A corpus investigation of PP-fronting in Dutch

Gosse Bouma

Rijksuniversiteit Groningen

Abstract

A long-standing discussion in Dutch syntax concerns the question whether PP dependents of a noun may be fronted. Although examples which apparently illustrate this pattern can be easily found, it is difficult to come up with linguistic arguments which show once and for all that the fronted PP is actually a dependent of a noun, and not of a verb. In this paper, we investigate to what extent corpus data can be used to decide on this matter, and conclude that the data suggest that the PP is a dependent of the verb.

1 Introduction

A long-standing discussion in Dutch linguistics is concerned with the status of the PP in sentences like (1). In (1-a), a full PP appears in sentence initial position, and in (1-b), the initial pronoun is interpreted as the object of the preposition *naar*. The PP can be seen as a dependent of the noun *onderzoek* or of the main verb.

(1)   a.   **Naar** deze pijnlijke gebeurtenis wordt nu   nader **onderzoek**
            Into   this painful event   is   now further research
            gedaan.
            done
            Further research on this painful event is now done
   b.   Daar is echter   nauwelijks **onderzoek naar** verricht
            There is however hardly   research   into carried out
            However, hardly any research on this has been carried out

Bach and Horn (1976) argued that extraction from Dutch NPs should not be possible, and thus, that neither fronting of an NP-internal PP nor fronting of an object of a PP contained within an NP, should be possible. At first sight, the examples in (1) clearly seem to falsify this claim. However, arguments that the PP in (1-a) and the preposition in (1-b) are actually (heading) a dependent of the verb have also been put forward.

The N+PP analysis is intuitively plausible, as there seems to be a strong semantic relation between the noun and preposition. Furthermore, N+PP may precede the finite verb in main clauses, and thus clearly forms a constituent in some cases. Also, when N is preceded by certain definite determiners, fronting of the PP is almost impossible. This suggests PP-fronting is subject to a constraint on extraction from

NP, something which seems highly problematic for a V+PP analysis. The V+PP analysis, on the other hand, is supported by the fact that PP-fronting seems to occur only with certain verbs. Furthermore, some nouns clearly select a PP, but do not allow fronting of this PP.

The proper analysis of examples like those in (1) has been the topic of a heated discussion (in Klein and van den Toorn (1977, 1979), and Kooij and Wiers (1979), among others). Coppen (1991) reviews most of the arguments presented earlier, and concludes that the most convincing arguments point towards a V+PP analysis.

Although many of the theoretical assumptions which played a role in the original discussion have changed, the question whether fronting of PP-dependents from an NP is possible, is still relevant. The syntactic annotation of the Corpus of Spoken Dutch (CGN) (Moortgat, Schuurman and van der Wouden 2000), for instance, adopts an N+PP analysis for the following type of example:

(2) a. daar heb ik helemaal geen **zin in**
         there have I totally no desire for
         I have desire for that at all
     b. m-hu wij uh zullen daar goede **nota van** nemen
         Uhm we uhm shall there good notice of take
         We shall take good notice of that

The Alpino-grammar (van der Beek, Bouma and van Noord 2002), on the other hand, does not allow extraction out of NPs, and thus has opted for the V+PP analysis. As one of the design goals of the Alpino system was to produce output compatible with CGN, it seems that either the Alpino-grammar should be modified, or the CGN-annotation guidelines need to be reconsidered.

In this paper, we investigate to what extent corpus data can be used to decide on this matter. A corpus-based approach seems appropriate for at least two reasons. First, the claim that certain determiners block PP-fronting as well as the claim that PP-fronting occurs only with certain verbs, can be verified using corpus data. Second, there has been considerable disagreement between authors on the status of examples that were crucial in arguing for one or the other position. Examples marked with a star in one paper were considered to be acceptable by authors arguing for a different analysis.[1] Coppen notes that the examples in his paper show varying acceptibility, and that linguistic intuitions with respect to these data even seem to change over time.

In section 2, we describe the construction and annotation of the corpus. Next, we investigate the role of the verb in PP-fronting. We find that some verbs are far more frequent in PP-initial sentences containing the relevant N+P combination than in general sentences with this N+P combination. In section 4, we look at the determiner preceding the noun which apparently selects for the PP. There is a strong preference for indefinite determiners in PP-initial sentences, while possessive pronouns and genetive NPs are almost absent. We argue that this is not necessarily evidence for a constraint on extraction from NP. Instead, it seems that the verbs

---

[1] I.e. see Klein and van den Toorn (1977, p. 432), Klein and van den Toorn (1979, p. 105) and Kooij and Wiers (1979, p. 488).

which easily admit PP-fronting, also have a preference for indefinite NP-dependents. In section 5, we observe that PPs may also be included in relatives modifying the noun, as in:

(3)   de enige **relatie** die er   **tussen** haar en   Van Kooten bestaat
      the only relation that there between her   and Van Kooten exists
      the only relation that exists between her and Van Kooten

This seems highly problematic for an N+P analysis.

In section 6, we note that a number of patterns which have been used as arguments for a particular analysis, are practically absent in the corpus.

The data suggest that the verb plays an important role in the question whether PP-fronting is possible, and that the PP is thus best seen as a dependent of the verb. In section 7, we note that many of the nouns combining with a prepositional complement in CGN behave similar to the examples we investigated and that many of the frequent N+P+V combinations found in PP-initial sentences have been identified as phrasal verbs taking a prepositional complement by other authors.

We conclude therefore that the corpus data suggest that the V+PP analysis is more likely than the N+PP analysis, and that these expressions are best analyzed as phrasal verbs involving a prepositional complement.

2   Corpus Construction

We used the newspaper sections of the Twente News Corpus[2] (TWNC) as our initial corpus. The corpus contains text from major Dutch newspapers in the period 1994-2001, and has a size of approximately 300 million words. We believe that, at least for the phenomenon we are interested in, this corpus is representative for Dutch in general.

2.1   Selection of relevant N+P combinations

Sentence-initial PPs in many cases are clearly dependents of a verb, and in many other cases, could equally well be seen as a dependent of a verb or a noun. The discussion referred to in the introduction has focused on N+PP combinations displaying a strong semantic relation between the noun and the PP. Our first goal was to identify a number of such nouns in the corpus.

In principle, one might try to find nouns selecting for a PP by looking at frequent N+P bigrams in the corpus. However, raw frequency by itself is not a very good indication of the fact that the nouns actually selects for a PP (rather than just co-occurring regularly with a modifier PP headed by this P). Therefore, we used the *log-likelihood-test* of Dunning (1993) to find promising candidates. From the list of N+P bigrams with strong collocational properties according to this test, we manually selected 20 bigrams as suitable candidates for our research.[3] Three

---

[2]`wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html`

[3]Highly ranked bigrams which we discarded among others were parts of names (*ministerie van (ministry of)*), and parts of complex prepositions (*(in) tegenstelling tot ((as) opposed to)*).

behoefte aan	*need for*
belangstelling voor	*interest in*
bezwaar tegen	*objections against*
contact met	*contact with*
discussie over	*discussion about*
gebrek aan	*lack of*
gesprek over	*conversation about*
informatie over	*information about*
kritiek op	*critique on*
onderzoek naar	*investigation into*
protest tegen	*protest against*
relatie tussen	*relation between*
sprake van	*talk of*
twijfel aan	*doubt about*
verhaal over	*story about*
verschil tussen	*difference between*
vraag naar	*demand for*

Table 1: Selected N+P collocations

bigrams turned out to be very infrequent in PP-initial sentences, or gave rise to a large number of errors in automatic syntactic analysis (see below). These were discarded, and thus we used 17 bigrams (given in table 1) for further research.

2.2 Construction of Analyzed Subcorpora

Using the N+P collocations in table 1, we constructed a syntactically analyzed general corpus as follows:

1. From the TWNC, we initially extracted per N+P collocation maximally 10.000 sentences containing both N and P. 155.000 sentences were selected in total (as some bigrams do not occur 10.000 times).

2. The *dependency tree* for each sentence was computed using the Alpino-system.[4].

3. From the syntactically analyzed sentences, we selected[5] the examples which satisfied either one of the criteria below:

   - The NP headed by N and the PP headed by P are both dependents of the same verb, or
   - The PP is a dependent of N, and the NP headed by N is a dependent of a verb (i.e. and not part of a PP or other non-verbal constituent).

---

[4]`www.let.rug.nl/~vannoord/alp`

[5]using the XML-tool for searching dependency trees described in Bouma and Kloosterman (2002).

Syntactic analysis is important for our purposes for two reasons. First, not all sentences containing N and P are actually valid instances of the pattern we are interested in (P might be heading a PP containing an NP headed by N, or the PP might be part of another NP, for instance). Second, we want to investigate which verbs co-occur with these N+P collocations. Therefore, we must be able to determine which verb actually selects for the NP headed by N. As we are interested in investigating the status of the PP, we need to consider both the case where the PP is analyzed as a dependent of N and the case where the the PP is analyzed as a dependent of V. The selection step in 3 retrieved 56.000 examples containing between 2.000 and 4.000 examples per N+P collocation on average.

The general, 56.000 sentence, corpus described above was contrasted with a second corpus, which consisted of PP-initial sentences only. This 'P1' corpus was constructed by exhaustively searching the TWNC for sentences containing both N and P, but where P was also the first word in the string. Initially, the P1 corpus consisted of almost 10.000 examples. After syntactic analysis and selection, this was reduced to approximately 5.000 cases. The corpus is dominated by *sprake van*, which occurred no less than 3.872 times in P1. For the other collocations, between 30 and 326 examples were found per collocation.

One might wonder whether automatic analysis is sufficiently reliable to create a representative corpus. Even though automatic analysis is not completely error-free,[6] the effect it has on the task we are interested in seems small. Automatic analysis does reliably filter cases where the NP is not a dependent of a verb, or where the NP and PP are dependents of a different verb. Also, the main verb selecting NP or both NP and PP is identified reliably. Nevertheless, errors do sometimes occur, and thus we did manually inspect many of the results found in the experiments below, especially cases involving small numbers.

3 The role of the verb

The idea that fronting of a PP is possible only with certain verbs, has been used as argument for the V+PP analysis. A problematic aspect of this argument is the fact that linguistic intuition alone does not seem to be sufficient to draw the line between those verbs that do allow PP-fronting and those that do not. In this section, we investigate whether corpus-data provide a clearer answer.

Using the information provided by automatic syntactic analysis, as described in the previous section, we we counted how often a specific verb occurs with a specific N+P collocation in the general and in the P1 corpus. In particular, we counted the verbs with a dependent NP headed by N and containing a PP or with a dependent NP headed by N and a dependent PP headed by P. To avoid inclusion of (verbs functioning as) auxiliaries and modals, verbs with a VP-dependent were excluded. If PP-fronting is determined by the verb, only a limited number of verbs should be found in the P1 corpus, and in P1 these verbs should occur more frequently than in the general corpus. In table 2, we present an overview of verbs found more than

---

[6]Malouf and van Noord (2004) report that the Alpino-system identifies dependency relations with an accuracy of 87.8% on a representative 500 sentence subset of the TWNC.

once in P1 and in the general corpus, for the first 7 N+P combinations on our list and the (idiomatic) *sprake van.*

Table 2 shows that the combination *behoefte aan* mainly occurs with *hebben* en *zijn* in P1, but the same is true in the general corpus. The verb *bestaan* appears significantly more often with *behoefte aan* in P1 than in general. The absolute number of occurrences of other verbs cooccurring with *behoefte aan* is very small, so the differences in distribution are not statistically significant. For allmost all N+P collocations we investigated, statistically significant differences in distribution can be observed for the most frequent verbs.

The verbs *hebben, zijn* and *bestaan* are special in that they seem to allow PP-fronting with almost all investigated N+P combinations. The role of *bestaan* is remarkable: this otherwise rather infrequent verb occurs frequently with 10 of the 17 investigated N+P combinations. There are also a number of verbs in P1 which clearly seem to form a phrasal verb with the N+P combination, e.g. *een gesprek voeren met (be engaged in a conversation with), informatie verstrekken over (provide information on), een onderzoek instellen naar (start an investigation into), een onderzoek loopt naar (an investigation is being carried out into), protest rijst tegen (protest is raised against), een verhaal gaat over (a story is about), een verhaal doet (de ronde) over (a story goes around about),* en *(er) zit een verschil tussen (there is a difference between).*

The V+PP analysis also predicts that for some verbs, PP-fronting should be impossible. This prediction is hard to test, as the absence of a verb in P1 might be due to lack of data. Nevertheless, in table 3 we provide a list of verbs missing in P1 which occur with more than 1% of the relevant N+P example sentences in the general corpus. All verbs listed for *gebrek aan* seem to resist PP-fronting. In other cases, fronting seems marked (*aan* NP *groeit er behoefte (for* NP *grows the demand), naar* NP *leidt/eist* NP *een onderzoek (into* NP, NP *demands an investigation), naar* NP *kondigt* NP *een onderzoek aan (into* NP, NP *announces an investigation), over* NP *nam* NP *alle twijfel weg (on* NP, NP *took all doubts away)*). For other verbs and N+P combinations, it seems that fronting is at least theoretically possible. The limited size of the P1 corpus might be the reason why these are absent in our data.

The corpus data clearly suggest that the verb plays a role in PP-fronting. The distribution of verbs in P1 and the general corpus shows large differences for most investigated N+P combinations. For frequent verbs, these differences are often statistically significant. Furthermore, there seem to be a number of verbs which easily combine with certain N+P combinations, but which do not allow PP-fronting.

4   The role of the determiner

It has been argued that so-called *specified subjects* within the NP block extraction:

(4)   a.   Over   Piet herinnerde   hij zich   een verhaal.
              About Piet remembered he REFL a   story
              He remembered a story about Piet
        b. *Over   Piet herinnerde   hij zich   Jans verhaal.
              About Piet remembered he REFL Jan's story

		P1	Gen
behoefte aan N=		241	5699
hebben (*have*)		56.0	53.8
zijn (*be*)		26.6	24.5
bestaan (*exist*)	●	11.2	4.6
blijken (*turn out to be*)		1.2	0.5
toenemen (*increase*)		0.8	1.7
blijven (*remain*)		0.8	0.6
belangstelling voor N=		326	5124
zijn (*be*)	●	37.4	23.5
bestaan (*exist*)	●	27.0	5.2
hebben (*have*)	●	19.3	28.4
tonen (*show*)		6.4	7.4
komen (*come*)		1.8	1.1
blijken (*turn out to be*)		1.2	0.8
verwachten (*expect*)		0.9	0.4
ontstaan (*come up*)		0.6	1.1
blijven (*remain*)		0.6	0.5
bezwaar tegen N=		163	3772
hebben (*have*)		38.0	35.9
maken (*make*)		32.5	36.6
aantekenen (*register*)		12.3	13.0
bestaan (*exist*)	●	8.6	1.1
zijn (*be*)		4.9	8.0
aanvoeren (*raise*)	●	1.2	0.1
contact met N=		304	3645
hebben (*have*)	●	56.2	29.3
zijn (*be*)		9.2	7.1
houden (*keep*)		5.6	3.8
zoeken (*search*)	●	4.9	13.2
onderhouden (*maintain*)		3.9	4.3
leggen (*lay*)		3.6	5.0
opnemen (*take up*)	●	3.6	14.2
maken (*make*)		2.6	2.9
krijgen (*get*)		1.6	2.3
verbreken (*break*)		1.3	0.9
willen (*want*)		1.0	0.8
verliezen (*loose*)	●	1.0	3.2
herstellen (*reestablish*)		1.0	0.8
verlopen (*decrease*)		0.7	0.7
komen (*come*)		0.7	0.4

		P1	Gen
discussie over N=		137	3857
zijn (*be*)	●	42.3	15.3
voeren (*be engaged in*)	●	12.4	7.3
bestaan (*exist*)	●	12.4	0.7
woeden (*rage*)	●	5.1	2.4
ontstaan (*come up*)		5.1	3.7
gaan (*go*)		4.4	7.4
hebben (*have*)		3.6	2.0
losbarsten (*burst out*)		2.9	1.9
ontbranden (*ignite*)		1.5	0.6
houden (*hold*)		1.5	0.6
gebrek aan N=		297	2574
zijn (*be*)	●	63.6	32.8
hebben (*have*)	●	26.9	9.1
bestaan (*exist*)	●	2.7	0.6
liggen (*lay*)	○	1.0	0.3
heersen (*rule*)		1.0	0.5
lijken (*seem*)		0.7	0.5
gesprek met N=		82	2067
hebben (*have*)		39.0	34.3
voeren (*be engaged in*)	●	20.7	12.8
zijn (*be*)		6.1	7.4
worden (*become*)	●	3.7	0.6
verlopen (*develop*)	●	3.7	0.5
aangaan (*engage in*)		3.7	4.0
komen (*come*)		2.4	0.8
volgen (*follow*)		2.4	2.2
informatie over N=		63	3492
geven (*give*)	●	30.2	16.8
verstrekken (*provide*)	●	17.5	4.8
hebben (*have*)		7.9	4.4
zijn (*be*)		7.9	5.6
krijgen (*get*)		4.8	9.7
vinden (*find*)		3.2	3.7
verschaffen (*provide*)		3.2	3.3
ontbreken (*lack*)		3.2	0.5
sprake van N=		3872	5128
zijn (*be*)	●	98.1	98.6
lijken (*seem*)	●	1.3	0.9
blijken (*turn out to be*)		0.6	0.4

Table 2: Distribution of verbs for several N+P collocations. Differences marked with ● (○) are significant according to the chi-square test at p=0.05 (p=0.10).

behoefte aan	N=	5699
onstaan (*come up*)	○	1.4
groeien (*grow*)	○	1.3
belangstelling voor	N=	5124
wekken (*wake*)	○	1.1
discussie over	N=	3857
beginnen (*start*)	○	2.7
brengen (*bring*)		1.9
aanzwengelen (*start up*)		1.8
volgen (*follow*)		1.0
aangaan (*engage in*)		1.0
krijgen (*get*)		1.0
gebrek aan	N=	2574
verwijten (*blame*)	●	8.7
compenseren (*compensate*)	●	2.4
opbreken (*stumble over*)	●	1.6
noemen (*mention*)	●	1.6
leiden (*lead to*)	●	1.5
vinden (*find*)	●	1.3
spelen (*play*)	●	1.3
hekelen (*criticize*)	○	1.2
worden (*become*)	○	1.0

kritiek op	N=	4077
toenemen (*increase*)		1.3
onderzoek naar	N=	3570
leiden (*lead*)	○	1.8
willen (*want*)	○	1.8
gelasten (*demand*)	○	1.5
eisen (*demand*)		1.2
aankondigen (*announce*)		1.1
twijfel over	N=	1714
uiten (*utter*)	○	2.5
uitspreken (*pronounce*)	○	2.3
wegnemen (*take away*)		1.9
groeien (*grow*)		1.6
verschil tussen	N=	3925
bedragen (*amount to*)	○	2.7
worden (*become*)		2.0
weten (*know*)		1.8
kennen (*know*)		1.5

Table 3: Frequent N-P-V-combinations in the general corpus, absent in P1. Differences marked with ● (○) are significant according to the chi-square test at p=0.05 (p=0.10).

An NP contains a specified subject if its determiner is a genitive NP or a possessive pronoun. The existence of a constraint like this would be a strong argument for the N+PP analysis.

In this section, we first investigate whether there is a relationship between the distribution of determiners and PP-fronting. Next, we discuss how the scarcity of genitives and possessive determiners in the P1 corpus might be explained.

4.1 Definite and indefinite NPs

In table 4, a comparison of the frequencies in P1 and general is made of the most common determiners preceding the relevant noun.[7] Table 4 suggests that the indefinite determiners *geen, veel* and *weinig* occur relatively frequently in P1, whereas the definite determiner *de/het* is relatively infrequent in P1.

We believe that the difference in distribution of determiners in P1 and the

---

[7] The idiomatic *sprake van* (*talk of*) was not included, as it is far more frequent than the other combinations, and would distort the results too much. However, in the general corpus we find that 80% of the examples consists of *NULL sprake van* whereas 20% is *geen (no) sprake van*. In P1 sentences this is 70% for *geen sprake van* and 30% for *NULL sprake van*.

	P1	**Gen**			**P1**	**Gen**
determiner N=	2.144	50.892	determiner N=	2.144	50.892
geen (*no*)	30.7	8.0	weinig (*few/little*)	3.8	0.7
NULL	27.7	31.8	enkele (*some*)	2.1	0.8
een (*a*)	14.4	16.5	meer (*more*)	0.8	1.0
veel (*many/much*)	7.7	2.1	minder(*less*)	0.6	0.2
de/het (*the*)	7.3	32.9			

Table 4: Frequency of determiners preceding the relevant noun in P1 and the general corpus.

general corpus can be explained to a large extent by the fact that the verbs in P1 and general have a very different distribution (as shown in the previous section). If we restrict our attention to N-P-V combinations that contain a verb which is relatively frequent in P1, we see that the definite determiner is much less frequent in the general corpus as well. This is illustrated in table 5.

The combination *verhaal vertellen over* is one of the few examples where the definite determiner is relatively frequent in the general corpus. In this case, the P1-data show an even stronger preference for the definite determiner: 25 out of the 42 cases of PP-fronting with *verhaal over* contain a definite NP.

The conclusion to be drawn from these data is that the preference for indefinite determiners in the P1 data correlates strongly with the preference for indefinite determiners in the general corpus, if one restricts attention to those verbs which are frequent in P1. It seems therefore that the differences in determiner distribution are for the most part a consequence of the differences in the distribution of the verbs in both corpora.

4.2   Possessive pronouns and genitive NPs

At first sight, the corpus seems to confirm the observation that PP-fronting requires an NP which does not contain a 'specified subject' in the form of a possessive pronoun or genitive NP. Table 4 does not contain any of these determiners. Genitives are in fact absent in P1, while possessives are scarce, and restricted to the N+P combinations *verhaal over* en *twijfel over*:

(5)   a.   Over die   worsteling gaat mijn verhaal
             about that struggle   goes my   story
             My story is about that struggle
        b.   Over   adverteren in de verzorgingssfeer heeft hij zijn twijfels
             About advertising in the health sector   has   he his doubts
             He has his doubts about advertising in the health sector

Only the phrase *twijfels hebben over* is relatively frequent in P1.

One might argue that the absence of genitives and the apparently highly restricted use of possessives, is evidence for the claim that PP-fronting is blocked

N+V+P	N=	determiners
behoefte hebben aan *have need for*	3001	NULL 60.4, geen 25.5, ...,de 1.0
behoefte zijn aan *be need for*	1051	NULL 71.6, een 10.5, geen 5.6, ..., de 2.1
behoefte bestaan aan *exist need for*	259	NULL 52.1, een 18.9 geen 11.6, de 5.8
belangstelling hebben voor *have interest in*	1343	NULL 70.3, geen 12.9 ..., de 0.5
bezwaar hebben tegen *have objection against*	1431	geen 53.9, NULL 34.2 ..., het 0.0
contact zoeken met *seek contact with*	462	NULL 93.9, geen 4.1 het 1.1
discussie zijn over *be discussion about*	257	NULL 36.6, de 16.3 geen 14.0, een 12.8
gesprek voeren met *be engaged in discussion with*	250	een 90, het 4.8 geen 1.6
informatie geef over *give information about*	552	NULL 76.4, geen 8.3 ...,de 1.6
onderzoek lopen naar *carry out research on*	96	een 77.1, het 18.8 NULL 3.1, geen 1.0
verhaal vertellen over *tell story about*	291	een 51.5, het 33.3 ..., geen 1.4
twijfel bestaan over *exist doubt about*	402	geen 41.8, NULL 38.6 de 1.5

Table 5: Frequency of common indefinite and definite determiners in the general corpus for frequent N-P-V-combinations in P1.

for certain NPs. It should be noted, however, that NPs introduced by a possessive pronoun or genitive are not very frequent in the general corpus either: 2.1% of the relevant NPs in the general corpus contains a possessive pronoun and 0.9% a genitive NP. Furthermore, those verbs which do occur with this type of NP seem to be highly infrequent in P1. The absence of NPs introduced by a genitive and the restricted possibilities for using possessive pronouns can therefore also be attributed to properties of the N+P+V combination as a whole.

5   PPs in relative clauses

It has been argued that pronominalization provides an argument for the V+PP analysis of PP-fronting. If a PP is a dependent of the noun, pronominalization of that noun requires the PP to disappear as well:

(6)   a.   Hij heeft artikelen tegen   die   stelling gelezen
            He has   articles   against that thesis   read

He has read articles against that thesis

b. *Hij heeft ze   tegen   die stelling gelezen
   He has   them against that thesis   read

If the PP can remain without giving rise to ungrammaticality, this shows that the PP can also be interpreted as a dependent of the verb

(7) a. Hij heeft alle boeken van Vestdijk gelezen
         He has   all books by Vestdijk read
         He has read all books by Vestdijk
     b. Hij heeft ze   van Vestdijk (allemaal) gelezen
         He has   them by Vestdijk all   read
         He has read them (all) by Vestdijk

As this seems to be the case for nouns that allow PP-fronting, it suggests that the PP in those cases is actually a dependent of the verb. A problem with this argument, from our point of view, is that pronominalization of the nouns investigated here is scarce, and hard to locate reliably, even in a syntactically analyzed corpus.

However, there is a related construction, involving relative clauses, which also provides evidence that the PP can be interpreted as a dependent of the verb. In relative clauses modifying the noun, the PP is sometimes clearly embedded in the relative clause:

(8) a. Een **bezwaar** dat je **tegen** deze boeken zou   kunnen
         An objection that one against these books   should can
         aanvoeren
         raise
         an objection that one might raise against these books
     b. de **belangstelling** die   Eduard **voor** het nazisme toonde
         the interest   which Eduard for   the nazism   showed
         the interest which Eduard showed for Nazism

For PPs which are unambiguously part of the NP (and which cannot be fronted) this is not possible:

(9) a. *een **demonstratie**, die   **tegen** de hoge werkdruk in   chaos
         a   demonstration   which against the high work-load into chaos
         ontaardde
         turned
     b. ***Tegen** de hoge werkdruk ontaardde een **demonstratie** in   chaos.
         against the high work-load turned   a   demonstration into chaos

Thus, the possibility of a PP to appear inside a relative clause is evidence for the fact that the PP can be interpreted as a dependent of the verb.

In the general corpus, for most of te N+P combinations we investigated, several examples can be found where the PP is included in a relative clause. Some more examples are given below.

(10) a. het laatste **gesprek**   dat ik **met** hem heb gehad
          the last   conversation that I with him have had
          the last conversation that I had with him
      b. de **informatie** die   hij **over** zijn patiënt heeft
          the information that he about his patient has
          the information that he has on his patient
      c. De **kritiek** die   hier **op** het boek wordt uitgeoefend
          the critique that here on the book is   offered
          the critique on the book which is offered here

For a few combinations (*twijfel over, vraag naar, verhaal over*), only exhaustive search in the Twente News Corpus provided us with some examples. For *sprake van* and *gebrek aan*, we did not find examples where the PP is unambiguously part of a relative clause.

The fact that most PPs which can be fronted also may occur within a relative clause seems problematic for a N+PP analysis. Under such an analysis, it seems that the relative pronoun would have to inherit the selection or subcategorization properties of the noun it modifies. Furthermore, a mechanism needs to be established which allows the PP to appear in a position non-adjacent to the relative pronoun (i.e. head-movement, remnant movement, or argument transfer from the pronoun to the verbal head). We believe the syntactic literature does not provide evidence for assuming that such processes are at work here.

6   Infrequent and Missing Patterns

As argument for the V+PP analysis, it has been suggested that there are word orders which seem incompatible with the idea that the PP is a dependent of a noun. In this section, we observe that these patterns are practically absent in even a large corpus of Dutch. However, the word order PP+NP is relatively frequent in the Dutch Mittelfeld, and seems problematic for an N+PP analysis.

One argument for the V+PP analysis has been the suggestion that, in analogy to PP-fronting, one also finds cases of NP-fronting, where the PP occupies a position in the 'Mittelfeld':

(11) Een **roman** heb ik **van** Vestdijk gelezen
      A   novel   have I of   Vestdijk read
      I have read a novel by Vestdijk

Such examples are practically absent in the general corpus. The most convincing cases are given below:

(12) a. Verder **contact** is er   **met** Den Haag niet geweest
          Further contact   is there with The Hague not been
          There has not been further contact with The Hague
      b. Dat **gesprek**   zou   vandaag ook **met** de ouders van de
          That conversation should today   also with the parents of   the

> mogelijk   misbruikte kinderen worden gevoerd
> potentially abused   children be   conducted
> *Today, that conversation with the parents of the potentially abused children shall be conducted*

    c. Meer **informatie** kunt u   er   **over** krijgen bij een notaris
        More information can   you there about get   at the notary
        You can get more information about this at the notary

A few other cases involving an NP containing the relevant noun in first position and a PP in the Mittelfeld, had to be discarded as false positives, as they most likely involved a verb selecting for a PP-complement (i.e. *gaan over (is about)*). Examples such as (13), which are more frequent, have to be discarded as well, as they may be the result of extraposing a PP-complement of NP in initial position.

(13)   Hoeveel   **behoefte** is er   in de toekomst nog **aan** de diverse
        How much need   is there in the future   still for   the various
        joodse zorg-, onderwijs- en welzijnsorganisaties, ...?
        Jewish care   education and wellfare organizations
        How much need is there in the future for the various Jewish care, education, and wellfare organizations,....?

The difference in frequency between PP-fronting and 'NP-fronting' is puzzling.

    Another argument for the V+PP analysis has been the claim that the NP and PP may be separated from each other within the Mittelfeld. In the general corpus, we did not find a single example of an NP-XP-PP word order, however. We found only 4 examples of PP-XP-NP word order, one of which is given below:

(14)   Gek   genoeg bestond er   **voor** zijn oorlogsfoto's   *tot   voor*
        Funny enough existed there for   his war photography until for
        kort   weinig **belangstelling**.
        recently little   interest
        Curiously, until quite recently there was little interest in his war photography

On the other hand, PP-NP orders, as in (15), are relatively common in the general corpus (with 10-50 examples per N-P combination, except for *protest tegen*, for which we found only a single example):

(15)   De Marokkanen hebben **aan** dergelijke groepsvorming   geen
        The Maroccans have   on   such   group formation little
        behoefte.
        interest
        the Maroccans have little interest in such group formation

Although this pattern seems equally problematic for an N+PP analysis as PP-XP-NP order, it has not been mentioned as such in the literature.

7 Related Work

As was noted in the introduction, the Corpus of Spoken Dutch (CGN) frequently annotates nouns as taking a prepositional complement. We collected all N+P combinations which are annotated at least three times in this way in CGN. The resulting list of 35 N+P combinations contained 8 combinations which were also selected as collocational N+P combinations in section 2. For the other combinations (e.g. *vertrouwen in (confidence in), problemen met (problems with), inzicht in (insight in), waardering voor (appreciation for), etc.*) we checked in the Twente News Corpus whether they allow PP-fronting. This was true for almost all N+P combinations. For a few combinations (*verschil met/in (difference with/in), kans op (chance on), deelname aan (participation in), parodie op (mockery of), interpretatie van (interpretation of)*) no examples of PP-fronting could be found. This suggests that the majority of the N+P combinations annotated as involving a prepositional complement in CGN are comparable with the expressions we investigated, and thus are problably best analyzed as V+PP in cases where the PP has been fronted. For the remaining cases, the CGN-annotation (i.e. N+PP) is most plausible.

Our corpus based analysis suggests that PP-fronting is best analyzed as involving a PP which is a dependent of the verb. The strong semantic relation between the noun and the preposition suggests that the examples we have investigated are examples of *phrasal verbs*, involving a verb with a more or less fixed NP-complement and a PP-complement. This analysis is compatible with observations in the ANS (Haesereyn *et al.*, 1997), Broekhuis (2003) and Loonen (2003). These works all provide (non-exhaustive) lists of phrasal verbs involving a PP-complement. A considerable number of V+N+P combinations which we found in the P1 corpus are mentioned by these authors as well. Also, a large number of the N+P combinations annotated as involving a noun with a prepositional complement in CGN, is listed with one or more verbs as a phrasal verb in one of these sources.

8 Conclusion

In this paper, we have reported on research in which a number of claims with respect to PP-fronting in Dutch were checked against corpus data. The results can be summarized as follows:

1. Certain verbs are far more frequent in P1 sentences than in general sentences containing the relevant N+P combination.

2. There is a difference in the distribution of determiners in P1 and in the general corpus. The difference seems to be mainly a consequence of the difference in distribution of verbs in both corpora.

3. PPs that occur sentence initially may also appear within relative clauses.

4. A number of patterns which have been used as argument for a specific analysis of PP-fronting are hardly encountered in a large corpus.

Conclusions 1-3 suggests that the V+PP analysis is most plausible. This confirms the conclusion in Coppen (1991), who argues for an analysis which treats the PP as an argument selected by the combination of NP+V, i.e. a (pseudo) phrasal verb.

References

Bach, E. and Horn, G.(1976), Remarks on conditions on transformations, *Linguistic Inquiry* **7**, 265–299.

Bouma, G. and Kloosterman, G.(2002), Querying dependency treebanks in XML, *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, Gran Canaria.

Broekhuis, H.(2003), Het voorzetselvoorwerp, *Nederlandse Taalkunde*. to appear.

Coppen, P.-A.(1991), Over vooropstaande PP's is het laatste woord nog niet gesproken, *Gramma* **15**(3), 209–225.

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

Haesereyn, W., Romijn, K., Geerts, G., De Rooy, J. and Van den Toorn, M.(1997), *Algemene Nederlandse Spraakkunst*, Martinus Nijhoff Uitgevers Groningen / Wolters Plantyn Deurne. Tweede, geheel herziene druk.

Klein, M. and van den Toorn, M.(1979), Van NP-beperking tot XP-beperking; een antwoord op Kooij en Wiers 1978, *De Nieuwe Taalgids* **72**, 97–102.

Kooij, J. and Wiers, E.(1979), Beperkingen en overschrijdingen: een antwoord aan Klein en Van den Toorn, *De Nieuwe Taalgids* **72**, 488–493.

Loonen, L.(2003), *Stante pede gaande van dichtbij langs AF bestemming @*, PhD thesis, Universiteit Utrecht, Utrecht.

Malouf, R. and van Noord, G.(2004), Wide coverage parsing with stochastic attribute value grammars, *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, Hainan.

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.

van der Beek, L., Bouma, G. and van Noord, G.(2002), Een brede computationele grammatica voor het Nederlands, *Nederlandse Taalkunde* **7**(4), 353–374.