words in the collocational PP

```

dp -->      det   bnp -->  dp ap* noun
dp -->      poss  bnp -->      ap* noun
ap  -->  adv* adj

```

Figure 1: Definition of non-recursive NPs

constitute a lexical unit. Optional complementation, extraposition and pronominalization shown by some phrases rule out the treatment of the phrase as a lexical unit and therefore, it is best treated as a [ Prep NP ] lexical unit.

### 3 Acquisition

We described what features characterize the two proposed types of collocational PPs. Next, we discuss how to automatically acquire such phrases. Automatically acquiring idiomatic expressions implies the need to either access electronically available lexica or to apply techniques that facilitate the automatic extraction of idiomatic expressions from corpora (Krenn, 2000a). Paardekooper (1973) and Geerts et al. (1984) propose a list of 86 *voorzetseluitdrukkingen* (‘idiomatic prepositional phrases’) in Dutch, however, we believe that more than 86 phrases exist. Here, we explore the usefulness of data-driven methods for the acquisition of Dutch collocational PPs.

**Setup** The 1997 CD-ROM version of the newspaper *de Volkskrant* was used as data. The corpus consists of over 16M words and over 1M sentences. The chosen methodology requires a little a priori grammatical knowledge (Brent, 1993). The corpus was tagged by a part-of-speech tagger with the WOTAN tagset described in van Halteren, Zavrel, and Daelemans (2001) and originally proposed by Berghmans (1994). Tagging was performed automatically, using a Brill-tagger for Dutch (Drenth, 1997). The accuracy of the tagger is around 95%.

Extraction of instances of the chosen pattern was done by using a corpus query tool named Gsearch (Corley et al., 2001). Gsearch allows the extraction of strings that match a user’s query from a part of speech tagged corpus.<sup>1</sup> The tool uses a context-free grammar defined by the user, a parser and the user’s query to search through the tagged corpora. Figure 1 shows the definition of a non-recursive ‘base’ NP.

Preprocessing with a tool like Gsearch filters out unwanted data to extract a more reliable data set. We set as constraint that the words in the string be adjacent to each other without crossing

---

<sup>1</sup>But also raw text.

sentence boundaries. We believe the candidate data to be syntactically, rather homogeneous (Evert and Krenn, 2001). However, errors may still surface in the datasets due to tagging mistakes or to the fact that the second preposition may belong to a different constituent.

The output of Gsearch are all found instances of the pattern [ prep b(ase)np prep ] in the corpus. All extracted instances were sorted and assigned frequencies. A total of 285.027 [ prep bnp prep ] strings were extracted instantiating 163K different types (137K strings occur only once, 2333 strings occur at least ten times).

**Statistical model** Once the frequency of the candidate collocates was computed, we used Ted Pedersen’s Bigram Statistic Package in order to compute the log-likelihood, mutual information and  $\chi^2$  score of each bigram.<sup>2</sup>

The length of the extracted strings varies from 2 to 4 or sometimes more words. Consequently, standard statistical tests needed to be adjusted since they are usually applied to bigrams or trigrams. In our case, the extracted data sets were accordingly formatted and treated as bigrams. We treated each string as a bigram  $(w_1, w_2)$ . As an example, *in tegenstelling tot* allows two possible bigram combinations: either  $w_1$ =in and  $w_2$ =tegenstelling\_tot or  $w_1$ =in\_tegenstelling and  $w_2$ =tot. In a more general way, each string is represented by two possible bigrams: ((P NP) P) and (P (NP P)). We applied the three statistical tests to both bigram combinations for each extracted string and then, we added up the ranks of the two partial bigrams. A frequency threshold of 10 was used to discard very low frequency data.

**Results** For a preliminary evaluation we collected the 100, 300, 1000, 2000 and 100000 best ranked collocation candidates result of applying the mutual information (MI) (Church and Hanks, 1990), log-likelihood (ll) (Dunning, 1993),  $\chi^2$  test and co-occurrence frequency for two different frequency thresholds (10 and 40). The abovementioned *n-best* lists were compared to the list of 86 *voorzetseluitdrukkingen* (‘idiomatic prepositions’) proposed in the ANS (Geerts et al., 1984)(See Figure 2).

Mutual information (MI) applied to candidate strings with a frequency threshold of 10 performs worse than the other statistical tests in general. Only if we consider 2000 best ranked candidates can we compare the results of MI to the other tests. However, the accuracy of all tests decreases in smaller *n-best* lists. The accuracy of  $\chi^2$  is slightly lower than that of log-likelihood with a low frequency threshold. A higher frequency threshold of 40 leads to improved results of all the

---

<sup>2</sup>The Bigram Statistic Package is available at <http://www.d.umn.edu/~tpederse/code.html>.

N-BEST=		100	300	1K	2K	100K
mi	Freq $\geq$ 10	2	13	55	68	
ll	Freq $\geq$ 10	37	51	62	68	
$\chi^2$	Freq $\geq$ 10	29	50	62	68	
mi	Freq $\geq$ 40	32	51			
ll	Freq $\geq$ 40	37	51			
$\chi^2$	Freq $\geq$ 40	37	51			
raw freq		35	48	63	67	74

Figure 2: Accuracy results for *n-best* [Prep NP Prep] strings.

tests, with MI still doing worse than log-likelihood and  $\chi^2$  in the first part of the data. Log-likelihood and raw frequencies lead to the best results overall. A close look at the top 100 highly ranked candidates returned by the LL test and mere raw frequency counts reveals that, the strings returned by the LL test and not by the raw frequency test are better collocational candidates (qualitatively speaking) than the strings returned by raw frequency.

## 4 Conclusion and future work

The work reported in this paper proposes a dual classification of collocational prepositional phrases in Dutch. We have shown that standard association measures, in particular log-likelihood helps to extract fixed collocational PPs.

In the future, we want to treat the extracted strings as trigrams and apply the statistical tests to them. The question we want to answer is whether the accuracy of statistical tests increases that way. We also aim at applying other statistical tests such as the *Dice* coefficient and phrase entropy (Krenn, 2000b).

*Phrase Entropy* calculates the entropy observed in (prepositional) phrases where potential collocation candidates may occur. Krenn (2000b) extracts triples consisting of [Prep Noun Verb] and measures the rigidity exhibited by the tuple [Prep Noun] focusing on modification, quantification, etc. Krenn’s purpose is closely comparable to our aim of distinguishing which strings in our data set constitute clearly fixed collocational PPs, intermediate collocational PPs or compositional PPs. Given the good results reported by Krenn (2000b), we believe that this method serves well to determine the degree of variation inside potential collocations. In fact, preliminary experiments done with phrase entropy to discover true collocations and their syntactic variation, return strings which are strong collocation candidates and are encouraging. In future work we plan to investigate whether phrase entropy can serve to establish natural classes of collocational phrases with regards

to the syntactic variation observed inside the NP. We also want to implement collocational PPs in the Alpino grammar<sup>3</sup> and carry out human evaluation.

## References

- Berghmans, J. 1994. Wotan, een automatische grammatikale tagger voor het nederlands. Masters Thesis, Dept. of Language and Speech.
- Breidt, E., F. Segond, and G. Valetto. 1996. Local grammars for the description of multi-word lexemes and their automatic recognition in texts. In *COMPLEX96*, Budapest.
- Brent, Michael R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational linguistics*, 19(2):243—262.
- Briscoe, Ted and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL conference on applied Natural Language Processing*, pages 356–363, Washington, D.C.
- Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22—29.
- Corley, S., M. Corley, F. Keller, M. W. Crocker, and S. Trewin. 2001. Finding syntactic structure in unparsed corpora. *Computers and the Humanities*, 35(2):81—94.
- Drenth, Erwin W. 1997. Using a hybrid approach towards dutch part-of-speech tagging. Master's thesis, Alfa-Informatica, University of Groningen.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61—74.
- Evert, Stefan and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.
- Geerts, G., W. Haeseryn, J. de Rooij, and M.C. van den Toorn. 1984. *Algemene Nederlandse Spraakkunst*. Wolters-Noordhoff, Groningen.

---

<sup>3</sup>The Alpino computational parser and grammar are being developed in the framework of Gertjan Van Noord's PIONIER project *Algorithms for Linguistic Processing*. More information at <http://odur.let.rug.nl/~vannoord/alp>

- Krenn, Brigitte. 2000a. CDB – a database of lexical collocations. In *Proceedings of the 2nd International Conference on Language Resources & Evaluation*, Athens.
- Krenn, Brigitte. 2000b. Collocation mining: Exploiting corpora for collocation identification and representation. In *Proceedings of KONVENS 2000*, Ilmenau, Germany.
- Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.
- 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.
- van Halteren, Hans, Jakub Zavrel, and Walter Daelemans. 2001. Improving accuracy in word class tagging through the combination of machine learning systems. *Computational Linguistics*, 27(2):199–230.