Abstract

Preposition-noun combinations (PNCs) are compositional and productive, but not fully regular. In school grammars and many theoretical approaches, PNCs are neglected, but they have recently been addressed in an HPSG analysis by Baldwin et al. (2006). After discussing some basic properties of PNCs, we show that statistical methods can be employed to prove that PNCs are indeed productive and compositional, which again implies that PNCs should receive a syntactic analysis. Such an analysis, however, is impeded by the limited regularity of the construction. We will point out why adding semantic conditions to syntactic schemata might be necessary but not sufficient and turn then to a framework which allows the derivation of syntactic (and semantic) generalizations from linguistic data without taking recourse to introspective judgments.[1]

1 Introduction

Combinations of a preposition with determinerless nominal projections have been neglected in theories of grammar for some time. But with increasingly blurring boundaries between core and periphery in grammar, a growing interest in preposition-noun combinations can be observed. Minimally, a preposition-noun combination consists of a preposition and an unadorned count noun in the singular, as illustrated in (1). Minimal combinations can be extended in various ways: the noun can be modified, as illustrated in (2); it may – and in some cases even must – realize a complement, as illustrated in (3).

(1) *auf Anfrage* (after being asked), *auf Aufforderung* (on request), *durch Beobachtung* (through observation), *in Anspielung* (alluding to), *mit Vorbehalt* (with reservations), *ohne Vorwarnung* (without warnings), *unter Androhung* (under threat)

(2) *auf parlamentarische Anfrage* (after being asked in parliament), *auf diskrete Aufforderung* (on discreet request), *durch kritische Beobachtung* (through critical observation), *in untertreibender Anspielung* (in an allusion to understate ...), *mit leisem Vorbehalt* (with quiet reservations), *ohne mündliche Vorwarnung* (without verbal warnings), *unter sanfter Androhung* (under gentle threat)

[1] I would like to thank Francis Bond, Takao Gunji and Shuichi Yatabe for kindly inviting me to HPSG 2008 in Japan, and thus making it possible to discuss the work reported here. The present results would not have been possible without the assistance of Katja Keßelmeier, Antje Müller, Claudia Roch, Tobias Stadtfeld, and Jan Strunk. Special thanks to Stefan Müller for his help and patience.

(3) Experten, die von Anreizen reden, sollten diese unter Annahme
experts who of incentives talk should these under assumption
realistischer Bedingungen durchrechnen.
realistic conditions calculate
'Experts who talk of incentives should calculate on the basis of realistic conditions.'

The characteristic difference between a preposition-noun combination on the one hand and an ordinary PP on the other hand is the missing determiner in the nominal projection. This property has led some linguists to call such constructions somewhat erroneously *determinerless PPs* (cf. Quirk et al. 1985). Since determiners combine with nominal projections, and not with prepositions, we will refrain from using this terminology and call the combination in (1) to (3) *preposition-noun combinations* (henceforth: PNCs). The missing determiner might also be one of the main reasons for neglecting the construction: it makes the construction look like an irregular sequence in languages that require the realization of a determiner together with a count noun in the singular. By the same line of reasoning, constructions like the ones presented in (4) and (5) do not form exceptions. The nouns in question are not classified as count nouns or not realized in the singular.[2]

(4) Sie befanden sich unter Druck.
They found themselves under pressure

(5) Die wechselnden Ursachen verbieten es, bei Annahmen über
the changing causes prohibit EXPL at assumptions about
künftige Bewegungen eine einfache Fortschreibung der
future movements a simple continuation the
Vergangenheit zugrunde zu legen.
past base to place.
'The ever-changing causes put a ban on a simple continuation of past activities as a basis to determine future movements.'

The German Duden grammar (Duden 2005) offers an exception-based treatment of PNCs. According to Duden rule 442 (Duden 2005:337), the realization of a determiner is mandatory for count nouns realizing the feature singular. In order to deal with constructions like (1), (2) and (3), the Duden introduces rule 395 (Duden 2005:306). It provides a list of exceptions to rule 442, thus suggesting that PNCs are restricted to sublanguages and registers and that they do not form a productive subclass of prepositional phrases. Such a treatment is not an oddity of the Duden grammar or of German. Himmelmann (1998) reports the universal tendency that singular count nouns have to be accompanied by determiners; but also that determinerless count nouns are often combined with prepositions. More recently, Baldwin et al.

[2] Bare plurals and mass terms form NPs without determiners. Hence the relevant phrases in (4) and (5) have to be analyzed as ordinary PPs.

(2006) have claimed that a subclass of English PNCs must be analyzed as productive.
As a second reason for neglecting PNCs, we may consider the observation that at least certain combinations of a preposition and a noun are idiomatic. An illustration is given in (6).

(6) Alles ist unter Kontrolle.
Everything is under control

Combinations like the one in (6) are often identified with PNCs as defined above although they do not strictly belong to this set. Typically, nouns found in constructions like (6) have to be analyzed as mass terms. This is obscured by the fact that the property of being a count noun cannot be attributed to words, but must be attributed to word senses. So while *Kontrolle* in one of its senses can be a count noun (as in *Eingangskontrolle,* i.e. *reception inspection)*, this is not the pertinent sense in (6).
A third reason for the neglect might stem from the observation that PNCs are known to be less regular than ordinary PPs. The frequency of PNCs when compared with prepositional occurrences in general is indicative. Table (7) lists the proportion of PNCs for 20 high frequency prepositions in a newspaper corpus of 310 million words (Neue Zürcher Zeitung, 1993-1999).

(7) Proportion of PNCs for 20 high-frequency Ps

Preposition	Frequency	P-N Proportion
in	2.127.029	0,76 %
mit	1.233.962	2,46 %
auf	1.094.267	1,45 %
für	940.824	2,02 %
an	547.787	1,93 %
nach	460.080	2,79 %
bei	383.172	2,32 %
über	379.538	1,93 %
um	268.384	2,22 %
vor	264.178	2,15 %
durch	249.353	4,27 %
unter	199.232	2,08 %
gegen	179.375	3,33 %
seit	120.517	1,26 %
ohne	93.219	11,56 %
wegen	66.973	5,25 %
während	45.170	0,38 %
neben	38.804	3,71 %
gemäß	36.878	4,82 %
dank	26.217	8,58 %

With the exception of *ohne (without), dank (thanks to)* and *wegen (because)* PNCs make up less than 5 % of the respective occurrences of prepositional phrases, and in many cases, the proportion falls below a value of 3 %.
What is more, speakers show great reluctance and cannot easily decide whether a PNC should be considered acceptable. Baldwin et al. (2006) point out that combinations might be constrained by further semantic conditions, but it seems that the pertinent conditions are not available to speakers in judgement and production tasks. Since speakers are not able to judge newly coined PNCs, taking recourse to introspective judgments or judgment tasks cannot substantiate the productivity of the construction.
The following sections will address these issues in turn. In the second section, we will report results from Kiss (2007) and Dömges et al. (2007) showing that PNCs can neither be classified as non-compositional, nor as non-productive. From an empirical perspective, PNCs in German are no more idiomatic than other regular phrasal combinations, and from the same perspective, they can be classified as productive, supporting the claim made in Baldwin et al. (2006) for English. In the third section, we will review the proposal made in Baldwin et al. (2006) that PNCs are in fact completely regular but the rules have to be amended by semantic conditions. In the final section, we will suggest that in the absence of clear judgments, annotation mining (Chiarcos et al. 2008) will be useful to arrive at results concerning the latent properties, which determine the combination of prepositions and determinerless count nouns in the singular.

2 Compositionality and Productivity

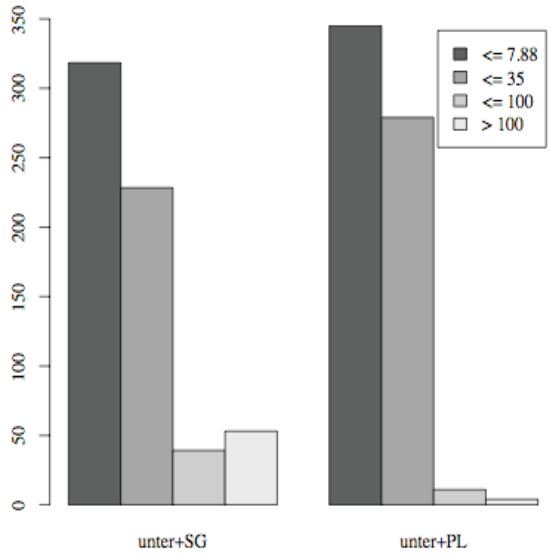
2.1 Compositionality

In a corpus-based study, Kiss (2007) has investigated whether PNCs of the type *unter+noun* should be classified as mainly compositional or not.
To assess the compositionality of PNCs, Kiss (2007) makes use of a structural analogy between PNCs and ordinary collocations. Methods to detect collocations can be used to determine whether PNCs behave like collocations.[3] A high degree of non-compositional combinations among PNCs would entail a high degree of fixed expressions and hence a high degree of collocations, which would be found my statistical methods for the identification of collocations. Kiss (2007) employs Dunning's *log likelihood ratio* (Dunning 1993) and compares the distributions of log likelihood ratios for combinations of *unter* with a noun in the singular and PPs headed by *unter* where the NP-complement is a bare plural. Since combinations with bare plurals are phrases that do not show a particularly high degree of idiomatic combinations, deviations between this class and combinations of

[3] For a discussion of the relation between collocations and idioms, cf. Burger (2007), Deuter (2005), and Smadja (1993).

prepositions and a singular noun would allow the conclusion that the latter class does indeed show a higher degree of idiomatic members.

(8) Collocation Detection for *unter+noun*$_{sg}$ *vs.* [$_{PP}$ *unter noun*$_{pl}$]



Following the analysis suggested in Dunning (1993), we may assume a basic threshold value of 7.88, which means that structural dependency between two adjancent words makes their occurrence in the corpus $e^{7.88/2}$ times more likely than assuming that the words are structurally independent. However, as has already been pointed out by Dunning (1993), the absolute values are of much lesser relevance than either an ordering reached among the candidate pairs or a comparison of values between one set of candidates and another set, whose properties are known. In addition, the basic value of 7.88 does not take into account the influence of morphosyntax and grammar, so that a more plausible threshold could be placed at a level of 35.

Given these assumptions, the figures summarized in (8) are even more indicative: 40 % of candidate pairs of type *unter+noun*$_{sg}$ show a log likelihood value *below* the basic threshold of 7.88. 75 % show a value below the more plausible threshold. What is more, the distribution between the singular and the plural types shows a similarity in the first two columns, mostly deviating if values larger than 35 are considered. This deviation indicates that there is a larger number of collocations among combinations of type *unter+noun*$_{sg}$ than among combinations of the plural type. But the total number of presumed collocations is small in both classes. The results show that most instances of *unter+noun*$_{sg}$ cannot plausibly be analyzed as non-compositional combinations. While there are more candidates with high log likelihood values among *unter+noun*$_{sg}$, their number is still small and does not justify the claim that the combination is idiomatic in general.

2.2 Productivity

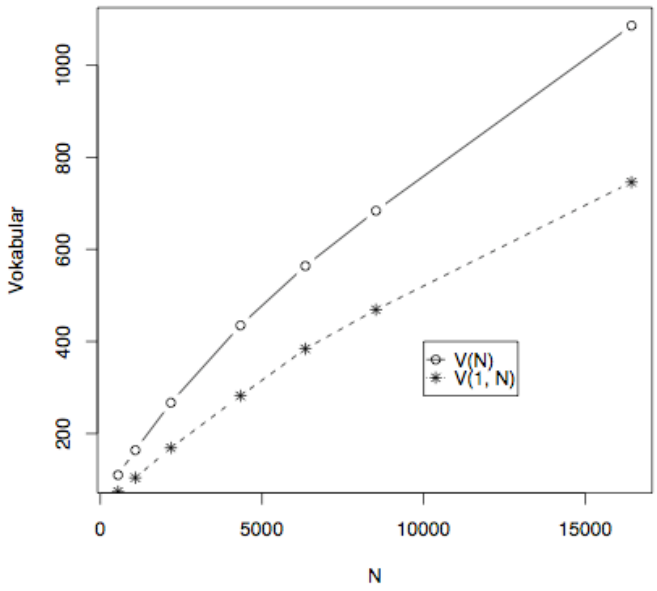
The empirical productivity of PNCs has been investigated in Dömges et al. (2007), following and extending the calculations for morphological productivity in Baayen (2001) and Evert (2004). Baayen (2001) has proposed that a process can be considered productive if the number of *hapax legomena* produced by the process will not drop below a threshold, as the corpus gets larger. The basic insight is that a process is still productive if more and more new instances are coined. If an instance is truly new, it will be encountered only once (it is already known when encountered a second time), making it a *hapax legomenon*. If a process cannot produce new instances, there will be no further *hapax legomena*. True productivity is thus indicated by three measures: to be productive, the vocabulary V(N) must not decrease as the corpus size N grows, i.e. $V(N) \leq V(M)$ if $N < M$; the number of hapax legomena V(1, N) must not decrease as the corpus increases, i.e. $V(1, N) \leq V(1, M)$ if $N < M$; and finally, the productivity as measured on the basis of the *hapax legomena* and the corpus size must not fall below a threshold. The measure for productivity is calculated as illustrated in (9).

(9) Baayen's (2001) measure for productivity: $\mathbf{P(N)} = \mathbf{E}[V(1, N)]/N$

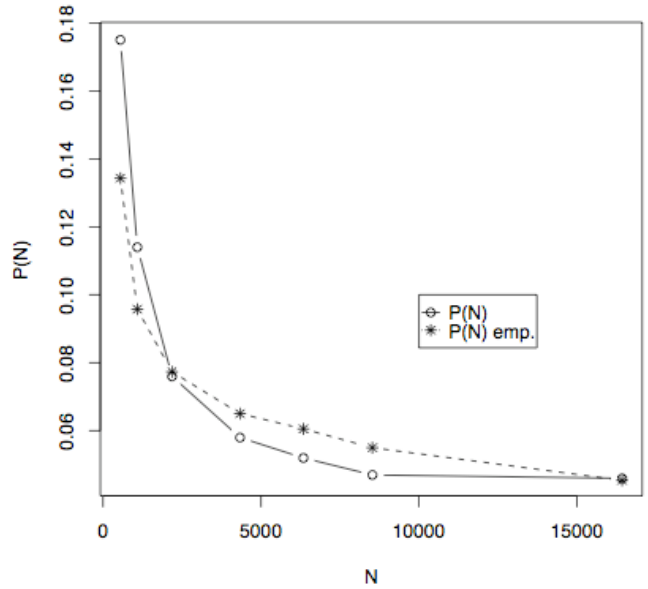
The measure in (9) also has a probabilistic interpretation: it provides the likelihood that a new instance can be observed after a corpus of N token instances has already been considered. With regard to the values for V(N) and V(1, N), the following illustration shows that both values increase as the corpus of candidates gets larger, already suggesting that the process is productive.

Yet, **P(N)** has to be determined. In its calculation we require the true expectation of the *hapax legomena* $\mathbf{E}[V(1, N)]$, which is not known for the sample corpus. Dömges et al. (2007) suggest the following approximation: They calculated the empirical productivity of the pertinent construction, i.e. $\mathbf{P(N)} = V(1, N)/N$ for fixed values of N. Dömges et al. (2007) employ two different regression models and use the empirical productivity over a large sample to determine which of the two models offers a better fit. The two models are a finite Zipf-Mandelbrot model (fZM) and a generalized Zipf-Mandelbrot (ZM) model (for a detailed discussion of the models, cf. Evert 2004). The crucial difference between the finite and the generalized model concerns the cardinality of categories employed by the two models.

(10) Development of V(N) and V(1, N)



(11) Fitting to empirical **P(N)** from Dömges et al. (2007)



While an fZM assumes a finite number of categories, a ZM allows for infinitely many categories. If a better fit is reached by an fZM, this would indicate an upper limit of different instances of the basic process. But if a

better fit is reached by the ZM, infinitely many instances of the basic process are predicted, which yields true productivity. Dömges et al. (2007) show that ZM provides the better approximation, as is illustrated in (11). According to this result, PNCs are not properly analyzed by a finite set of instances, i.e. the combination is productive.
Summing up, the investigations in Kiss (2007) and Dömges et al. (2007) have shown that PNCs in German are compositional and productive, thus supporting the proposal by Baldwin et al. (2006) that subclasses of English PNCs have to be analyzed as productive.

3 Semantic Conditions and Syntactic Combinations

Baldwin et al. (2006:175f.) presuppose the results stated in section 2. They conclude that an exception-based proposal *"will not extend to the productive constructions ... in which a particular preposition ... selects for an exclusively countable noun that cannot project a determinerless NP in other syntactic contexts."* They assume that at least certain prepositions can be described by a lexical entry as the one given in (12).

(12) Lexical entry of P (Baldwin et al. 2006)

$$\left[\text{SYN|CAT}\left[\begin{array}{l}\text{HEAD prep}\\ \text{VAL|COMPS}\ \left\langle\left[\text{SPR}\ \left\langle\text{Det}\right\rangle\right]\right\rangle\end{array}\right]\right]$$

But a lexical entry like the one given in (12) would justify the conclusion that PNCs are fully regular. Thus, it leaves open why speakers cannot form clear judgments and are uneasy to coin new combinations. Baldwin et al. (2006:176) note that *"[t]hese productive [determinerless PPs] 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."*
This amendment does not seem to be sufficient, both from a conceptual and an empirical perspective. Conceptually, adding semantic conditions to a rule, schema, or general lexical entry may affect the generality of a rule; it does not affect its regularity (and a lexical entry is already quite specific. Constraining it further does not actually change its status in the architecture of the grammar). If rule conditions are met, a rule can and has to be applied.
An HPSG of PNCs should not only offer a grammatical description but should also account for speakers' judgments of the pertinent construction. Speakers cannot easily discern acceptable from unacceptable PNCs, and this does not seem to be a question of generality, but of regularity.

We will leave this issue open and turn to the empirical perspectives of the proposal presented in Baldwin et al. (2006), reminding us of the polysemy of prepositions. In addition to the two alternatives suggested by Baldwin et al. (2006), a third possibility is conceivable: it might be that the noun imposes constraints on the interpretation of the preposition. Such a treatment would in fact require changes of the rule schema responsible for complementation, and imply further ramifications for the principles of semantic combination. But we will ignore these issues presently, especially since the second amendment suggested by Baldwin et al. (2006) would require similar changes.
An illustration of the application of the third alternative can be given by considering interpretation options of a preposition if either realized with an NP complement or with a determinerless nominal complement (DNC). With regard to the possible interpretation of the preposition *unter,* the dictionary *Duden Deutsch als Fremdsprache (Duden German for Foreign Learners;* Duden 2002) offers eleven top level definitions, many of which show fine-grained subdivisions and further qualifications. The top-level definitions are listed in (13).

(13) *spatial, temporal, circumstantial, comtemporaneity, subordination, association, presence among other things, picking an individual from a set, mutual dependency, state, causality*

In a further corpus study, we have investigated interpretation options of *unter* in combination with NPs and determinerless nouns. The corpus contains 29 million words and 650 different types of *unter* combined with an unadorned noun. It turns out that in relation to combinations of *unter+NP,* spatial and temporal interpretations are underrepresented in combinations of the type *unter+noun.* PNCs that require a spatial interpretation are highly restricted and can only be found in headlines – which generally seem to offer a natural habitat for otherwise problematic PNCs. An illustration is given in (14).

(14) Fußweg unter Brücke gesperrt.
footpath under bridge barred-for-traffic
'The footpath under the bridge is barred for traffic.'

This small study illustrates that certain interpretations of a highly polysemous preposition seem to be shadowed if the preposition is used in a PNCs. The results, however, are not accidental. Müller (2008:330) reports that uses of the preposition *unter* in support verb constructions involve a suppression of the spatial interpretation of the preposition, thus mirroring the present results. An analysis of PNCs should thus not only constrain the semantics of the preposition's complement but also account for a suppression of one of the preposition's senses when used in a PNC.
The cross-linguistic perspective offers a further empirical challenge. If the occurrence of PNCs is largely restricted by semantic conditions, we would expect that PNCs occurring in one language are mirrored in other, closely

related languages. But this does not seem to be the case, as can be illustrated with the examples in (15).[4]

(15) a. Mijn auto is proper. Ik smijt alles op straat.
b. Mein Auto ist sauber. Ich schmeiße alles auf *(die) Straße.
c. My car is clean. I throw everything on *(the) street.

While (15a) shows that the determiner can be dropped in the combination *on straat,* leaving out the determiner in similar constructions is neither possible in German, nor in English (15b, c). If semantic conditions govern the omission in Dutch and Flemish, why does the same condition not apply to German or English? Interestingly, a Dutch grammar offers an explanation for the grammaticality which is in direct opposition to the analysis suggested for PNCs in the Duden, in that the grammar turns PNCs into regular citizens, once a semantic condition is fulfilled: *"We gebruiken ook geen lidwoord als het zelfstandig naamwoord een meer algemene betekenis heeft." (We do not use a determiner if the noun receives a generic interpretation.* Grammatica in gebruik: Nederlands for anderstaligen, p. 42).
It should be noted, however, that the implicational relationship between a generic interpretation of the noun and a determiner omission cannot always be established. A generic interpretation is not sufficient to drop the determiner in German and English, as has been illustrated in (15b, c). Moreover, many examples with non-generic interpretations of the noun can be found, illustrated with *auf Initiative (on initiative)* and *unter Voraussetzung (presuming that)* for German in (16) and (17)

(16) Im Januar 1996 hat sich dort auf (die) Initiative der ehemaligen
in January 1996 has REFL there at (the) initiative the former
Bob-Vizeweltmeisterin Erica Fischbach eine Bob- und
bobsled-vice-world-champion EF a bob and
Rodelabteilung formiert.
toboggan-department constituted
'On initiative of former vice-world champion Erica Fischbach, a new department for bobsled and toboggan has been constituted there in January 1996.'

(17) Auch Philipp Egli besteht auf einer eigenen Handschrift – unter
also Philipp Egli insists on a own signature under
Voraussetzung des Einverständnisses des Ensembles.
prerequisite the acceptance the ensemble
'Philipp Egli insists on his own style as well, provided that the ensemble accepts.'

[4] Example (15a) is used as an ironic slogan against waste prevention on Belgian highways.

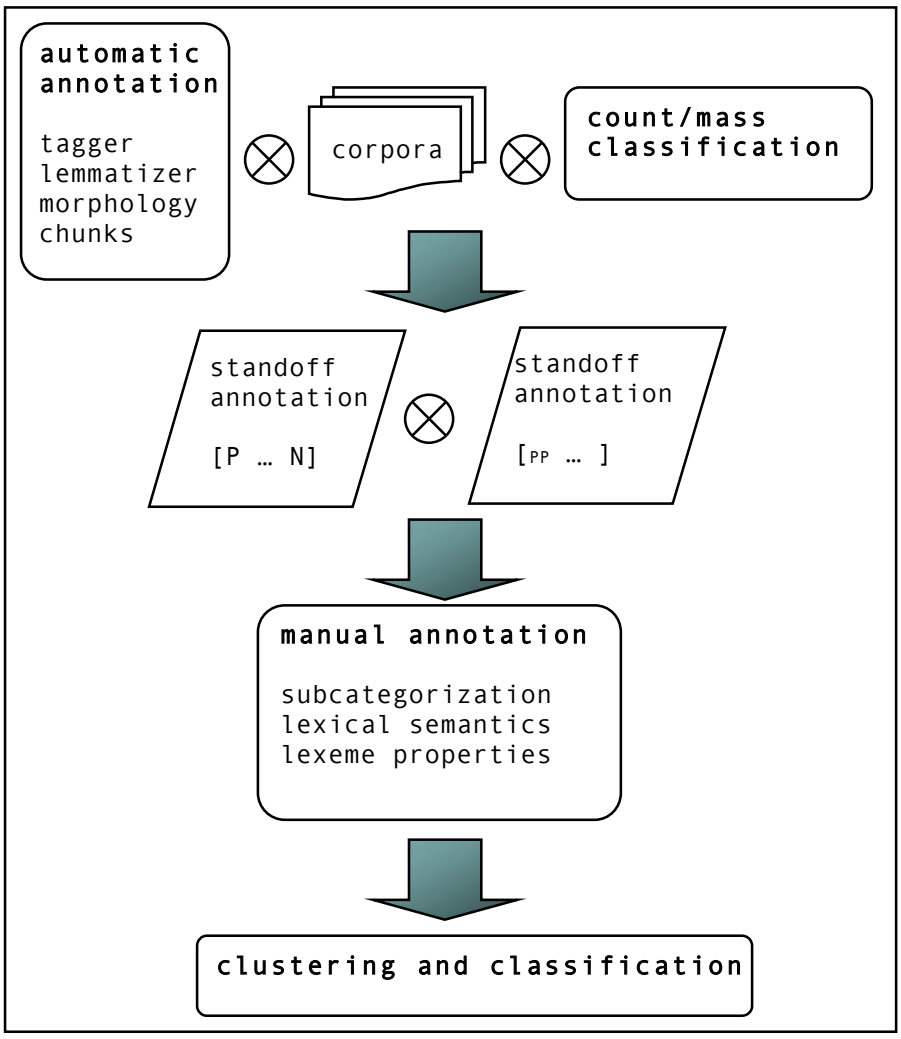
It is interesting that the use of an article is in fact optional in example (16), while its omission leads to strong unacceptability in (15b). The example (17) further illustrates with the preposition *unter* that PNCs cannot be tied to genericity in German.
It is indicative that for both the prepositions *auf* and *unter* a spatial interpretation is blocked if the prepositions are used in PNCs. A similar condition may apply in English but it obviously not active in Dutch and Flemish. Possibly, the semantic conditions active in the determination of acceptable PNCs must be described as language-specific.

4 Where is the method in this madness?

While the regularity of PNCs was neglected for a long time (and sometimes is still today), current analyses assume that the construction should in fact be described as regular, and that PNCs are compositional. Support for both assumptions come from corpus-based studies as presented in section 2. Despite a growing consensus that the constructions are regular, it is accepted that the constructions are just not as regular as other combinations – such as an ordinary preposition and an NP. Yet, grammar theory has not been able to pin down the factors that distinguish grammatical from not so grammatical combinations of a preposition and a determinerless nominal projection. Standard methods for the determination of grammaticality and the identification of features and factors, which make a construction acceptable cannot be applied to PNCs. In particular, speakers are extremely uneasy to produce acceptability judgments in isolation and normally cannot coin new combinations. A variety of factors may account for this lack on the speaker's side. To begin with, prepositions are highly polysemous, and only certain senses seem to be available in PNCs. Choosing a sense, however, largely depends on the local and non-local context in which a PP or preposition-noun combination can be embedded. Secondly, the distinction between mass nouns and count nouns interferes. Only combinations with the latter should lead to imperfect combinations, but this conclusion already assumes that speaker's have knowledge of the count/mass-distinction that is independent of contextual clues (cf. the recent discussion in Borer 2004, where this assumption is explicitly denied). Additional factors may depend on different senses of the nouns involved. Taken together, it does not come as a surprise that speakers become reluctant. For the linguist, the question remains how to tackle these constructions and how to identify the discerning factors.
A solution to this problem comes from the area of *annotation mining* (Chiarcos et al. 2008). Annotation mining combines large corpora with classification tools and annotations to produce large annotated corpora, ideally in a stand-off format allowing further extension of the annotation without affecting the other layers. After automatically and manually annotating the corpora, they can be used as input for clustering and categorization tools, such as Weka (Witten and Frank 2005). Since raw data have been annotated on various strata from morphology to semantics, and

since many instances have been annotated, classifier and clustering tools receive a robust multidimensional representation of the data. In the present setting, raw corpora are combined with lemmatizers, morphological analyzers, taggers and chunkers, and in particular, with a classification system to determine the count/mass-distinction, an annotation of realized syntactic arguments, as well as annotations on the sense level for nouns and prepositions. From the initial corpora, we extract all cases of PNCs (appropriately chunked), all cases of ordinary PPs, in which the same preposition and noun appears, and also all NPs outside of PPs, in which the noun appears. By extracting not only PNCs, but also PPs, and NPs, we hope to find characteristic properties that are present with the former but are possibly missing with the latter. The identification of characteristic yet latent traits is not carried out by manual inspection, but by feeding the different subgroups into a classification algorithm and extract the rules for classification from the classifier – particularly well-suited to this task are decision tree classifiers, such as Weka's J4.8, which is a re-implementation of the standard decision tree algorithm C4.5 (cf, Quinlan 1993). Decision tree classifiers, possibly amended with a Principal Components Analysis (Baayen 2001), are useful in that they allow the derivation of a probabilistic rule system from the classification. The following schema (18) gives an illustration of the annotation task.

(18) Annotation Mining



In an on-going project, we are working on the identification characteristic properties of PNCs in the aforementioned manner. The results will form the basis for further analysis in terms of controlled experiments. The result of this process will most likely be a probabilistic analysis of PNCs. Yet the results can be turned into a categorical analysis by using a threshold value to turn continuous probabilities into clear-cut categories, thus offering a broader empirical coverage of PNCs in terms of a refined HPSG analysis. As not only syntactic properties, but also semantic and other influences play a role in determining whether or not a preposition may combine with a determinerless nominal projection, a model like HPSG is clearly more appropriate for a representation of the latent generalizations than a framework that relies on purely syntactic means only.

References

Baayen, Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer.

Baldwin, Timothy, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger and Ivan A. Sag. 2006. In Search of a Systematic Treatment of Determinerless PPs. In Patrick Saint-Dizier (Ed.): *Syntax and Semantics of Prepositions*. Dordrecht: Springer, pages 163-179.

Borer, Hagit. 2004. *In Name Only. Structuring Sense, Vol. I.* Oxford: Oxford University Press.

Burger, Harald. 2007. *Phraseologie. Eine Einführung am Beispiel des Deutschen.* Berlin: Erich Schmidt Verlag.

Chiarcos, Christian, Stefanie Dipper, Michael Götze, Ulf Leser, Anke Lüdeling, Julia Ritz and Manfred Stede. 2008. A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. *Traitement Automatique des Langues*. *Special Issue Platforms for Natural Language Processing*. *ATALA* 49 (2).

Deuter, Margaret (Ed.). 2005. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Dömges, Florian, Tibor Kiss, Antje Müller and Claudia Roch. 2007. *Measuring the Productivity of Determinerless PPs*. In: Fintan Costello, John Kelleher and Martin Volk (Eds.): Proceedings of the ACL 2007 Workshop on Prepositions, Prag, pages 31-37.

Duden. 2002. *Duden. Deutsch als Fremdsprache*. Mannheim: Bibliographisches Institut & F.A. Brockhaus AG.

Duden. 2005. *Duden. Die Grammatik*. Duden Band 4. Mannheim: Bibliographisches Institut & F.A. Brockhaus AG.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19, 61-74.

Evert, Stefan. 2004. A Simple LNRE Model for Random Character Sequences. In *Proceedings of the 7mes Journees Internationales d'Analyse Statistique des Donnees Textuelles,* pages 411-422.

Himmelmann, Nikolaus. 1998. Regularity in Irregularity: Article Use in Adpositional Phrases. *Linguistic Typology* 2, 315-353.

Kamber, Alain. 2008. *Funktionsverbgefüge – empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen.* Tübingen: Max Niemeyer Verlag.

Kiss, Tibor. 2007. Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. *Zeitschrift für Sprachwissenschaft* 26, 317-345.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning.* San Francisco: Morgan Kaufman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19, 143-177

Witten, Ian H. and Eibe Frank. 2005. *Data Mining. Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufman.