In Search of a Systematic Treatment of Determinerless PPs

Timothy Baldwin[†], John Beavers[†], Leonoor van der Beek[‡], Francis Bond[*],
Dan Flickinger[†] and Ivan A. Sag[†]

† CSLI
Stanford University
Stanford, CA 94305 USA
{tbaldwin,jbeavers,danf,sag}@csli.stanford.edu

‡ Alfa-informatica
RuG, Pb 716
9700 AS Groningen
The Netherlands
vdbeek@let.rug.nl

∗ NTT Communication Science Labs
2-4 Hikari-dai, Seika-cho, Soraku-gun
Kyoto, 619-0237, JAPAN
bond@cslab.kecl.ntt.co.jp

Abstract

This paper examines Determinerless PPs in English 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 three 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 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 & Pullum 2002), as in Table 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 *down pat*) 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

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: Examples of PP−Ds

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*).

Despite these 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 (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 Section 5.1.[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 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.

The remainder of the paper is structured as follows. We describe the syntax and semantics of PP−Ds in Sections 2 and 3 respectively. In Section 4 we sketch our analysis, and we present a

[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.

Preposition	FREQUENCY $\geq$ 20						FREQUENCY $\geq$ 5	
	BNC		*COMLEX coverage*		*WordNet coverage*			
	Types	*Tokens*	*Types*	*Tokens*	*Types*	*Tokens*	*Types*	*Tokens (%)*
as	41	7,292	.00	.00	.00	.00	484	12,686 (0.02%)
at	54	18,948	.15	.17	.22	.59	289	28,580 (0.04%)
by	71	8,327	.35	.48	.01	.01	1,023	15,493 (0.02%)
in	237	113,235	.29	.45	.09	.14	1,918	113,582 (0.13%)
on	99	25,097	.26	.44	.07	.09	964	28,204 (0.04%)

Table 2: Coverage and corpus occurrence of PP−Ds

number of methods for extracting PP−Ds in Section 5. Our results are summarised in Section 6.

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, and the data is not uniform across dialects (Chander 1998).

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). One such 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 office*, although context often improves these readings). However, this contrast mirrors the contrast between *school is over* and **office is over*: the nouns that can appear in this type of PP−D have an uncountable sense and can therefore also appear without a determiner outside of PPs, and in this way these PP−Ds are not syntactically marked. Likewise some prepositions simply 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. Again, 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.

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 un-

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

marked (with an uncountable noun). For example, means/manner *by* as in *by car*, *by computer*, 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. Some allow no modification at all (*in *children's/*mental/*small hospital*[4]), some only allow idiosyncratic modification (*at full/complete/?total/?absolute/*partial liberty*), and 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.

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'll 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 metaphoric (*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",

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

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. (PP Comp)

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 1998:331). 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:p.127, (43))

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

[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.

[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.

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 large, on ice, at leave*, 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.

It likewise appears that 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, inboth temporal and stative uses, although this is hardly surprising since cross-linguistically spatial prepositions frequently grammaticalise into temporal and stative/metaphoric uses independent of PP−D constructions (correspondingly to a low degree of markedness) (see e.g. Haspelmath (1997)). However, within these broader semantic classes there appear to be further relevant semantic dimensions. 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 highest frequency heads while the negative pairs were generally much rarer (virtually no examples occurred in our corpora with *off, out, away (from)*, although these do exist, e.g. *away from work, out of 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*. One possible explanation for this is that generic readings are more commonly attested in negative contexts. 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.[8] 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 1998:319).

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 which 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:[9]

[7] Acceptable in some American dialects

[8] 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.

[9] 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.

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

To a degree such PP−Ds are more semantically marked than the institution class since most of these nouns rarely occur with these 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 4). These are just a few of the myriad levels of (semi-)regularity in the PP−D system. Although 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. In Section 5, we explore how extraction of these fixed combinations can be automated, so that PP−Ds can be listed in a systematic fashion when the need arises. But first, we will compare a simple lexical listing approach with two other readily available alternatives that are capable of expressing more systematic properties, arguing that each of the three kinds of analyses is well suited for a large class of PP−Ds.

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.[10] Within a syntactic approach, one might consider positing a general rule: NP → $\underline{\text{N}}$. 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.

In fact, we believe that many PP−Ds can be analysed as simple syntactic combinations of a preposition and an NP complement. The key motivation for such an analysis, as noted in Section 3 above, is the fact that many nouns that are typically countable live a second life as an uncountable noun,

[10] An alternative approach to these transitive PPs is to analyse them as complex prepositions. 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*].

as shown by their ability to function 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 uncountable nouns *church*, *school*, etc. that refer to (the appropriate mereological constructs based on) the relevant church and school activities. The uncountable noun *work* may in fact not even be synchronically related to its countable homophone. In any case, our account of these requires no new apparatus: since the uncountable nouns in question exist independently, and give rise to determinerless subject and object NPs (as shown by the grammaticality of examples like *School is over*, cited in Section 2 above), 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 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, given that *in*/*at*/*after*/*before*/*during school* are all well-formed and easily interpretable.

The approach just sketched will not extend to the syntactically marked constructions discussed earlier (e.g. *by car, as president*). The nouns in these constructions are exclusively countable and cannot project a determinerless NP in other syntactic contexts:

(4) a. They arrived by train/plane/bus/hydrofoil/pogo stick...

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

c.*Train/plane/bus/hydrofoil/pogo stick could save us some 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)
$$\left[\begin{array}{ll}\text{SYN} & \left[\begin{array}{ll}\text{CAT} & \left[\begin{array}{ll}\text{HEAD} & \textit{prep} \\ \text{VAL} & \left[\text{COMPS} \left\langle \left[\begin{array}{ll}\text{SPR} & \langle\ \text{Det}\ \rangle \\ \text{KEY} & [\textit{X-rel}]\end{array}\right] \right\rangle\right]\end{array}\right]\end{array}\right]\end{array}\right]$$

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

coercion, 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]

As we have seen, we cannot treat all types of PP−Ds in a unified manner. Lexical listing would undergenerate in the syntactically unmarked or productive classes and a syntactic approach would overgenerate with respect to the fully fixed PP−Ds. We have instead proposed to lexicalise only the nonproductive, syntactically marked combinations and to account for all other combinations in two more flexible ways: either assuming the noun leads a double life as an NP with particular semantic interpretations that can combine with prepositions as other NPs, or by having the preposition select for certain nouns, imposing particular syntactic and possibly additional interpretive constraints.

In both 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 or a specially selected, unsaturated nominal phrase ($\bar{\text{N}}$) (the determinerless NP in non-prepositional contexts, restricted choice of P).

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 & Sailer (2003) provides a third alternative: an analysis in terms of selectional restrictions imposed by the noun. Our hope is that 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. Needless to say, the fulfilment of this hope rests on the detailed analysis (or reanalysis) of data examined by Chander (1998), Himmelmann (1998), and Bond (2001), among many others.

5 Extraction

Above, we have proposed lexicalisation and selection as mechanisms for capturing non-productive PP−Ds. We have also provided evidence for the existence of a wide lexical variety of these PP−D types and patchy coverage in existing resources, begging the question of exactly how feasible it is to produce a static repository of such expressions. In this section, we suggest a number of avenues that could be explored to extract PP−Ds from corpora and reduce the manual annotation overhead.

We propose two basic methods for extracting PP−Ds, drawing on distinct syntactic and semantic properties of the construction. Below, we outline each in turn. Each of the proposed methods assumes a pre-processing step to extract P + N_{Sing} bigrams from corpus data, and selects genuine instances of PP−Ds from amongst these. Pre-processing can take the form of full parsing, although here the quality of the output is conditioned on the ability of the parser to analyse PP−Ds correctly, leading

[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.

to circularity. Chunk parsing offers a more robust, knowledge-light form of analysis at the cost of not being able to resolve phrasal attachment ambiguities. By setting aside instances of PP−D internal attachment ambiguity (e.g. *at home by the sea*) for PP−Ds, we can maximise precision and focus on the unambiguous data.

5.1 Countability-based extraction

Syntactically-marked PP−Ds are characterised by the head noun being strictly countable. Assuming we have access to countability information, therefore, a simple filter on head noun countability should be sufficient to identify syntactically-marked PP−Ds. In order to carry out a countability-based classification of PP−Ds, we need access to a large-scale inventory of simplex noun countability judgements. Fortunately, such information exists in **COMLEX** and the transfer dictionaries used in the **ALT-J/E** system (Ikehara et al. 1991) for English, and the Alpino lexicon (Bouma et al. 2000) for Dutch. Alternatively, or in addition, the countability of nouns can be reliably extracted from corpora (Baldwin & Bond 2003; van der Beek & Baldwin 2003).

Countability-based extraction was the method that formed the basis of the BNC analysis presented above. As our preprocessor, we used a POS tagger and chunker built using fnTBL 1.0 (Ngai & Florian 2001), and **COMLEX** was used as the source of noun countabilities. In order to gauge the range and relative occurrence of all PP−D types, we simply flagged PP−Ds not containing strictly countable nouns, and manually inspected the nouns occurring with each preposition type in decreasing order of frequency.

5.2 Substitution-based extraction

In order to detect semantically-marked and institutionalised PP−Ds, substitution seems to offer considerable promise. Substitution involves replacing words with synonyms or near-synonyms and ascertaining whether the derived form is attested in corpus data (Pearce 2001). The expectation is that for semantically-marked PP−Ds (e.g. *on board*) the opacity of the semantics means that synonym derivatives are anomalous (e.g. **at board *on plank*). For institutionalised PP−Ds (e.g. *at work*), synonym substitution leads to jilted or blocked expressions (e.g. *#in work, *at job*) which one would not expect to find in a corpus. Synonyms or near-synonyms can be obtained from a thesaurus or other semantic knowledge-base.

5.3 Complications

The principal complication in extracting PP−Ds comes in the pre-processing step, in lifting P + N_{Sing} instances out of corpus data. As noted above, there are a number of complex constructions that license fully productive PP−Ds (e.g. *from X to Y, X to X*). Given the simple-minded nature of the methods proposed above, we have to be careful not to misanalyse the component PP−Ds of such constructions as standalone entities. This could be achieved in one of two ways: (a) to have a prescribed listing of complex constructions which lead to PP−Ds, and ignore any data corresponding to those constructions, and (b) to look at the word-entropy to the immediate left and right of the PP−D candidate and ignore all PPs where the entropy is below a certain threshold and the most commonly-occurring context is a word rather than a punctuation mark.[12]

[12] We need to relax the constraint on the entropy for punctuation marks due to the large numbers of adverbial PP−Ds which occur most prevalently sentence-initially or sentence-finally.

5.4 Automatic Analysis

The final stage of extraction is to attempt to identify regular semantic types (such as Temporal and Institution). A rough attempt can be made using the extracted data and cross classifying the prepositions against the noun's semantic classes (obtained from thesauri or other lexical knowledge bases). Better results can be expected from disambiguating the nouns in context, and extracting them labelled with their semantic class. Alternatively, the co-occurence of nouns with prepositions and articles can be used to help define semantic classes.

6 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 three analyses was proposed: simple lexical listing, selection for determinerless nominal phrases ($\bar{\text{N}}$s) by the preposition, or treating the noun as a full NP. The analyses we have outlined 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 outlined two possible methods for extracting PP−Ds from corpora, and the next step would appear to be 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 of the proposed analyses over open data and build up a rich inventory of lexicalised PP−Ds to supplement existing resources.

Acknowledgements

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 two anonymous reviewers for their valuable input on this research.

References

Baldwin, Timothy & Francis Bond: 2003, 'Learning the countability of English nouns from corpus data', in *Proc. of the 41st Annual Meeting of the ACL*, Sapporo, Japan, (to appear).

van der Beek, Leonoor & Timothy Baldwin: 2003, 'Crosslingual countability classification: English meets Dutch', *LinGO Working Paper No. 2003-03*.

Bond, Francis: 2001, 'Determiners and number in English, contrasted with Japanese, as exemplified in machine translation', Ph.D. thesis, University of Queensland, Brisbane, Australia.

Borthen, Kaja: 2003, 'Norwegian bare singulars', Ph.D. thesis, Norwegian University of Science and Technology.

Bouma, Gosse, Gertjan van Noord & Rob Malouf: 2000, 'Alpino: Wide coverage computational analysis of Dutch', in *Computational Linguistics in the Netherlands (CLIN 2000)*.

Burnard, Lou: 2000, '*User Reference Guide for the British National Corpus*', Tech. rep., Oxford University Computing Services.

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

Chander, Ishwar: 1998, 'Automated postediting of documents', Ph.D. thesis, University of Southern California, Marina del Rey, CA.

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

Grishman, Ralph, Catherine Macleod & Adam Myers: 1998, *COMLEX Syntax Reference Manual*, Proteus Project, NYU, (`http://nlp.cs.nyu.edu/comlex/refman.ps`).

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

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

Huddleston, Rodney & Geoffrey K. Pullum: 2002, *The Cambridge Grammar of the English Language*, Cambridge, UK: Cambridge University Press.

Ikehara, Satoru, Satoshi Shirai, Akio Yokoo & Hiromi Nakaiwa: 1991, 'Toward an MT system without pre-editing – effects of new methods in **ALT-J/E**–', in *Proc. of the Third Machine Translation Summit (MT Summit III)*, Washington DC, USA, pp. 101–106.

Ngai, Grace & Radu Florian: 2001, 'Transformation-based learning in the fast lane', in *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, Pittsburgh, USA, pp. 40–7.

Pearce, Darren: 2001, 'Synonymy in collocation extraction', in *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, USA.

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

Ross, Háj: 1995, 'Defective noun phrases', in *Papers of the 31st Regional Meeting of the Chicago Linguistics Society*, pp. 398–440.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger: 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)*, Mexico City, Mexico, pp. 1–15.

Soehn, Jan-Philipp & Manfred Sailer: 2003, 'At first blush on tenterhooks. About selectional restrictions imposed by nonheads', in Gerhard Jäger, Paola Monachesi, Gerald Penn & Shuly Winter, eds., *Proceedings of Formal Grammar 2003*, pp. 149–161.

Stvan, Laurel Smith: 1998, 'The semantics and pragmatics of bare singular noun phrases', Ph.D. thesis, Northwestern University.