Chapter 4

Determinerless PPs

In this chapter we will use a large, automatically parsed corpus to extract the lexical information needed to facilitate an account of determinerless PPs in a (computational) grammar of Dutch. Determinerless PPs (PP-Ds) are a heterogenous group of constructions which pose problems for formal and computational grammars. We will describe the different types of PP-Ds and indicate how a grammar could account for them. For these accounts, information is required about the prepositions and nouns that participate in the construction. This information is not generally available, but with the use of corpus data a base-repository of PP-Ds is generated semi-automatically.

4.1 Introduction

Combinations of prepositions and singular nouns show many of the problematic characteristics of multiword expressions (MWEs): the syntax and semantics of the construction is often—but not always—idiosyncratic, and at the same time the constructions are to some degree productive or allow modification (Baldwin et al., 2003). With these characteristics, PP-Ds pose problems for formal and computational grammars: the grammar should allow *op reis* 'on journey' but not **op stoel* 'on chair' or **ik maak reis* 'I make journey'. It should analyze *in zwang* 'in fashion', but not allow the string *zwang* in any other context. It should not parse or generate the unmodified **op wijze* 'on way', even if the modified *op slinkse wijze* 'on sneaky way' is fine and other PP-Ds allow both modified and unmodified versions (*op (lange) termijn*) 'later/long-term', lit. 'on (long) term'.

In this chapter we present a general characterization of PP-Ds and we describe ways in which different types of the construction could be handled by a grammar. We will see that all of these accounts share the prerequisite

that information about which prepositions and which nouns participate in which type of PP-D, as well as the modifiability of the P-N combination, should be available. Unfortunately, this is generally not the case. The second half of this chapter will show how this lack of information can be overcome by using a large, automatically parsed corpus to compose the lexical resource semi-automatically. Section 4.2 is partly based on Baldwin et al. (2003) and Baldwin et al. (to appear).

4.2 The Syntax of Determinerless PPs

In earlier work (Baldwin et al., 2003) it was shown that PP-Ds do not form a homogeneous group. They argued that in principle, each combination of a preposition and a singular noun without a determiner is a PP-D, but these P-N combinations differ with respect to their syntactic and semantic markedness. In addition, it may be either the noun or the preposition which selects for the lack of a determiner. We will argue that at least one more distinction has to be made, namely whether or not the PP as a whole is dependent on a verbal or nominal head.

A PP-D minimally consists of a preposition and a noun. If this noun is a plural or an uncountable noun (*suiker* 'sugar'), which in itself may constitute a saturated noun phrase, the resulting structure is syntactically unmarked (*met suiker* 'with sugar'). But if the noun is a countable noun, which in itself does not constitute a saturated noun phrase (**huis is mooi* 'house is beautiful'), the resulting structure (*naar huis* 'to house') is syntactically marked. In this chapter the focus will be on syntactically marked determinerless PPs and from now on we will use the term PP-D to refer to this subset of all P-N combinations. Although there are criteria for distinguishing countable from uncountable nouns, e.g. only countable nouns co-occur with numerals, and only uncountable singular nouns combine with *veel* 'much', distinguishing countable from uncountable nouns is not a simple task. More about countability information and distinguishing marked from unmarked determinerless PPs in section 4.3.2 and more about countability classification in general in chapter 5.

The absence of a determiner may come with idiosyncratic semantics, such as in *buiten spel* 'not in a position to influence the matter', lit. 'offside'. Although we will mention semantic effects in PP-Ds on occasion, the focus of this chapter is on the syntactic properties of PP-Ds. For a more thorough discussion of the semantics of PP-Ds, we refer to Stvan (1998) and Baldwin et al. (to appear).

The syntactically marked PP-Ds may be further subdivided in four classes:

the fully fixed PP-Ds, PPs with bare noun NPs, compositional PP-Ds and prepositions selecting for determinerless NPs. The following four sections each discuss the syntactic properties of one class of PP-Ds. Furthermore, each type of PP-D may be either independent or dependent on a verbal (or a nominal) head.

4.2.1 Fixed determinerless PPs

The first type of PP-D is the class of fully fixed PP-Ds such as *in zwang* 'in fashion'. Modification of this type is either excluded or fully fixed, as in *naar *(eigen) zeggen* 'according to him/herself', lit. 'after own saying', and the construction is non-compositional and non-productive. This class contains PP-Ds which contain lexical items that do not occur outside of PP-Ds (anymore). For example, the word *a* from the PP-D *a priori* is not used as a preposition in regular, compositional PPs. Similarly, *lieverlee*, from the construction *van lieverlee* 'gradually' is not used as a noun elsewhere. The strings *a* and *lieverlee* can be classified as a preposition and a noun on the basis of information from other languages or historical variants of Dutch, but do not behave as such in present-day Dutch.

Lexical listing is a simple and sufficient solution for this type of PP-D. Instead of breaking the string down into a preposition and a noun and including both separately in the lexicon—which would incorrectly predict both items to occur without the other—we analyze the string as a word with spaces. The syntactic category may be adjective or adverb (1), depending on the syntactic contexts in which the PP-D occurs, and additional annotations may be added where appropriate. For example, we added the annotation (↑ATYPE) = pred in (1) to indicate that the PP-D is only used in predicative constructions. The string as a whole is associated with a unique predicate, which does not bear any formal relation to the meaning of the subparts.

(1)

in zwang:	A	(↑PRED)	= 'in_zwang'
		(↑ATYPE)	= pred
van lieverlee:	Adv	(↑PRED)	= 'van_lieverlee'

In order to include these words with spaces in the lexicon, one needs a repository of PP-Ds which are fully fixed. Section 4.3 is devoted to the semi-automatic construction of such a repository.

4.2.2 Independent bare noun NPs

A second type of PP-Ds consists of a preposition and a bare noun NP. An example of such a bare noun NP is *school* in the English *in/at/after school*.

Stvan (1998) describes different types of what she calls defective NPs in English, and focuses on the semantic effects of the determinerless use in PPs. For example, the 'institutionalized location denoting nouns' such as *school* in (2-a) do not refer to the building as such, but to the related activity, in this case attending classes. This is illustrated by the fact that it is not appropriate to use the determinerless construction for the mayor visiting the school. The crucial difference between these PP-Ds and others is that the noun occurs without a determiner outside of PPs as well, and in these sentences we observe the same semantic effects as Stvan observed for the PP-Ds (2-b) (Baldwin et al., 2003). That means that strictly speaking, the construction is not syntactically marked, but comparable to a PP with a bare plural object.

(2) a. John is at school.
b. School is over.

PPs with bare noun NPs are much more common in English than they are in Dutch (or German). Furthermore, there is little evidence for a particular semantics associated with this type of PP-D in Dutch. In the category of the institutionalized location denoting nouns we find *school* 'school' and *kantoor* 'office', which can both be used determinerless to refer to an activity, but only one of them (*school*) can be used determinerless outside of PPs: *kantoor* is only used without a determiner in PP-Ds (3) and is thus not a bare noun NP. It is better analyzed as part of a compositional PP-D, a type that we will discuss in the next section. For other classes of bare noun NPs, for example crime names in Dutch, no particular semantic effect of the absence of the determiner is observed: there is no difference in meaning between *doodslag* 'homicide' in example (4-a) and in example (4-b).

(3) a. Meteen na kantoor wandel ik met de hond.
directly after office walk I with the dog
I walk the dog directly after work.
b. *Ik vind kantoor niet leuk
I like office not PART
I do not like work

(4) a. Hij is veroordeeld wegens doodslag.
he is convicted for homicide
He has been convicted for homicide.
b. De rechter acht doodslag niet bewezen.
the judge holds homicide not proved
The judge holds homicide not proved.

To account for PPs with bare noun NPs, no special machinery is needed. The defective noun phrases may be special, but the combination of this noun phrase and a preposition is the same as for other uncountable words. We assume a grammar for Dutch which is equivalent to the toy grammar in (6) with respect to PP-Ds. We will use the sample lexical entries in (5), which capture those aspects of Dutch nouns and determiners that are relevant in the PP-D analysis, while abstracting away from the characteristics that are not of direct interest here.

We assume an NP analysis for all nominals. The determiner contributes the attributes DETFORM and DETTYPE to the NP. The value of the first is the surface string of the determiner. The latter has one of the values 'definite', 'indefinite', 'demonstrative' or 'null'. The value 'null' is explained in detail below, the other values are for indefinite, definite and demonstrative determiners respectively. By setting this value, the existential constraint on the noun in the NP rule is satisfied. This constraint states that the f-structure projected from the noun (i.e. the NP) should be defined for DETTYPE.

Parentheses around a functional equation indicate that it is optional.[1] For example, the determiner is optional in the NP rule. This facilitates NPs without a determiner, for example NPs with uncountable or plural nouns. But these NPs still need to be defined for DETTYPE to satisfy the existential constraint on the noun. Countable nouns cannot satisfy this constraint, because they do not have this feature. As a result, they cannot constitute an NP by themselves. But uncountable (mass) nouns and plurals are optionally defined DETTYPE=indef, either in the lexicon or via a lexical/morphological rule, and thus satisfy the DETTYPE constraint on NPs, even if no determiner is present.

(5)

auto:	N	(↑PRED)	= 'auto'
suiker:	N	(↑PRED)	= 'sugar'
		((↑DETTYPE)	= indef)
een:	D	(↑DETFORM)	= een
		(↑DETTYPE)	= indef
de:	D	(↑DETFORM)	= de
		(↑DETTYPE)	= def

[1] The optionality can be resolved by writing two separate entries, one with the optional equation and one without.

(6)
$$\begin{array}{lcccc} \text{NP} & \Rightarrow & \text{(D)} & & \text{N} \\ & & \uparrow{=}\downarrow & & \uparrow{=}\downarrow \\ & & & & \uparrow\textsc{dettype} \end{array}$$

$$\begin{array}{lccc} \text{PP} & \Rightarrow & \text{P} & \text{NP} \\ & & \uparrow{=}\downarrow & \uparrow{=}\downarrow \end{array}$$

Words that can form independent bare noun NPs, such as *school* in Dutch and English and crime names like *moord* 'murder' in Dutch are assigned a lexical entry similar to uncountable nouns (7), allowing for occurrence with and without a determiner in and outside of PPs. The optionality of the DETTYPE annotation ensures that definite NPs such as *de school* 'the school' are still allowed.

(7)
$$\begin{array}{lll} \textit{school}: & (\uparrow\textsc{pred}) & = \text{'school'} \\ & ((\uparrow\textsc{dettype}) & = \text{indef}) \end{array}$$

4.2.3 Compositional determinerless PPs

The third type of PP-D is syntactically marked: it consists of a preposition (e.g. *in*) and a true count noun, which only occurs without a determiner inside a PP, such as *termijn* 'term'. The meaning of the sentence is composed from the regular meaning of the preposition and the regular meaning of the noun, but in some cases semantic effects occur. For example, the meaning of *naar huis* means 'to one's own house', 'home'. The PP must be headed by a particular preposition (8-a) or a member of a particular set of prepositions (8-b). The productivity of this construction varies from nouns occurring without a determiner with only one preposition to nouns that occur with a wide range of prepositions, but productivity is never unrestricted.

(8) a. op/*in/*na termijn
on/in/after term
b. in/op/naar/*onder/*naast bed
in/op/to/under/next to bed

Modification may be excluded (9-a), optional (9-b) or obligatory (9-c). Orthogonal to this three-way distinction, the modification may be restricted (9-d) or virtually unrestricted (9-e), resulting in 5 different modification patterns.

(9) a. in *zacht bed
in soft bed

b. op (hoog) niveau
   on high level
c. op *(slinkse) wijze
   on sneaky way
d. op ski-/*lange/*mooie vakantie
   on ski/long/beautiful vacation
e. op vegetarische/politieke/water-/... basis
   on vegetarian/poltical/water/... basis

The possibility for these nouns to occur in PP-Ds, and the restrictions that apply to the determinerless occurrences of these nouns, can be represented in their lexical entries. First, we allow for an optional DETTYPE annotation, which will satisfy the f-structure constraint in the NP rule. The value of DETTYPE is 'null'. In this, the structure differs from the syntactically unmarked PPs consisting of a preposition and a bare plural, which are annotated DETTYPE=indef. The null value facilitates an implementation of semantic effects such as a familiarity effect (Stvan, 1998), which contradict a value 'indef': *naar huis* 'to house' does not mean to some indefinite house, but to one specific house, namely one's own (i.e. home). These effects disappear if a determiner is added.

Secondly, we conjoin this optional DETTYPE annotation with restrictions on the syntactic category of the mother (PP), the head of the phrase, and the presence of adjuncts. If the optional annotation is instantiated, then all conjoined annotations are instantiated as well.

If modification is not allowed, e.g. for *bed* (9-a), the noun cannot have a feature ADJ. This is indicated by the negated existential constraint in (10). The lexical entry for *bed* further requires that the PP it is contained in is headed by one of the prepositions *in* 'in', *op* 'op' or *naar* 'to'. This is achieved by the inside-out function application ((OBJ↑) PTYPE) = {in|op|naar}, that states that the f-structure of which 'bed' is the OBJ should have a PTYPE 'in', 'op' or 'naar'.

Termijn 'term' (9-b) can only form a PP-D with the preposition *op* 'on'. The lexical entry is optionally specified DETTYPE=null, and if this is annotation is instantiated, the PTYPE of the PP has to be 'op'. The noun can optionally be modified by any modifier, so no further constraints are necessary. Note, though, that the heavier the modification, the less acceptable the PP-D becomes, and the higher the probability of finding a determiner. We do not account for this heaviness effect here.[2]

[2]This heaviness effect may be accounted for with a general Optimality Theoretic constraint. Alternatively, one could specify the possible type(s) of modification on each lexical entry separately. For instance, one could restrict modification to attributive adjectives only

Wijze 'way' is obligatorily modified, but the modifier is unrestricted (9-c). This is modeled by the existential constraint (↑ADJ). Again, the heaviness constraint applies here, effectively ruling out postnominal modification.

(10)

bed:	N	(↑PRED)	= 'bed'	
		{((OBJ↑) PTYPE)	$=_c$ {in|op|naar}	&
		(↑DETTYPE)	= null	&
		¬(↑ADJ)}		
termijn:	N	(↑PRED)	= 'term'	
		{((OBJ↑) PTYPE)	$=_c$ op	&
		(↑DETTYPE)	= null}	
wijze:	N	(↑PRED)	= 'way'	
		{((OBJ↑) PTYPE)	$=_c$ {in|op|naar}	&
		(↑DETTYPE)	= null	&
		(↑ADJ)}		

More complicated are the compositional PP-Ds with restricted modification. The noun *evenwicht* 'balance' can form a PP-D with the preposition *in* 'in'. There is only one adjective that can (optionally) modify the noun: *wankel* 'unstable'. Similarly, only the adjective *onmiddellijke* 'immediate' can modify the noun *ingang* 'start', but now the modification is obligatory (*met *(onmiddellijke) ingang* 'immediately', lit. 'with immediate start').[3] In the lexical entry of the noun, we constrain the predicate of its adjunct as in (11). In case the modification is optional, we only state that the modifier cannot have a PRED other than the fixed modifier *wankel* 'unstable'. For obligatory modification, we again restrict the predicate value of the adjunct and additionally state that the NP must have a feature ADJ. The only way to satisfy both constraints is to realize the fixed modifier, in this case *onmiddellijke* 'immediate'.[4]

with a constraint (ATYPE = 'attributive') on the adjunct.

[3] *Ingang* 'start' can also be followed by a PP headed by *van* 'from' (*met ingang van* 'starting', let. 'with start from'). This is assumed to be a collocational preposition, but could alternatively be analyzed as a PP modifier in the PP-D, in which case the obligatory modifier is either the adjective with PRED 'immediate' or a PP with PTYPE 'from'.

[4] This formalization still allows for multiple occurrences of the lexically selected modifiers. It is difficult to get definite grammaticality judgments for those sentences, but it may well be that they have to be considered out. We believe that this is not a characteristic of PP-Ds, but rather a more general phenomenon that penalizes literal repetitions of words. Tracy H. King (p.c.) furthermore noted that it is technically possible to avoid having more than one adjunct by blocking the presence of the attribute SCOPE, which determines the relative scoping of the adjuncts.

(11)

evenwicht: N	(↑PRED)	= 'balance'	
	{((OBJ↑) PTYPE)	$=_c$ in	&
	(↑DETTYPE)	= null	&
	¬((↑ADJ ∋ PRED)	≠ 'unstable'}	
ingang: N	(↑PRED)	= 'start'	
	{((OBJ↑) PTYPE)	$=_c$ met	&
	(↑DETTYPE)	= null	&
	(↑ADJ)		&
	¬((↑ADJ ∋ PRED)	≠ 'immediate')}	

In our account of compositional PP-Ds, non-heads (nouns, NPs) pose restrictions on their head (the preposition). A similar approach was advanced by Soehn and Sailer (2003) in HPSG. This work focuses on so called *unique nominal complements*, nouns which never occur outside of PPs. This is in contrast to our case of compositional PP-Ds, where the nouns do occur independently as well as inside PPs. Ideally, one would generate and analyze the use of the noun in and outside of PPs as instances of one and the same lexical entry, avoiding duplication of identical information (e.g. PRED information) in the lexicon. The LFG framework facilitates such an analysis via the mechanism of optional (complexes of) f-structure annotations: if the noun is used outside of a PP, the optional annotations are not realized. As a consequence, it is not defined for DETTYPE and can only form an NP after combining with a determiner. If the noun occurs inside a PP-D, the optional annotations have to be instantiated to satisfy the DETTYPE constraint, and consequently the structure has to satisfy all other conjuncts of the complex of optional annotations, restricting the head of the PP and modification. We see again that in order to implement an account of this type of PP-Ds, we need detailed information about which nouns combine with which prepositions in a PP-D and the modifiability of the resulting construction.

4.2.4 Prepositions selecting for determinerless NPs

In the fourth and last type of PP-D, it is the preposition that licenses determinerless NP complements. This type of PP-D again has two subtypes. On the one hand we have the prepositions *per* 'per', *ter* and *ten* 'at' (with different archaic case-markings), which obligatorily select for a PP-D (12).[5]

[5]There is one use of *per* with a determiner:

(i) Ik stop per de eerste van de volgende maand.
I quit from the first of the next month
I will quit on the first of next month.

In the case of *per*, this is a fully productive process. *Ter* and *ten* are historically contractions of the preposition *te* 'at' and an article, and their use is restricted. *Te* is still used productively, but only with city names, which are always determinerless. Some fixed expressions with *te* do contain determiners, e.g. *te allen tijde* 'at all times' (but *te voet* 'on foot' and *te midden van* 'amidst'). This is in contrast with *ten* and *ter*, which occur in many fixed combinations, but never with a determiner.

On the other hand we have prepositions that license PP-Ds, but can occur with regular NPs as well, such as *zonder* 'without' (13). Sometimes, the determinerless occurrences are restricted to a certain semantic domain. An illustration of this phenomenon is the preposition *in* 'in', which combines with any bare noun indicating a piece of clothing, but requires a 'regular' NP elsewhere (14).

(12) a. Ik zal het per koerier laten bezorgen
I will by courier let deliver
I will have a courier deliver it.
b. *Ik zal het per een koerier laten bezorgen
I will by a courier let deliver

(13) a. Ik kan best zonder auto.
I can fine without car
I can live just fine without a car.
b. Ik kan best zonder een auto.
I can fine without a car
I can live just fine without a car.

(14) a. Zie je die student in
see you that student in
pak/uniform/spijkerbroek/bloemetjesjurk?
costume/uniform/jeans$_{sg}$/flower dress?
Do you see that student in costume/uniform/jeans/flower dress?
b. *Ken je die student in kerk/gebouw/klas/groep/kamer?
Know you that student in church/building/class/group/room
c. *Ik vind pak/uniform/spijkerbroek/bloemetjesjurk niet
I consider costume/uniform/jeans$_{sg}$/flower dress not
mooi.
nice

Prepositions selecting for 'real' PP-Ds should be distinguished from preposi-

We do not account for this use of *per* in this chapter.

tions that often occur with uncountable nouns. An example of such a preposition is *wegens* 'on account of' or the collocational preposition *op verdenking van* 'on suspicion of', which occur very frequently with the names of crimes, which we saw can form NPs by themselves 4.2.2. If one combines these prepositions with true count nouns, the determiner is again obligatory (15).

(15) a. A. werd veroordeeld wegens een autokraak
A. was sentenced one account of a car break-in
A. was sentenced on account of breaking into a car
b. *A. werd veroordeeld wegens autokraak
A. was sentenced on account of car break-in

Treating *per* and *zonder* 'without' as regular prepositions, the PP-Ds in (12-a) and (13-a) violate the grammar rules in (6): the bare count nouns are not specified for DETTYPE and thus cannot form an NP. But prepositions do not combine with bare nouns or N′s. We solve this by specifying the value for the nouns DETTYPE attribute *on the preposition*. By doing so, the count nouns can form an NP, but only if this NP functions as the complement of this particular preposition.[6]

The preposition *per* specifies its complement to be DETTYPE=null (16-a). Thus NPs with a determiner, as well as bare plurals, mass nouns and proper names, are correctly excluded. For *zonder* 'without', this would not be correct: example (13-a) and (13-b) are both grammatical and can be used interchangeably. We model this with an optional annotation ((↑OBJ DETTYPE)=indef). All this annotation does is providing the necessary DETTYPE attribute if it is not provided by a determiner. The value is the same as for NP complements with an indefinite determiner, correctly predicting identical semantics for example (13-a) and (13-b). The optionality of the annotation ensures that definite prepositional complements can still be derived.

(16) a. *per*: P (↑PRED) = 'per(OBJ1)'
(↑OBJ DETTYPE) = null
b. *zonder*: P (↑PRED) = 'zonder(OBJ1)'
((↑OBJ DETTYPE) = indef)

A different approach is necessary for semantically restricted PP-Ds. One could try to add the semantic restriction on the lexical entry of the preposition, but this would imply that *all* members of the semantic class allow for the determinerless construction. This appears not to be the case. Compare (14-a) with (17-a) and (17-b) with (17-c).

[6]This solution is supported by the fact that the prepositions *ter* and *ten* are historically a combination of a preposition and an article.

(17) a. ??Zie je die student in broek?
see you that student in trousers$_{sg}$
b. Hij is op reis/tournee/pad/weg/vakantie/expeditie.
he is on voyage/tour/path/way/vacation/expedition
c. *Hij is op trip/tocht.
he is on trip/journey

Alternatively, one can treat each example as an instance of a compositional PP-D. Although this appears to be empirically correct, it does not capture the semantic generalization. We will get back to these semantically restricted PP-Ds in section 4.3.5.

4.2.5 Determinerless PPs as dependents

Orthogonal to the distinctions made in Baldwin et al. (2003, to appear), a number of (Dutch) PP-Ds function only as dependents of a verbal (or nominal) head. In this case, the preposition does not combine with detless nominals unless it is an arguments of this particular verbal or nominal head. The PP-D itself may be fully fixed (18-a) or compositional (18-b). In some cases, the head of the prepositional complement selects for any nominal object, as long as it is determinerless (18-c)-(18-d). We will refer to dependent PP-Ds as determinerless prepositional complements. Dependent PP-Ds should be distinguished from independent PP-Ds, because their distribution is much more restricted. The lexical entry which licenses the occurrence of such PP-D is not the entry of the preposition or the noun in this PP-D, but the entry of the verb (or the noun) which selects for the PP-D complement: they will be analyzed as fixed prepositional arguments Villada Moirón (2004). Table 4.1 presents an overview of the different types of PP-Ds that we have distinguished.

(18) a. Iemand in toom houden
someone in bridle hold
To restrain someone
b. In première gaan
in premier go
To have its opening night
c. Van auto veranderen
of car change
Change cars
d. De functie van voorzitter
the function of president

The function of president

4.3 Extraction of PP-Ds

4.3.1 Introduction

In the previous section we identified various types of PP-Ds: fully fixed and compositional PP-Ds, bare noun NPs and bare noun selecting prepositions, all of which could be independent PPs or dependents of verbal/nominal heads. We indicated how each of these types could be accounted for in a grammar. Although the analyses differ in many respects, they all share one prerequisite: that information is available about which nouns and which prepositions participate in which type of PP-D and which kind of modification is allowed.

This information is not available, generally. The Dutch part of EuroWordNet Vossen and Bloksma (1998) lists a total of seven PP-Ds. The electronic dictionaries Celex (Baayen et al., 1993) and Parole[7] do not include any information on PP-Ds. Even the Dutch reference grammar (Haeseryn et al., 1997) and the main dictionary (Geerts and Heestermans, 1992) do not include information about PP-Ds systematically, although the grammar includes a list of collocational prepositions (i.e. fixed combinations of a preposition, a—possibly determinerless—NP and another preposition). Only the Alpino lexicon, which is part of the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b) contains more information about marked PPs. The lexicon lists about 95 fixed PP-Ds, such as *a priori* 'a priori', *in feite* 'in fact' and *van nature* 'by nature' and 132 (semi-automatically extracted) collocational prepositions consisting of a preposition, a bare noun and another preposition, such as *in antwoord op* 'in reply to', (Bouma and Villada, 2002). Furthermore, almost fifty fixed parts of larger idiomatic expressions, for example phrasal verbs, contain PP-Ds.

The work presented in this section aims at overcoming the lack of systematic lexical information by means of (semi)automatic extraction of PP-Ds from corpus data. We want to identify prepositions that select for determinerless NPs, nouns that participate in compositional PP-Ds, and fixed P-N tuples. Furthermore, subcategorized PP-Ds should be distinguished from independent ones. Like regular uncountable nouns, the independent bare noun NPs from section 4.2.2 will be discarded.

[7] `http://www.inl.nl/corp/parole.htm`

Type	Example	Characteristics
Selection by P (obligatorily)	*per te vroeg geboren kind* (per premature child)	modification allowed high prep-noun entropy no PP+D
Selection by P (optionally)	*zonder hoed* (without hat)	modification allowed high prep-noun entropy also PP+D
Fixed PP-D	*in principe* (in principle)	zero modification entropy high verb entropy no PP+D
Idiosyncratic P-N pairs	*op straat* (on street)	no NP-D restricted modifiability high verb entropy
Bare noun NPs	*wegens moord* (for murder)	also NP-D high preposition entropy high verb entropy
PP-D phrasal verbs	*in premiere gaan* (in premier go, have opening night)	low dep-head entropy
PP-D verbal complements	*van auto veranderen* (of car change)	low dep-head entropy

Table 4.1: Overview of PP-D types.

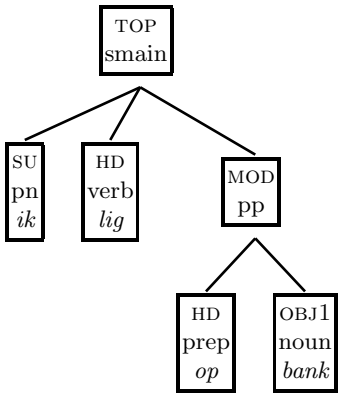
4.3.2 Preliminaries

The PP-D extraction methods proposed in this work are all based on parsed corpus data. We used a 75M word newspaper corpus that was automatically annotated with dependency structures by the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b). Examples of dependency trees are in example (19)-(21). The parser has an overall accuracy of 85.5%, measured over the dependency relations. The syntactic annotation allows us to extract information about (different types of) modification and about the verb that governs the PP-D. Both types of information are difficult to extract with shallower forms of annotation.

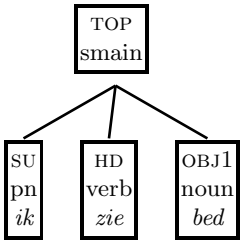
Baldwin et al. (2003) argued that chunked data should be prefered over parsed data for extraction of PP-Ds. The motivation for this is that the quality of the output of the parser is conditioned on the ability of that parser to analyze PP-Ds, giving rise to circularity. However, in our case we used the Alpino parser, for which this circularity does not arise. The Alpino parser overgenerates in that it generally allows bare nouns to form an NP. Although this is not always correct, it allows every P+N combination to be analyzed as a PP-D, even if the noun in itself does not constitute a saturated noun phrase. The only PP-Ds that will not be retrieved systematically are those that feature in collocational prepositions or function as a fixed part of a larger idiomatic expression, because these are analyzed as a single lexical item. This analysis as a word with with spaces is based on the annotation guidelines for collocational prepositions of the Corpus of Spoken Dutch (Moortgat et al., 2001). As a result, these PP-Ds will not be retrieved by querying for P+N patterns. Instead, the collocational prepositions will show up in the results as prepositions.

As a result of the overgeneration strategy of the Alpino parser, implementing a detailed account of PP-Ds will not improve coverage: Alpino already assigns the grammatical *op reis* 'on journey' a PP analysis. However, it assigns the same parse to ungrammatical PP-Ds (19) and to the ungrammatical use of the bare noun outside of the PP-D (20). If one aims at a parser which parses all and only grammatical strings or a generator which generates all and only grammatical sentences, one needs to replace the overgenerating NP $\Rightarrow$ N rule by a detailed account of PP-Ds (and an extensive list of uncountable nouns, a named entity recognition module, etc.).

(19) *Ik lig op bank.
I lie on couch

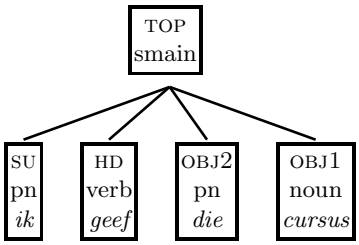


(20) *Ik zie bed.
I see bed



Such a restriction to the grammar is expected to reduce ambiguity and improve accuracy of the parser, as incorrect application of the NP ⇒ N rule may result in incorrect full parses of grammatical sentences. For illustration, if regular count nouns are prohibited to form a NP by themselves, then the noun *cursus* 'course' cannot form an NP and the incorrect parse in (21) is ruled out.

(21) Ik geef die cursus
I give that course
I teach that course



From the parsed data we extracted the preposition heading the PP-D, the noun object of the preposition and the verb that heads the PP-D.[8] The nouns were restricted to singular common nouns. Determiners and appositions were not allowed in the NP. Furthermore, we extracted the list of modifiers (both post- and prenominal) and the list of dependency triples corresponding to the heads of the modifiers (allowing for generalizations over types of modifiers).

As the goal of this research is the extraction of syntactically marked PP-Ds, we are only interested in PP-Ds with a countable noun, which in itself cannot constitute an NP. Unfortunately, the availability of reliable count-

[8]Or the verb that heads the NP in which the PP-D is embedded. We included these verbs in order to recognize fixed verbal arguments that are misanalyzed as NP internal modifiers.

ability information is limited. Nouns in the Alpino lexicon (14K words) are labeled with countability information. We found a 81.1% agreement between the countability judgments in Alpino and a 196 word hand annotated gold standard. The judgments in the gold standard were based on actual occurrences of the nouns in the Twente Nieuws Corpus.[9] In addition to this, we used a list of 6K nouns that were automatically classified as countable, uncountable or both according to the method described in Baldwin and van der Beek (2003) and chapter 5 of this thesis. The accuracy of this list is 85.7%. Note furthermore that nouns may have both countable and uncountable usages. In this research, any noun which has at least one uncountable sense is considered syntactically unmarked whenever it occurs without a determiner (in or outside of a PP). For example, *buiten beeld* 'offscreen' is not considered syntactically marked, because *beeld* has an uncountable use e.g. *goed beeld hebben* 'to have good reception'. The reason for this approach is that word sense disambiguation is not yet feasible in broad scale computational grammars, thus countable and uncountable senses cannot be distinguished from eachother.

Using the extracted data, we first identify prepositions that (optionally) select for determinerless objects, then fully fixed PP-Ds and nouns that select for occurrence in compositional PP-Ds. Finally, we illustrate how the same methods can be applied to extract a set of PP-Ds that are selected for by particular verbs, forming phrasal verbs or determinerless prepositional complements.

4.3.3 Prepositions selecting for determinerless NPs

It was shown in section 4.2.4 that the prepositions that select for determinerless objects can be subdivided in prepositions that only occur in PP-Ds and those that optionally select for a determinerless NP. The prepositions *ter* and *ten* 'at' (with archaic casemarkings) and *per* 'per' obligatorily select for PP-Ds and can simply be listed in the lexicon with the lexical entry given in example (16-a). We excluded them from our data in further experiments.

Other prepositions optionally select for a determinerless complement. To extract these prepositions, we calculated the ratio between the number of noun types in PPs with a determiner headed by preposition P and the number of types in PP-Ds headed by the same preposition. We excluded nouns that were flagged as uncountable and used a frequency cutoff of 50. This relatively high cutoff was used to filter out many of the infrequent collocational prepositions, while keeping the very frequent simplex prepositions. As

[9] `http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html`

Preposition	Example Noun	Ratio	Countability
vol 'full of'	*liefde*'love'	0.00	uncount
bij wijze van 'by means of'	*geintje* 'joke'	0.01	count
qua 'wrt'	*lichaam* 'body'	0.02	count
op verdenking van 'on suspicion of'	*moord*'murder'	0.29	uncount
richting 'towards'	*schatkist*, 'treasury'	0.37	count
op het gebied van 'concerning'	*kunst* 'art'	0.81	uncount
zonder 'without'	*paspoort* 'passport'	0.86	count
als 'as'	*banneling* 'exile'	0.90	count
bij gebrek aan 'lacking'	*ervaring* 'experience'	1.04	uncount
tot 'to'	*bedelaar* 'beggar'	1.93	count

Table 4.2: Prepositions that select for PP-Ds

the list of uncountable nouns is far from complete, many uncountable nouns still show up in the data. We therefore also extracted a sample of 15 noun objects of these prepositions and manually classified them as countable or uncountable, making it possible to distinguish prepositions that select for 'real' PP-Ds from those that simply select for uncountable objects. Table 4.2 lists the ten prepositions with the lowest type-ratio, including both prepositions that frequently co-occur with uncountable nouns and prepositions that allow uncountable nouns to form a determinerless prepositional object.

Finally, there are prepositions which generally select for regular NP objects, but form PP-Ds with bare singular nouns from a particular semantic class to be their object. An example is the preposition *in* with clothing items such as *pak, bloemetjesjurk, uniform, overhemd* 'suit, flower dress, uniform, shirt'. We argued that semantically restricted selection by the preposition would overgenerate in section 4.2.4. We therefore treat this type of PP-D as a compositional PP-D in section 4.3.5, where we hope to capture the semantic generalization by clustering individual PP-Ds on the basis of EuroWordNet synsets.

4.3.4 Fixed determinerless PPs

The second group of PP-Ds to be extracted is the fully fixed PP-Ds. A very simple heuristic was applied: we extracted all preposition-noun tuples (F>10) with 'nouns' that only occur grammatically in PP-Ds. At least 99% of all occurrences of the noun had to be determinerless and at least 90%

Candidate	F	Modifier	P	Fixed
in afwachting (in anticipation)	610	van (of)	0.91	√
volgens/naar zeggen (according to)	564	eigen (own)	1.00	√
in opspraak (compromised)	366			√
op voorhand (in advance)	347			√
bij uitstek (pre-eminently)	345	in (in)	0.02	√
op jaarbasis (on a yearly basis)	264	op (on)	0.02	√
van nature (by nature)	200			√
in diskrediet (into disrepute)	172	bij (by)	0.01	
op straffe (under penalty)	147	van (of)	0.94	√
naar hartelust (to ones heart's content)	115	met (with)	0.03	√
in zwang (in fashion)	83	in (in)	0.04	√
sinds mensenheugenis (within living memory)	77	in (in)	0.03	√

Table 4.3: Fixed PP-Ds

of the occurrences inside a PP[10]. The difference between the two cutoffs is motivated by the fact that very little parse errors are made with respect to the recognition of the determiner and grouping Det + N in an NP. On the other hand, the parser does not always identify the PP correctly, so we slightly relaxed the condition on determinerless occurrence outside of PPs. Finally, we set the maximum noun-preposition entropy at 1.00 as a filter for uncountable nouns.

In total, 45 candidate fixed PP-Ds were extracted. The candidates were manually checked by three native speakers, including the author. For 36 (80%) of them, at least two of the informants indicated that they knew the noun and that it only occurred in a PP context. Nine candidates were considered false positives. Some of the true positives could be considered collocational prepositions (see section 4.3.1), but as they conform to the definition of fully fixed PP-Ds that we formulated, we included them in the results.

The modifiers of the candidate PP-Ds were extracted and the entropy of the modifier given that there is one) was calculated. Table 4.3 lists the most frequent fixed PP-Ds with their frequency, the most frequent modifier and the probability of this modifier. We included all types of modification, but abstracted away from the objects in PP modifiers: all PPs headed by a particular preposition are regarded the same. The distribution of the modifiers

[10]A cutoff of 3 was used for PP-Ds with a frequency lower than 30, allowing for a maximum of 3 typos or parse errors

confirms that modification is fully fixed for this type of PP-D: the probabilities of the most frequent modifier are either very high, between .90 and 1.00, or else less than .05.

4.3.5 Compositional determinerless PPs

The largest set of PP-Ds is the group of compositional PP-Ds. Clearly, the simple heuristic applied for the fixed PP-Ds will not work for PP-Ds with 'regular' nouns: the nouns occur outside of PPs and in regular PPs as well as in PP-Ds. Instead, we used the data from our automatically parsed corpus to calculate the association between the absence of a determiner and the occurrence in a PP. The count noun may occur without a determiner outside of PPs, for example as a result of the universal grinder or a parse error, but it will appear determinerless in PPs much more frequent than expected based on the ratio of NP/PP contexts. The association is measured with the log-likelihood ratio, implementation based on the NSP package Banerjee and Pedersen (2003).

We excluded the nouns of the fixed PP-D category, the prepositions that obligatorily select for PP-Ds and prepositions from table 4.2 that optionally select for PP-Ds. We used a frequency cutoff of 10. We further restricted our candidates by setting a verb entropy minimum of 2.00. This is aimed at excluding phrasal verbs.[11] Table 4.4 lists the nouns for which the association is the strongest.

We see that 7 (47%) of the top 15 are nouns that do not occur without a determiner outside of PPs, making their occurrence in PP-Ds syntactically marked.[12] Only two of the nouns that were judged uncountable were listed as such, as well as one noun (*mate*, 'measure') that was judged countable. With large amounts of high quality countability information, precision can increase considerably.

The members of semantic classes that select for occurrence in PP-D, such as pieces of clothing (plus the preposition *in*) in Dutch, are not represented in the output. A possible explanation is that that each of these nouns is infrequent, often not passing the frequency cutoff. This problem

[11] We included nouns which occur with a low entropy with *zijn* ('to be'), as these tend to be PP-Ds with predicative uses, not phrasal verbs.

[12] Again, nouns were marked countable if at least two out of three informants knew the word and indicated it was strictly countable. This only reflects the basic countability of the nouns. Countable words can still be used in uncountable contexts by way of coercion, the *universal grinder* being the most well known example. Despite this, we still assume that nouns have a basic countability and that this influences the possibility to occur with or without a particular determiner.

Noun	LL	P ent	P max	Count
huis (house)	2994.1	1.08	naar (to)	√
belang (interest)	2226.4	0.18	van (of)	
beeld (view)	2161.3	0.58	in (in)	
verwachting (expectation)	2101.0	0.77	naar (to)	
straat (street)	1904.2	0.31	op (on)	√
voorbeeld (example)	1877.9	0.52	bij (by)	√
druk (pressure)	1636.8	0.51	onder (under)	
plaats (place)	1604.0	1.62	van (of)	
dienst (service)	1330.6	0.20	in (in)	√
voorkeur (preference)	1251.7	0.22	bij (by)	
principe (principle)	1206.4	0.12	in (in)	√
kracht (strength)	1058.2	1.32	met (with)	
leeftijd (age)	1050.1	0.31	op (on)	√
totaal (total)	973.51	0.02	in (in)	√
school (school)	930.67	1.21	op (on)	

Table 4.4: Flexible PP-Ds

can be circumvented by clustering together all members of the semantic class, thus increasing the amount of data per item. We combined the data of all hyponyms of *reis* (journey), including *tournee, safari, vakantie, kruistocht* (tour, safari, vacation, crusade) in EuroWordNet and we did indeed find a positive correlation between occurring inside a PP and lacking a determiner. The log likelihood of this cluster is 259.0, which is higher than that of the most frequent member of this set, *vakantie* (vacation, 195.5), but lower than that of the number two, *reis* (journey, 290.7). However, for the semantic class *kledingstuk* (piece of clothing), we found a *negative* correlation between the lack of a determiner and being the object of a preposition. For other clusters, no appropriate hyperonym could be found in EuroWordNet. An example is the set *op verzoek/bevel/aanbevelen/aanraden/initiatief* (on request/order/advise/recommendation/initiative).

Not all nouns form PP-Ds with various different prepositions: many nouns combine with only one or two prepositions in a PP-D. This is illustrated by the low preposition entropy in table 4.4. As the nouns combine with other prepositions in regular, saturated NPs, their ability to occur in preposition specific PP-Ds is not (optimally) reflected in the results of table 4.4, which generalize over all prepositions. Instead, we should measure for each noun the association between the absence of a determiner and the occurrence in a PP headed by a specific preposition. Again, the association was measured

Tuple	LL	Ent	Mod max	Count
naar huis (to house)	4972.6	0.23	met (with)	√
van belang (of interest)	4299.2	1.86	groot (great)	
op straat (on street)	2927.6	0.30	in (in)	√
onder druk (under pressure)	2865.3	1.28	van (of)	
naar verwachting (to expectation)	2846.2	0.21	over (about)	
in dienst (in service)	2756.5	1.18	bij (at)	√
bij voorbeeld (for example)	2515.3	1.22	in (in)	√
in principe (in principle)	2075.2	0.33	voor (for)	√
bij voorkeur (by preference)	1783.7	1.01	in (in)	
op bezoek (at visit)	1744.3	0.74	bij (at)	
op leeftijd (at age)	1725.9	4.48	jong (young)	√
in totaal (in total)	1706.3	0.27	voor (for)	√
na afloop (after ending)	1440.5	0.33	in (in)	√
in werkelijkheid (in reality)	1285.3	0.18	in (in)	
voor rekening (on account)	1276.6	0.91	eigen (own)	√

Table 4.5: Flexible PP-Ds

with the log likelihood ratio. The results are listed in table 4.5.

Table 4.5 also list for each PP the modifier entropy and the most probable modifier. As expected, modification is somewhat more flexible than in the fully fixed PP-Ds. However, we see that modification is still very much restricted.

Comparing the tables 4.4 and 4.5 we see that the association calculation per preposition leads to a higher number of syntactically marked PP-Ds in the top 15: 9 (60%) instead of 7 (47%). Furthermore, looking at the 50 highest ranked nouns for both methods we see that the nouns that were found by the more general approach are almost exclusively uncountable nouns (*contrast, grond, lucht* 'contrast, ground, air', among others). In contrast, the nouns that were retrieved by the preposition specific method but not by the general approach are almost all count nouns (*gesprek, reis, slot* 'conversation, journey, lock'). We conclude that the preposition specific method is less sensitive to the availability of high quality countability information, although many uncountable nouns were still retrieved.

With the preposition specific approach, performance also increases with respect to EuroWordNet clustering: the combination of *op* ('on') with the journey-cluster now scores 947.1, whereas *reis* (journey) and *vakantie* (vacation), the two highest ranked members of the cluster, score 748.7 and 549.3. For the clothing class, the effect is not so strong: although the negative association is no longer present, we also do not find a strong positive association

PP-N	LL	Verb ent	Verb max
tot hand (to hand)	135.7	0.00	ga (go)
buiten functie (of duty)	142.3	0.00	stel (put)
bij stuk (at place)	431.4	0.06	houd (hold)
in bescherming (in protection)	113.5	0.09	neem (take)
in aarde (in ground)	151.7	0.11	val (fall)
in vervulling (in fulfilment)	194.0	0.14	ga (go)
van stemming (from voting)	328.9	0.16	onthoud (refrain)
in stilzwijgen (in silence)	108.9	0.18	hul (surround)
in rekening (in account)	170.4	0.20	breng (bring)
in première (in première)	963.0	0.21	ga (go)
op adem (on breath)	170.0	0.25	kom (come)
in opstand (in revolt)	622.7	0.28	kom (come)
bij kas (in treasury)	144.3	0.29	zit (sit)
met rust (in rest)	224.3	0.35	laat (let, leave)
in verlegenheid (in embarrassment)	151.2	0.36	breng (bring)

Table 4.6: PP-Ds with low verbal entropy

(LL 6.3). On top of that, a member like *uniform* scores much higher with 78.4. We can conclude that although positive clustering results may confirm existing intuitions about certain semantic classes optionally selecting for PP-Ds, the results are not good enough for identifying those semantic classes, even if a common class exists in EuroWordNet.

4.3.6 Dependent determinerless PPs

In the previous experiments we controled for selection by setting a minimum verbal entropy. If we focus on the other end of the scale, selecting for PP-Ds with a low verbal entropy, we find a list with a high density of phrasal verbs (see table 4.6).

Phrasal verbs are not the only verbal constructions containing PP-Ds. Some verbs take a prepositional complement which is a PP-D. In contrast to the phrasal verbs, the nominal in the PP is not fixed (22).

(22) Van auto/baan/jurk veranderen.
of car/job/dress change
Change cars/jobs/dresses.

To find out which verb preposition combinations select for PP-Ds, we used

V	P		-D	+D
inboeten	aan	'lose'	29	2
uitschelden	voor	'call (so) st'	26	4
wisselen	van	'change'	83	13
verwisselen	van	'change'	19	4
winnen	aan	'win'	48	12

Table 4.7: Verbs selecting for a determinerless prepositional complement

the same technique that we applied for the PP-D selecting prepositions, only changing the prepositions in verb preposition tuples: we calculated for each combination the ratio between the types with and without a determiner. The lowest ranked combinations (with proportionally the most determinerless occurrences) are listed in table 4.7. Although we excluded combinations with an uncountable noun, we still find verbs selecting for uncountable prepositional complements, such as *inboeten aan* 'lose', showing once again the influence of countability information on the results. As the precision of this list is low, we did not include them in further experiments.

4.4 Evaluation and Distribution of PP-Ds

We extracted a total of 363 PP-Ds from the automatically parsed corpus with various methods. Table 4.8 summarizes the PP-D types, the extraction methods and the results. To evaluate the classification methods proposed, we took all 3612 preposition+noun patterns from CGN, the syntactically annotated Corpus of Spoken Dutch Levelt (1998), removed all uncountable nouns, typos and obligatorily PP-D selecting prepositions, as well as PP-Ds containing phrases that Alpino analyzes as fixed, and classified the resulting 1510 PP-Ds according to PP-D type on the basis of our extracted data collection. From those 1510 PP-Ds, our methods classified 836 as a syntactically marked PP-D of a certain category. A total of 674 PP-Ds were *not* classified, resulting in a recall of 55.1%. We can compare this figure to a raw frequency baseline. If we take the 359 most frequent preposition+noun patterns in the training data instead of the 359 PP-Ds in our data lists, we get a recall of 34.3%. This indicates that our approach not only provides more detailed information than raw frequency (namely the PP-D type of a preposition+noun pattern), but also leads to a higher recall.

[13] 43% of the extracted PPs were syntactically marked PP components of phrasal verbs. In additional 12% of the cases the PP was not syntactically marked (the noun was un-

PP-D type	N	Test	Settings	Precision (%)
Preposition	10	Type Ratio	F>50	60
Fixed PP-D	45	Cut-off	F>10, -D>99%, +P>90%, Ent<1.00%	80
Idiosyncratic PP-D	200	Log-likelihood	F>10, V Ent>2.00	63
Phrasal Verbs	108	Entropy	F>10, V Ent<2.0, LL>100	43/55[13]
Total	363			

Table 4.8: Extracted PP-Ds

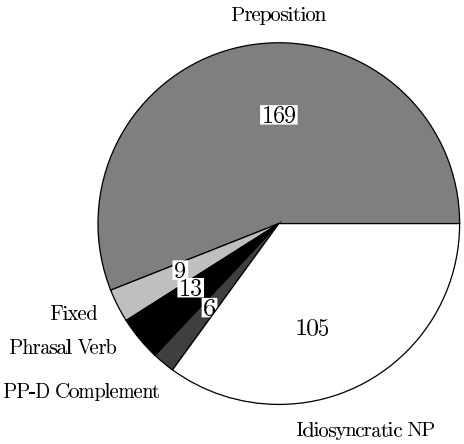


Table 4.9: Distribution of PP-D types in CGN.

The tables 4.9 and 4.10 show the results of the classification and give an impressions of the distribution of different PP-D types. We see that the prepositions that optionally select for PP-Ds and prepositions that select idiosyncratic NPs make up about 90% of the classified data. The fully productive PP-D selecting prepositions take up the largest part of the types, whereas most tokens are prepositions with idiosyncratic NPs.

Secondary evaluation on the written data from the Alpino Treebank van der Beek et al. (2002a) leads to higher recall (63.9% for semi-automatic extraction vs. 38.8% for raw frequency), but shows the same overall distribution of PP-D types.

We excluded the PP-Ds that Alpino classifies as fixed, because those PP-Ds were not represented in the training data either. This influenced the distribution in table 4.10: the 41 fixed PP-D types we excluded gave rise to 330 tokens in CGN. Among these excluded fixed PP-Ds are multiword adverbs, collocational PPs and fixed parts of larger idiomatic expressions, but by far the most tokens are phrasal verbs. The overall percentage of phrasal verbs in the data is thus higher than indicated in the pie charts.

As mentioned, 674 PP-Ds were *not* classified. These unclassified PP-Ds form a heterogeneous group. Among the negatives we find typical characteristics of spoken language (*in dinges* 'in what's-its-name') and typos, but also clear PP-Ds with idiosyncratic NPs (*in bad* 'in bath') and many instances of the preposition *met*, seventh on the list of prepositions that optionally select for NP-Ds. Interestingly, we also find members of the clothing class (*in pyjama/smoking/uniform* 'in pyjamas$_{sg}$/smoking/uniform') and the jour-

countable), but it was part of a phrasal verb.

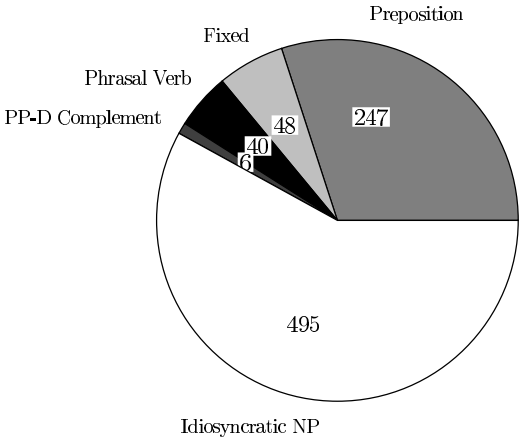


Table 4.10: Distribution of PP-D tokens in CGN.

ney class (*op tournee/trektocht/trouwreis* 'on tour/hiking tour/honeymoon'), showing that a proper treatment of semantically restricted PP-Ds has the potential of improving coverage. Another category of false negatives are compounds of words we identified as parts of PP-Ds (*in poedervorm, op beleidsniveau* 'in powder form', 'at the level of policy makers', lit. 'on policy level'). As compounding is generally possible in all PP-Ds except the fully fixed, these false negatives will be correctly analyzed by a grammar that includes a list of nouns occurring in PP-Ds.

4.5 Conclusion and Discussion

PP-Ds form a heterogeneous collection of constructions with varying degrees of syntactic and semantic markedness. A correct analysis of the different types of PP-Ds requires knowledge about which nouns and which prepositions participate in PP-Ds, and the modifiability of the resulting structure, which is generally not available. It was illustrated that a base repository of PP-Ds can be composed semi-automatically on the basis of automatically parsed corpus data. Information about the prepositions and the nouns, their modifiers and the governing verbs was extracted and used to calculate the association between the presence or absence of a determiner and the occurrence of the noun in or outside of a PP.

Baldwin et al. (2003) and Baldwin et al. (to appear) showed that PP-Ds are not just a Dutch problem, but that they occur in many languages. The methods can be applied to other languages, as long as either a large syntactically annotated corpus is available or an unannotated corpus and a

preprocessor which can extract prepositions, nouns, verbs and modifiers from sentences with PP-Ds. The resulting repository of PP-Ds may be fed into parsing systems to extend their coverage. In the case of the Alpino parser, the PP-D repository is a first step towards abolishing the NP ⇒ N rule, which furthermore requires more accurate countability data and high quality named entity recognizers.

The absence of gold standard data complicates thorough evaluation. Manual inspection showed that many uncountable nouns are included in the candidate lists, illustrating the importance of high quality countability information. Evaluation on CGN furthermore showed a considerable increase in recall compared to the raw frequency baseline, but with a recall of 55.3% on spoken language and 64.7% on written data, there is still plenty of room for improvement. It was shown that PP-Ds selected for by prepositions and PP-Ds composed of idiosyncratic preposition-noun combinations were the most frequent PP-D types: the two classes made up about 90% of the extracted corpus data.

In this paper, we made categorical decisions about prepositions and nouns: either they were classified as a particular type of PP-D or they were not. But virtually all preposition-noun combinations also occur *with* a determiner. Which items were included and which were excluded depended on the setting of parameters, such as N in selecting the N highest ranked PP-Ds with idiosyncratic NPs. Adjusting the parameters to include more candidates increased coverage. However, the more preposition-noun combinations are allowed to form a PP-D, the smaller the expected effect on ambiguity and the less ungrammatical PP-Ds we exclude from being parsed or generated. This trade-off between coverage and effect may be avoided if the rankings of the candidates are interpreted as weights that indicate the probability of that preposition-noun combination to participate in a PP-D.