Chapter 1

IN SEARCH OF A SYSTEMATIC TREATMENT OF DETERMINERLESS PPS

Timothy Baldwin,[1] John Beavers,[2] Leonoor van der Beek,[3] Francis Bond,[4] Dan Flickinger[2] and Ivan A. Sag[2]

[1] *University of Melbourne and NICTA Victoria Laboratory,* [2] *Stanford University,* [3] *Groningen University,* [4] *NTT Communication Science Laboratories*

tim@cs.mu.oz.au, {jbeavers,danf,sag}@csli.stanford.edu, vdbeek@let.rug.nl, bond@cslab.kecl.ntt.co.jp

Abstract This paper examines determinerless prepositional phrases in English and Dutch from a theoretical perspective. We classify attested P + N combinations across a number of analytic dimensions, arguing that the observed cases fall into at least three distinct classes. We then survey four different analytic methods that can predict the behaviour of the differing classes and examine various remaining difficult cases that may remain as challenges.

Keywords: determinerless PP, multiword expression, selection, noun countability

1. Introduction

There is a growing appreciation of multiword expressions (MWEs) as an obstacle to automated language understanding (Sag et al., 2002; Calzolari et al., 2002). In this paper, we highlight some of the peculiarities of MWEs, focusing on determinerless prepositional phrases (PPs). We then outline an analysis than can be used to systematically handle the phenomenon.

Determinerless PPs (henceforth PP−Ds) are defined to be made up of a preposition (P) and a singular noun (N_{Sing}) without a determiner (Quirk et al., 1985; Huddleston and Pullum, 2002), as in Table 1.1, organised roughly by semantic type (cf. Stvan, 1998). In the case that the noun is countable (e.g. *by bus, in mind*), a syntactically-marked structure results as the noun in itself does not constitute a saturated NP. This poses a problem for both parsing and generation unless we have

some explicit treatment of this unexpected grammaticality. Orthogonally, PP−Ds can occur with idiosyncratic semantics (e.g. *at bay* and *in kind*) which a system must have prior knowledge of to be able to analyse correctly.

PP−Ds exist in most languages with articles, and the same semantic types appear in a variety of languages: English, Albanian, Tagalog and German to name just a few (Himmelmann, 1998). Articles are generally used less frequently and less consistently in adposition phrases than in other syntactic environments. However, articles are regularly omitted in expressions of similar semantic types across languages: institution/location (*at school*), metaphor/abstract (*at large*), temporal (*in winter*), means/manner (*by car*). In this paper we will principally be concerned with English and Dutch data. Although the broad analysis is valid for other languages, the details will of course vary between languages (e.g. see Abeillé et al. (this volume) for an analysis of determinerless usages of *à* and *de* in French), and even across dialects of English (Chander, 1998).

Despite their regularities, PP−Ds tend to receive a simple 'words with spaces' treatment in lexical resources. **COMLEX**, for example, lists a total of 762 PP−Ds, in the form of a set of prepositions a given countable noun can occur with in a PP−D construction (Grishman et al., 1998). As **COMLEX** was developed as an exclusively syntactic resource, only syntactically-marked PP−Ds feature in the lexicon, and coverage tends to be patchy (e.g. in **COMLEX** 3.0, *tricycle* is listed as occurring in *via/by tricycle*, *motorbike* in only *by motorbike*, and *bicycle* has no annotated PP−D usages). **WordNet** (Fellbaum, 1998) is more ad hoc in its treatment of PP−Ds, listing around 80 PP−Ds in the adjective section and 330 in the adverb section. Predictably, the PP−Ds that are described in **WordNet** tend both to have predicative usages and to be semantically marked. The lexicon for the Japanese-to-English machine translation system **ALT-J/E** lists several classes of nouns that interact with prepositions and affect article usage, such as institutions and meals (Bond, 2001). However, the list is far from complete, and the classes are not explicitly linked to semantic classes.

To get a preliminary sense for the extent of the problem posed by PP−Ds and the relative success of **COMLEX** and **WordNet** at listing them, we carried out a semi-automated analysis of PP−D occurrences in the written component (80m words) of the British National Corpus (BNC, Burnard, 2000), using the method described in Baldwin et al.,

Institution	Media	Metaphor	Temporal	Means/Manner
at school	*on film*	*on ice*	*at breakfast*	*by car*
in church	*on TV*	*at large*	*at lunch*	*by train*
in gaol	*to video*	*at hand*	*on break*	*by hammer*
on campus	*off screen*	*at leave*	*by night*	*by computer*
at temple	*in radio*	*at liberty*	*by day*	*via radio*
...	...	...	...	...

Table 1.1. Examples of English PP−Ds, classified according to the system of Stvan, 1998

	Frequency ≥ 20						Frequency ≥ 5	
	BNC		*In* ***COMLEX***		*In* ***WordNet***			
	Types	*Tokens*	*Types*	*Tokens*	*Types*	*Tokens*	*Types*	*Tokens (%)*
as	41	7,292	0%	0%	0%	0%	484	12,686 (0.02%)
at	54	18,948	15%	17%	22%	59%	289	28,580 (0.04%)
by	71	8,327	35%	48%	1%	1%	1,023	15,493 (0.02%)
in	237	113,235	29%	45%	9%	14%	1,918	113,582 (0.13%)
on	99	25,097	26%	44%	7%	9%	964	28,204 (0.04%)

Table 1.2. Coverage and corpus occurrence of English PP−Ds

2003.[1] Focusing on the prepositions *as, at, by, in* and *on*, we first manually inspected all extracted PP−Ds which occurred at least 20 times in the corpus, and removed syntactically and semantically unmarked PPs (e.g. *at midnight*). These post-corrected sets were used to estimate the type and token coverage of **COMLEX** and **WordNet** over PP−D types in the BNC. Based on the relative error rates in each of these sets, we estimated the type and token frequencies of PP−Ds occurring at least 5 times in the BNC. The final results are presented in Table 1.2.

The coverage figures for **COMLEX** and **WordNet** vary according to the preposition, but **COMLEX** tends to have a token coverage of around 30% and **WordNet** a token coverage of around 15%, underlining the inadequacies of the two lexical resources with respect to PP−Ds. Turning next to the type and token frequency estimations, it becomes apparent that PP−Ds are a significant phenomenon in the BNC (accounting for over 0.2% of all tokens[2]). In summary, PP−Ds are surprisingly common in corpus data, and are treated inconsistently in lexical resources.

[1] Both countable and uncountable nouns were included in this data.
[2] Here, the percentages are calculated relative to the total token count in the BNC, not just tokens of frequency ≥ 5.

2. The Syntax of Determinerless PPs

The syntax of PP−Ds is not uniform. The constructions differ in their level of syntactic markedness, productivity and modifiability. On the one extreme, we have (typically Latinate) MWEs that are historically P + N combinations (*ex cathedra*, *ad hominem*, *ad nauseum*, etc.) but which, despite the erudition of certain speakers, are still best analysed as fixed expressions (Sag et al., 2002). These constructions are non-productive and non-modifiable. On the other extreme are fully productive and modifiable combinations of P + complement, where lexical selections[3] interact with a general head-complement construction to build standard PPs with compositional semantics (*per recruited student that finishes the project*). Much of English lies in between these two extremes.

We classify PP−Ds primarily in terms of their syntactic markedness, dependent largely on the nature of the prepositions and the uses of the nouns outside of these PPs. Syntactically unmarked PP−Ds are those where the N_{Sing} can occur without a determiner outside of the PP (i.e. the N_{Sing} is uncountable). For example, some prepositions select for an argument that is unbounded (uncountable or plural countable), e.g. *out of generosity* in English and *uit vrijgevigheid* "out of generosity" in Dutch. The determinerless nature of these PPs is not surprising and since these PPs are not marked syntactically (and often not semantically either as we'll discuss in the next section) they do not pose a significant problem for a (computational) grammar. A more interesting group is institutions (the social/geographic spaces in Stvan, 1998), which appear to be semi-productive. Some prepositions like *in* can combine with a range of these nouns (*in church, in school, in court, in gaol*), although other members of the same semantic class are not necessarily possible (**in library*, although context often improves these readings). However, this contrast mirrors the contrast between *school is over* and **library is over*: the nouns that can appear in this type of PP−D can also appear without a determiner outside of PPs, and in this way these PP−Ds are not syntactically marked.

On the other hand, there are prepositions that specifically require their argument to be both determinerless and countable, causing the PP to be syntactically marked. An example is the preposition *per* in both English and Dutch. Most prepositions do not specify the countability of their argument, so that the PP−Ds are sometimes syntactically marked (with a countable noun) and sometimes unmarked (with an uncountable noun). For example, means/manner *by* as in *by car*, *by computer*,

[3]Prepositions typically (but not always) select for an NP complement.

takes a wide, productive class of normally countable nouns that almost never occur without determiners. These are syntactically marked in the sense that the noun otherwise would require a determiner. But the same preposition combines with an uncountable noun in the syntactically unmarked PP *by public transportation*.

Another factor relevant to syntactic markedness is modifiability, and here most PP−Ds lie in the middle of the spectrum (Ross, 1995). Except for the fixed expressions mentioned above, most PP−Ds are modifiable to some extent. At the two extremes of modifiability are PP−Ds that allow no modification at all (*of course*, *in *children's/*mental/*small hospital*[4] and Dutch *in principe* "in principle") and PP−Ds that obligatorily require modification (*at great/public/considerable expense*, *for good/safe measure* and *op vreemde/Nederlandse bodem* "on foreign/Dutch soil", but not **at expense*, **for measure* or **op bodem* "on soil"). Between these two extremes, some PP−Ds only allow idiosyncratic modification (*at long/*great/*short last*), while others allow modification more freely (*at great/considerable/tedious/epic length*). Overall, though, modification is seldom unrestricted (in which case it tends to occur with fully productive constructions, e.g. *per recruited student that finishes the project* (from above)), and on this criterion virtually all PP−Ds are somewhat marked.

We can get some sense of the distribution of PP−Ds across the spectrum of relative modifiability by analysing the probabilistic predictability of modification patterns of different PP−D types. This is achieved via a process of cross-validation, whereby we partition up the BNC data into 10 contiguous segments of equal size, and compare the distribution of modifiers for a given PP−D in each of the 10 segments as compared to the remaining 9 segments. We determine the normalised distribution of modifiers in each case and calculate the Kullback Leibler (KL) divergence between the two distributions to determine their relative fit, averaging over the 10 iterations of cross-validation to attain a single divergence (D) value. Where the two distributions are identical, i.e. the exact same modifiers occur with the same relative occurrence, the KL divergence is 0, and failing this the magnitude of the divergence reflects the relative mismatch of the two distributions. In practice, there is a high correlation between the relative scope of modification and the KL divergence value as relative freedom of modification gives rise to greater variance in both the range of modifiers observed in a given partition and the relative frequency of each. By correlation, therefore, PP−Ds with

[4] *In hospital*, and the indicated judgements for modifiability, are particular to British English.

Prepositional Phrase	Divergence $D(PP\|\|PP)$	$D(PP\|\|NP)$
on horseback	0.00	0.04
before dawn	0.00	0.16
in reverse	0.00	0.51
by contrast	0.00	0.71
to hospital	0.02	0.32
into bed	0.02	0.56
up front	0.03	0.26
by marriage	0.05	0.29
on trial	0.07	0.21
on record	0.10	0.76
in readiness	0.11	0.50
in diameter	0.14	0.54
in school	0.18	0.26
on loan	0.18	0.71
in isolation	0.19	0.83
in disgust	0.22	0.34
in depth	0.27	0.50
in tone	0.87	1.08
by decree	1.62	2.07
on analysis	4.29	2.81

Table 1.3. A random sample of 20 PP−Ds occurring ≥ 100 times in the BNC

low KL divergence have restricted modifiability, and tend to occur unmodified the bulk of the time. In addition to analysing KL divergence relative to other instances of the same PP−D ($D(PP||PP)$), we calculate the divergence over NPs not selected for by prepositions ($D(PP||NP)$). This provides some insight into the relative markedness of modification relative to non-PP occurrences of the same noun. That is, we would expect to see relative low divergence for productive PP−Ds due to their greater compositionality, and relatively high divergence for PP−Ds with marked syntax and/or semantics.

In Table 1.3, we present a random sample of 20 PP−Ds occurring with frequency 100 or greater in the BNC, in increasing order of $D(PP||PP)$. Items higher in the list can be seen to resist modification, which in the case of *horseback*, e.g., is consistent with its behaviour outside of PP−Ds, whereas with *contrast*, the lack of modification appears particular to PP−Ds. At the end of the list, we see that with *on analysis*, there is greater variability in modification within the PP−D data than relative to non-PP usages. The relative increase in the value of $D(PP||PP)$ is slow, indicating that even for PP−Ds with scope for modifier variation, actual variation tends to be slight.

For PP−Ds that allow modification, there can be additional constraints on the word class of the modifier. Some PP−Ds allow only noun-noun compounds, as with *at eye/street level* but not **at higher level*, while others allow only adjective modifiers, as with *in sharp/pointed/rich contrast* but not **in color contrast*. The two dimensions of choice of modifier (noun, adjective, or either) and presence of the modifier (obligatory, impossible, or optional), combine to present seven logically possible subclasses of PP−Ds (since the subclass that disallows modifiers is indifferent to the dimension of modifier choice), as shown in Table 1.4. Each of these logically possible subclasses is instantiated in the BNC data. Other languages may have different constraints on modification. Some Dutch prepositions allow morphological but not syntactic modification, but select for a bare noun at the same time. Here, the prepositional object can only be modified via morphological rules, by forming a complex N (*op zeilkamp* "at sailing camp", *op ponykamp* "at pony camp" and *op schoolkamp* "at school camp", but not **op sportief kamp* "at sporty camp").

	Obligatory	Optional	Impossible
Noun	*at *(eye) level*	*on (summer) vacation*	
Adjective	*at *(long) range*	*in (sharp) contrast*	*on (*very) top*
Either	*at *(company) expense* *at *(considerable) expense*	*in (family) court* *in (open) court*	

Table 1.4. Variation in modification of determinerless PPs

Despite this rich spectrum of syntactically distinct PP−Ds, there are still some constructions that don't seem to fit in. In the first place there are some prepositional constructions consisting of two prepositions with determinerless arguments: *from X to Y, X by X*, e.g. *from mother to child*, *room by room*. Secondly, features of determinerless constructions may be distributed over both conjuncts of a coordination where only one fulfils the selectional requirements of the preposition. For example, *in* does not readily occur with the noun *brush* in a PP−D, and yet the coordination *in brush and ink* is perfectly acceptable (noting that *in ink* is also a grammatical PP−D). Finally, there is a class of coordinated PP−Ds in Dutch where neither one of the coordinated nouns can occur independently in a determinerless PP (e.g. *over mens en wereld* "about human being and world", *van stadion en hotel* "of stadium and hotel").

3. The Semantics of Determinerless PPs

Turning to the semantics of PP−Ds, Stvan, 1998 focused primarily on four natural semantic classes of nouns and a relatively small set of

prepositions (mostly locatives like *at* and *on*), classifying them by possible implicatures (or enrichments of content) and contrasts with full NPs. However, looking at a broader set of data shows considerable systematicity along many other semantic dimensions, and in this section we will highlight some of these relevant categories and outline a general classification of PP−Ds based on semantic markedness. As noted above, all PP−Ds show a certain degree of markedness in the form of metaphorical (*on ice* in the non-literal sense), institutionalised (*at school*), and generic uses (*by car*), which in many (but not all) cases is different from the basic simplex semantics of these nouns. Relative to this, however, they seem to follow a cline of markedness dependent on both lexical semantics and the overall compositionality of the PP, with certain natural semantic classes often clustering together.

Among the least marked semantic classes of PP−Ds are those formed with institutionalised nouns such as *in town*, *at school*, *at church*, a sizeable subset of Stvan's social/geographic spaces, which in the previous section were identified as the least syntactically marked since they occur both in and out of PP−Ds without determiners. Corresponding to this distributional property, not surprisingly, are similar semantic effects. In PP−Ds, these show a variety of special semantics including what Stvan refers to as activity and familiarity implicatures. Activity implicatures (or enrichments of content) occur when the PP seems to be referring to an activity associated with the institution, rather than a specific place (e.g. *in gaol* "while being a prisoner" and *in school* "while attending school", which can even be true of someone not located at a school, as opposed to *at a gaol/school* which is a simple locative). Familiarity arises from uses that seem to refer to specific entities familiar to a participant in the discourse (e.g. *John is in town* "John is in (my/his) town", as opposed to *John is in the/a town* which again is a simple locative).[5] However, most nouns in this institutionalised class have corresponding N_{Sing} non-PP uses that induce the same semantic effects, as in (1) (note that (1c) is particular to American English dialects where *school* can be synonymous with *university*):

(1) *a. While at school[=attending school], I learned the value of an education. (Complement of preposition)*

[5]This enrichment of content, however, seems to be somewhat intertwined with the 'activity implicature', since you can have this anaphoric reference even in activity senses, as in *his hair went grey in gaol*, which could mean *his hair went grey while serving time in his gaol* thus showing both enrichments. In other cases this is necessarily the case, as in *they had a bad day at work[=working at their workplace]*. In this regard the data is somewhat murky.

b. *School[=attending school] drains the best years of your life. (Subject)*

c. *Many students can't afford school[=to attend school] in the States. (Object)*

In (1) each use of *school* can induce the same reading, in this case the activity [enrichment], and likewise for other uses, like familiarity [enrichment] (e.g. *work wore him out* where *work* can be replaced by *his work*, as well as *working*).[6] Given the persistence of this kind of specialised semantics, their universally determinerless nature, and the large size and semi-productivity of this noun class, the semantics of these PP−Ds is unsurprising and thus relatively unmarked, being entirely predictable from the N. The fact that institutional nouns can occur without determiners in these environments is, however, a peculiarity of English; related Germanic languages such as German and Swedish require the definite article here (Himmelmann, 1998). Dutch examples of institutional nouns that can occur in determinerless environments are *school* "school" and *kantoor* "office". These examples show activity and familiarity implicatures similar to the English examples, but are less modifiable and less numerous. Norwegian has the intriguing property that PP−Ds tend to occur only in institutionalised contexts, e.g. the determinerless *i hengekøye* "in hammock" is grammatical only in combination with a verb such as *sove* "sleep" (Borthen, 2003).

Other nominal classes show varying degrees of semantic markedness, such as Stvan's class of media expressions, e.g. *in print*, *on film*, *on video*, involving media-related nouns. Here, too, we see similar nominal semantics in and out of PPs:

(2) a. *The Manchurian Candidate is my favourite film.*
[sense=content] [form=countable]

b. *I'd rather watch it on film than rent the video.*
[sense=material] [form=uncountable]

c. *I would always rather watch a film than a video.*
[sense=media form] [form=countable]

(Stvan, 1998)

In (2), *film* shows similar readings (specifically broadcast/media type, material, and content type) in a variety of positions, again showing a low

[6]This goes against Stvan, who argues that such nouns in subject position do not show familiarity, although as noted in fn. (5) the data in general isn't so clear.

degree of semantic markedness. However, unlike the institutional class, these uses rarely occur without determiners outside of PPs (although sometimes this is possible, e.g. *TV rots your brain* [sense=content]), indicating some degree of syntactic markedness. Another of Stvan's classes is "temporal interruptions", where the noun identifies a specific break in a particular routine, subdividing into two classes: shorter breaks marked by *at* (e.g. *at lunch*) and longer, more open-ended breaks with *on* (e.g. *on leave*). The nouns associated with short breaks occur frequently in similar uses outside of PP−Ds (e.g. *lunch starts at noon*), indicating less semantic markedness, whereas longer breaks involve nouns that rarely do (e.g. *??vacation lasts longer each year*,[7] **we want more holiday in our work year*), indicating more semantic markedness.

On the other end of the markedness scale is a class of non-compositional and relatively metaphorical PP−Ds, including *at hand* and *on ice*, largely corresponding to what Stvan labels "untethered metaphors", i.e. expressions formed by nouns that define states and generally have no referential properties. However, despite their non-compositionality, not all of these PPs are semantically unpredictable. In particular many adverbial and adjectival PP−Ds have synonymous, morphologically related adverb or adjective pairs, e.g. *lastly/at last*, *willfully/at will*, *effectively/in effect* and *handy/on/at hand*, *edgy/on edge*. While still idiosyncratic (e.g. *edgy/on edge* "nervy/excitable" is not entirely predictable from *edge*) the semantic relationship between these morphologically derived and analytic noun-centred forms is striking, showing some systematicity if not predictability.

Similarly, although prepositions do not cluster into fine-grained semantic classes like nouns, they show various semantic properties relevant to their distributions within PP−Ds. A significant number of spatial prepositions (e.g. *at*, *to*, *on*, etc.) occur in PP−Ds, in both temporal and stative uses, although this is hardly surprising since cross-linguistically spatial prepositions frequently grammaticise into temporal and stative/metaphorical uses independent of PP−D constructions (correspondingly to a low degree of markedness) (see e.g. Haspelmath, 1997). However, there are further semantic dimensions within these broader semantic classes. For example, a variety of interesting patterns are seen in antonymous pairs of prepositions. With locative prepositions, several antonymous pairs show stark differences in their distribution, e.g. *on/off*, *in/out*, *at/away (from)*, *near/far (from)*, etc. In our corpora, the inclusive or positive prepositions (e.g. *in*, *on*) were among the high-

[7] Acceptable in some American dialects

est frequency heads while the negative pairs were generally much rarer (there were surprisingly few corpus examples involving *off*, *out* and *away (from)*, although these certainly do exist, e.g. *off base*, *away from town*). Interestingly, antonymous pairs for which neither preposition had an inclusive/positive reading tended to show up infrequently, e.g. the relative infrequency of PP−Ds headed by *down/up*, *before/after*. Other antonymous pairs showed further interesting relationships. In our corpora, the relative frequency of *without* with uncountable nouns in generic readings (e.g. *without success*, *without fear*, *without help*) was roughly double that of *with*. Therefore it appears that cross-cutting semantic features such as inclusiveness/exclusiveness and negative polarity also play a role in the semantic regularity of PP−Ds. Synonymy, on the other hand, does not appear to be a relevant factor in determining grammaticality of PP−Ds. For example, *by* as in *by law*, where *by statute* is grammatical but not *??according to law* and *??according to statute*. This further highlights the generally lexicalised nature of PP−Ds. Crosslinguistically, primary adpositions (short monomorphemic adpositions with grammatical meanings) are more likely to be involved in PP−Ds than secondary adpositions (longer or complex adpositions with concrete meanings) (Himmelmann, 1998).

Finally, idiosyncratic prepositions sometimes form classes of PP−Ds all of their own. One of the most regular semantic classes is means/manner *by*, most of whose members are vehicular (e.g. *by car*, *by train*) although not always (e.g. *by hand*, *by post*, *by telephone*). In general these resist referential uses and familiarity enrichments, although they do allow generic and activity readings:[8]

(3) *I travelled to San Francisco by car. They're/It's a great way to travel/#It rattled a lot.*

Such PP−Ds tend to be nonreferential and more semantically marked than the institution class since most of these nouns rarely occur with the means/manner semantics in subject/object position (although it is possible, e.g. *car costs less than train for trips to the city*). On the other hand, this class shows a high degree of internal systematicity, particularly in excluding related readings with determiners (e.g. **by a/the car*) and some amount of productivity (e.g. *I arrived yesterday by carpet* in a context of having a flying carpet – see Section 1.4). These are just a few of the myriad levels of (semi-)regularity in the PP−D system. Al-

[8]PPs headed by *by* (and *via*) are not the only means/manner PPs, e.g. *on foot*, however we assume that cases such as this, which are non-productive and idiosyncratic, should be lexicalised.

though previous work has focused primarily on systematicity in relation to natural semantic class of the N_{Sing} and the small set of possible interpretations, it appears there is a wider set of generalisations, taking into account basic semantic features of the prepositions and broader lexical classes inside and outside of PP−Ds.

4. Analysis

As noted in the introduction, the coverage of existing resources is unsystematic and generally limited to more or less fixed preposition-noun combinations. We will introduce three more analyses to complement this: occurrence with defective noun phrases, selection for idiosyncratic noun phrases and selection for nominal phrases ($\bar{N}$s) by the preposition. Each of the four kinds of analyses is well suited for a large class of PP−Ds. The analyses are given in the framework of Head-driven Phrase Structure Grammar, and have been tested by implementing them in the English Resource Grammar (Flickinger et al., 2000), although with only with a few examples of each kind.

In all three kinds of syntactic analysis, the familiar HPSG head-complement construction will license all the PP−Ds in question. But the differing lexical specifications will modulate the relevant distributions appropriately. For any given PP−D, there should always be evidence (modification, productivity) to tell if it is to be lexically listed or treated syntactically. If it is treated syntactically, then there should be further evidence showing whether the prepositional object is a freely combining NP, a modified nominal phrase with idiosyncratic restrictions on the presense or type of modifier or a specially selected, unsaturated nominal phrase ($\bar{N}$) (the determinerless NP in non-prepositional contexts, restricted choice of P).

4.1 Lexical Listing

Lexical listing is the obvious approach for the syntactically and semantically marked class (e.g. *at large*, *on track*). For expressions such as these, it is entirely sufficient to simply list the P + N combinations in the lexicon, since the combination is non-productive and largely non-modifiable. In addition, the semantics is non-compositional and uniquely associated with a particular PP−D. Lexical listing is a simple approach that accurately reflects the inflexibility of these PPs.

For the other types of PP−Ds, lexical listing is more problematic. First, modification of the nominal within the PP can be possible (e.g. *as* **former** *president*, *at* **considerable** *length*). Simple lexical listing cannot handle this. Second, the syntactically marked class, e.g. *by car*,

by train, *by taxi*, is productive, which also makes a simple listing in the lexicon impossible. Moreover, the semantically unmarked constructions have compositional semantics. Hence any attempt to treat the preposition and noun as a multiword lexical unit would fail to express this compositionality. Finally, some of the PP−Ds (or rather the nominals within them) select for an optional prepositional complement (e.g. *in front of the children*). This selection is also hard to capture via simple lexical listing.[9] Within a syntactic approach, one might consider positing a general rule: NP → $\bar{\text{N}}$. However, such a rule would massively overgenerate, as any noun would be allowed to occur sans determiner in any context. Even if the rule were restricted to PP contexts, it would overgenerate, as not all prepositions and not all nouns allow the determinerless combination. Therefore, it would appear that a more fine-grained treatment is needed.

4.2 Prepositions that occur with Defective NPs

Some PP−Ds can be analysed as simple syntactic combinations of a preposition and an NP complement. The NP itself is defective and has no determiner. The key motivation for such an analysis, as noted in Section 1.3 above, is the fact that these noun phrases appear without a determiner in other (semantically appropriate) syntactic contexts, e.g. as subjects and objects. For example, *church*, *school*, etc. are countable nouns that refer to (sets of) churches, schools, etc. But these give rise to the determinerless noun phrases *church*, *school*, etc. that refer to the relevant church and school activities: for example *School is over*, cited in Section 1.2 above. Our account of these PP−Ds requires no new apparatus: since the determinerless noun phrases exist independently as subject and object NPs, it follows that they should also appear as prepositional objects in a standard head-complement construction. The semantics seems equally straightforward, in that the semantic composition of *in school* acquires the interpretation "in the appropriate school-related activity" in just the same way that *likes school* acquires its "likes the appropriate school-related activity" interpretation, as discussed in Section 1.3. This analysis also predicts that the determinerless NP in question will not be restricted to a single preposition. Though certain P + N combinations may give rise to semantic incompatibility, the general prediction made by this analysis seems right for this class of expression,

[9] An alternative approach to these transitive PPs is to analyse them as complex prepositions (prepositions with spaces). According to this analysis, *on top* is similar to *inside*, except that the former selects for a PP[*of*] and the latter for a complement that is either an NP or a PP[*of*].

given that *in/at/after/before/during school* are all well-formed and easily interpretable.

4.3 Prepositions that select idiosyncratic NPs

Next, we present an analysis for the more idiomatic PP−Ds where the nouns can take only a restricted set of modifiers. In this case the idiosyncratically modified nouns also construct defective noun phrases, but they are constrained to only appear as complements of prepositions, as with *at eye level* or *at considerable expense*.[10]

The syntactic analysis employs three unary rules similar to the bare-NP rule used for constructing full determinerless NPs from ordinary mass or plural nominal phrases. For each of these three additional rules, the daughter is constrained to be headed by a particular subclass of nouns, idiosyncratically marked in the lexicon for the property of being modifiable by a noun, an adjective, or neither. Two of the three rules require that the daughter be a nominal phrase containing a (pre-head) modifier, while the third rule constrains the daughter to be unmodified. On this account, a phrase like *at eye level* is thus analyzed as a head-complement structure combining the ordinary preposition *at* with the determinerless NP *eye level*, where this NP is constructed via a unary rule which constrains the daughter to be lexically headed by a noun which permits nominal modification, and moreover this daughter must indeed contain a modifier. The lexical entry for this idiosyncratic *level* is distinct from the entry for the ordinary count noun *level*, and is constrained so that (1) phrases that it projects will only appear as complements of prepositions, (2) its specifier (determiner) will never be expressed, (3) it must combine with a (pre-head) modifier before it can combine with the preposition, and (4) it can only appear with a nominal modifier, not an adjective. Of course, this analysis only ensures that the syntactic constraints are correctly imposed on these subclasses of PP−Ds containing modified nominals. We will still require additional semantic collocational constraints analogous to those for semi-productive idioms (cf. Riehemann, 2001), in order to reflect the collocational restrictions on which specific prepositions combine with which of these modified nouns, and which modifiers are possible.

[10]Note that we adopt a somewhat unconventional treatment of noun–noun compounds such as *eye level*, in treating the first noun as a modifier of the second.

4.4 Prepositions that select $\bar{\text{N}}$

The approaches just sketched will not extend to the productive constructions discussed earlier (e.g. *by car, as president*) in which a particular preposition (or preposition class) selects for an exclusively countable noun that cannot project a determinerless NP in other syntactic contexts:

(4) a. *They arrived by train/plane/bus/pogo stick/hydro-foil ...*

b. **I really like train/plane/bus/pogo stick/hydrofoil*

c. **Train/plane/bus/pogo stick/hydrofoil could save us money.*

When there is no evidence that a PP−D contains an NP-projecting uncountable noun, then it makes sense instead to posit a lexical entry or lexical type of preposition constrained as in (5):

(5)
$$\left[SYN \left[CAT \left[\begin{matrix} HEAD & prep \\ VAL & \left[COMPS \left\langle \left[SPR \langle\ Det\ \rangle \right] \right\rangle \right] \end{matrix} \right] \right] \right]$$

Prepositions of this type select a complement whose specifier is of type *Det*. As only nouns have specifiers of type *Det*, and NPs have an empty specifier, the complement is constrained to be an $\bar{\text{N}}$. By positing an entry of this sort for (one sense of) the preposition *by*, we can account for its special ability to combine with determinerless (unsaturated) nominal phrases that denote means/instruments but wouldn't normally occur in this interpretation. Crucially, in all such cases, the determinerless nominal is restricted to the preposition *by*, as predicted:

(6) **They arrived with/in/to train/plane/bus/hydrofoil/pogo stick...*

These productive PP−Ds seem further restricted to particular semantic domains, e.g. *on* + MEDIUM or *by* + MEANS/INSTRUMENT. These restrictions could be the result of selection for specific semantic classes of nouns by the preposition or they could alternatively be interpretations entirely contributed by the preposition on top of the nominal semantics. The Dutch construction *in* + PIECE OF CLOTHING is ungrammatical with anything that is not established as clothing and thus seems to suggest the former. However, examples like *From the train station to Hogwarts is 15 minutes by broom* suggest that the preposition supplies the interpretation, although it is a matter of descriptional granularity

and/or domain-specificity as to whether the noun enables a matrix transportation interpretation or not.[11]

4.5 Summary of Analysis

Finally, although we have suggested that there are three distinct kinds of analysis, there are a number of cases that present challenges to this simple picture of the world of PP−Ds. For instance, there are many different PP−Ds with the English nouns *sea* and *hand* or the Dutch nouns *zee* "sea" and *huis* "house". These PPs are semantically unmarked (the meaning is fully compositional) but syntactically marked (the nouns do not occur without a determiner outside of PPs). These are distinct from the *by car* type in that the determinerless P + N combination is not restricted to a particular preposition (e.g. *at sea*, *to sea*, *from sea to* ..., %*by sea*, **in sea*, **over sea*, ...). Perhaps these are idioms, whose common properties must be relegated to linguistic history; or perhaps there is some fine-grained semantic analysis that will account for the restricted distribution in synchronic terms. The work of Soehn and Sailer, 2003 provides a third alternative: an analysis in terms of selectional restrictions imposed by the noun. Our hope is that no such stipulations are required within an adequate grammar: in each such case there is some factor or factors to be discovered that interacts with the pristine picture of PP−Ds that we have sketched here.

5. Conclusion

We have presented PP−Ds as a commonly occurring, highly varied form of multiword expression, and documented their idiosyncratic syntax and semantics. Depending on the type of PP−D, one of four analyses was proposed: simple lexical listing, occurrence of the preposition for independently existing determinerless NPs, selection for idiosyncratic determinerless NPs or selection for nominal phrases ($\bar{\text{N}}$s). The analyses we have outlined cover a wide area, but do have yet to be reconciled with the full range of idiosyncratic restrictions on P + N combination that have been observed in the literature.

We have implemented these analyses in a computational grammar. The next step in our research is to extract determinerless PPs from corpora in volume and analyse each for such properties as modifiability and referentiality. Using this as a guide, we can determine the robustness

[11] Such an analysis could also be extended to cases of *from X to Y* and *like X like Y*, e.g. *from town to town* or *like father like son* by assuming *from*/*like* takes two complements, an $\bar{\text{N}}$ and a particular PP, providing the appropriate semantic relationship between them.

of the proposed analyses over open data and build up a rich inventory of lexicalised PP−Ds to supplement existing resources.

Acknowledgments

This material is based in part upon work supported by the National Science Foundation under Grant No. BCS-0094638 and also the Research Collaboration between NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation and CSLI, Stanford University. We would like to thank Ann Copestake and the anonymous reviewers for their valuable input on this research.

References

Baldwin, Timothy, Beavers, John, van der Beek, Leonoor, Bond, Francis, Flickinger, Dan, and Sag, Ivan A. (2003). In search of a systematic treatment of determinerless PPs. In *Proc. of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France.

Bond, Francis (2001). *Determiners and Number in English, contrasted with Japanese, as exemplified in Machine Translation*. PhD thesis, University of Queensland, Brisbane, Australia.

Borthen, Kaja (2003). *Norwegian Bare Singulars*. PhD thesis, Norwegian University of Science and Technology.

Burnard, Lou (2000). *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Calzolari, Nicoletta, Fillmore, Charles, Grishman, Ralph, Ide, Nancy, Lenci, Alessandro, MacLeod, Catherine, and Zampolli, Antonio (2002). Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1934–40, Las Palmas, Canary Islands.

Chander, Ishwar (1998). *Automated Postediting of Documents*. PhD thesis, University of Southern California, Marina del Rey, CA.

Fellbaum, Christiane, editor (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Flickinger, Dan, Oepen, Stephan, Uszkoreit, Hans, and Tsujii, Jun'ichi (2000). Journal of natural language engineering (special issue on efficient processing with hpsg).

Grishman, Ralph, Macleod, Catherine, and Myers, Adam (1998). *COMLEX Syntax Reference Manual*. Proteus Project, NYU.

Haspelmath, Martin (1997). *From Space to Time in The World's Languages*. Lincom Europa, Munich, Germany.

Himmelmann, Nikolaus P. (1998). Regularity in irregularity: Article use in adpositional phrases. *Linguistic Typology*, 2:315–353.

Huddleston, Rodney and Pullum, Geoffrey K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge, UK.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, and Svartvik, Jan (1985). *A Comprehensive Grammar of the English Language*. Longman, London, UK.

Riehemann, Susanne (2001). *A Constructional Approach to Idioms and Word Formation*. PhD thesis, Stanford, USA.

Ross, Háj (1995). Defective noun phrases. In *Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, pages 398–440.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan (2002). Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Soehn, Jan-Philipp and Sailer, Manfred (2003). At first blush on tenterhooks. About selectional restrictions imposed by nonheads. In Jäger, Gerhard, Monachesi, Paola, Penn, Gerald, and Winter, Shuly, editors, *Proceedings of Formal Grammar 2003*, pages 149–161.

Stvan, Laurel Smith (1998). *The Semantics and Pragmatics of Bare Singular Noun Phrases*. PhD thesis, Northwestern University.