Do we need a grammar of irregular sequences?

Tibor KISS
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
tibor@linguistics.rub.de

Abstract:
Combinations of a preposition and a singular count noun (P+Noun) occur with considerable frequency in German everyday speech. Yet it is unclear whether syntactic means, morphological operations, or other processes should handle the combination. If P+Noun combinations were analyzed as PPs, the grammar rule that handled singular count nouns and determiners would have to be amended. But a morphological analysis would require the formulation of a new word formation process, since P+Noun-formations are not known otherwise. It has thus been suggested that the combinations do not come about as result of an application of a rule of grammar, and perhaps that P+Noun combinations simply have to be listed. This assumption has been corroborated by the observation that the combination is not free in the sense that any P may combine with any Noun, and that the construction is not transparent to competent speakers in the sense that they cannot produce new instances on the fly. This observation leads to the conjecture that the set of P+Noun may be large, yet finite, and hence should be treated as listemes instead of being subject to rule application. The present paper argues on the basis of statistical evidence that the set of P+Noun combinations is infinite, and hence that P+Noun must be handled by a rule of grammar, be it syntactic or morphological.

1 Introduction

The combination of a preposition with a singular count noun, as illustrated in (1) with the preposition unter, is a frequent construction in written and spoken German. As with many other combinations, idiomatic interpretations can be found, such as unter Kontrolle (under control), but we also find many compositional instances, like the ones given in (1).

(1) unter Vorbehalt (subject to), unter Hinweis (with reference to), unter Verweis (with reference to), unter Beteiligung (with the participation of), unter Abschwächung (weakening)

At first sight, one may assume that the expressions in (1) can be categorized as PPs. Yet, this conclusion is problematic: PPs consist of prepositions combining with NPs. Singular count nouns like the ones given in (1) may not form an NP in itself, as we learn from the Duden-Grammatik:

Substantive mit Merkmalkombination ’zählbar’ plus Singular haben ... grundsätzlich immer ein Artikelwort bei sich, und wenn es als letzte Möglichkeit der indefiniten Artikel ist. (Duden 442)
(Hence, count nouns marked singular are always combined with a determiner, and it has to be an indefinite determiner if other determiners are blocked.)

Quite obviously, the constructions in (1) violate rule 442. In light of this observation, three positions can be established. Chafe (1968) brought attention to English word combinations that cannot be syntactic, because they are ungrammatical yet possible, such as by and large. Nunberg et al. (1994) suggest treating such combinations by listing them. Multiword lexemes as entities that have to be listed are explicitly addressed in Sag et al. (2002). While I am not aware of an explicit computational treatment of P+Noun combinations, their deviance could be handled in a framework that assumes the existence of so-called multiword lexemes, which presumably can be listed. It is more important that not much is gained by calling a combination a multiword lexeme, if one has to admit that the combination is regular in the sense that we find an infinite set of instances.

The contrary position would be to assume that P+Noun is admitted by the grammar of German, and hence that P+Noun is an instance of a rule of German grammar (one which counter rule 442 does not require the combination with a determiner before other combinations become possible). This position is at odds with speakers’ intuitions about P+Noun combinations (which seem to speak in favour of listing them). For compositional combinations, speakers should judge syntactic combinations as possible as long as P and Noun are semanti-
cally compatible. For idiomatic combinations, speakers should either be aware of the idiomaticity or else be able to assign at least a compositional reading to the combination. Irrespective of the question of compositionality and idiomaticity, speakers should be able to judge the combination as grammatical or not. But this does not seem to be the case. Speakers of German are not instantly able to coin new combinations of \( P+Noun \), and also find it highly problematic to judge the grammaticality of \( P+Noun \) combinations, if these are offered without a context. Although this problem is not discussed in the literature,\(^1\) it seems to me that it is the reason why \( P+Noun \) combinations are not taken into consideration in the formulation of rules like 442 above.

Helbig and Buscha (1998) take an intermediate position by assuming that \( P+Noun \) is regular, yet not phrasal combination. They claim that \( P+Noun \) combinations are regular. Still, they are reluctant to classify the combination as syntactic, instead calling them “\( \text{Zusammensetzungen und Wortgruppen aus Präposition}+\text{Substantiv (zumeist mit Nullartikel)} \)” (combinations and word groups made up from \( P+Noun \) (very often with zero-determiner)).

One may ask why reference to a concept of a word group is necessary, if one assumes a zero-determiner is present. Given the existence of a zero-determiner, the combination as a whole can be described as a proper PP, and hence as a phrase.\(^2\) Helbig and Buscha (1998:403f.) are aware of this inconsistency and offer the following list of criteria to identify word group combinations of \( P+Noun \):

(2) a. The N occurring in \( P+Noun \) can only appear as part of the combination.

b. The meaning of the noun is weakened in the combination.

c. The noun cannot be used with a determiner in the combination.

d. The combination of \( P+Noun \) is word like in that its complement can be substituted by many other complements.

e. The combination can be replaced by a preposition.

f. The noun is not capitalized in the combination or is written together with the preposition.

There are two major problems with the criteria given. Some of the criteria are insufficient. This holds for (2a-d). While there are several combinations of \( P+Noun \) where the noun acts like a cranberry word, there are much more combinations for which this is not true. Hence criterion (2a) will only help for the rather small group of cases where the noun may uniquely be combined with \( P \). Similar considerations hamper the second criterion (2b). It suggests that nouns turn into ‘light nouns’ if combined with \( P \), but this does not hold for many instances of the construction that are clearly compositional, where the meaning of the noun is carried over without any change. Again, the third criterion may hold for a minority of nouns, but in general, combinations of type \( P+Noun \) alternate freely with combinations of type \( P+\text{Det+Noun} \).\(^3\)

The second problem concerns the validity of the criteria itself. Not only that (2f) is insufficient in that many of the combinations show a regular – syntactic – orthography, but relying

\(^1\) I am grateful to Joachim Jacobs, who drew my attention to this observation.

\(^2\) One obvious problem of assuming the existence of a zero determiner is that a reason has to be offered why zero-determiners may not replace indefinite determiners in ordinary NPs.

\(^3\) With the forth criterion, Helbig and Buscha seem to suggest that \( P+Noun \) collectively select a complement. According to standard assumptions, nouns only optionally govern a postnominal genitive complement. If the complement of \( P+Noun \) were the complement of \( Noun \), we would expect that the complement of \( P+Noun \) could be dropped under the very same conditions as the complement of \( Noun \). At first sight, \( P+Noun \) complements seem to violate this assumption since they appear obligatorily. A closer scrutiny of large corpora, however, reveals that the initial assumption, i.e. the idea that complements of nouns are always optional, must be doubted in itself. As long as the exact conditions of noun complementation in German are not clear, claims to the apparent optionality of noun complementation are futile.
on orthography is problematic in itself. With respect to (2e), we could use the same line of reasoning to suggest that every NP is in fact a word, since any NP can be replaced by a pronoun.

Eisenberg (1999:224) suggests that certain P+Noun combinations (those which are written together) do not belong to the syntactic or morphological grammar of German. They come about as the result of a so-called Unverbierung, a kind of customary combination of two elements that happen to be juxtaposed quite often. While I would agree that certain monolithic P+Noun combinations could be conceived as the result of a customary combination of juxtapositions, I think that the idea begs the question when applied generally to all P+Noun combinations. It does not explain why P and the noun are juxtaposed in the first place.

If P+Noun combinations are not the result of a syntactic rule, all combinations of P+Noun perhaps make up a finite, possibly large set, which would imply that P+Noun combinations are listemes. When this conclusion is accepted, the deviance of the construction receives an explanation as not being a construction at all, and combinations of P+Noun can be handled by listing their properties. While this move may be appropriate for cases that make up a small set of instances that are non-regular both semantically and syntactically, it is doubtful that the same conclusion should be reached for P+Noun combinations in general.

The focus of the present paper resides on showing that combinations of P+Noun cannot be handled by listing them. If one accepts that the set of P+Noun combinations is infinite, one also has to accept that combinations of P+N can only plausibly be handled by rule, be this rule syntactic or morphological. Abstracting away from morphological and syntactic rules, I will claim that P+N combinations are instances of a rule of grammar. The following definition may illustrate this term:

(3) Instance of a rule of grammar:
A combination [A B] can be considered to an instance of a rule of grammar Y if [A B] can be derived from Y.

I am not using the term regular, because this term seems to imply that the pertinent construction does not show irregular characteristics. This is not the case for P+Noun combinations, which – even as instance of a rule of grammar – show irregular characteristics.

The remainder of this paper is structured as follows: in section 2 I will introduce statistical measures to determine whether a given combination is finite or not. I will apply Baayen’s (2001) apparatus for measuring morphological productivity to the domain of syntax. The application to syntax is feasible in the present case since the pertinent construction roughly behaves like a bigram. Hence an extension from unigram measures (for productivity) to bigram measures (for determining whether the construction is an instance of a rule of grammar) can be carried out if a large database is considered. The experiments carried out here are based on a corpus of the size of approx. 106 million words. Section 3 will present and discuss the results of the corpus analysis. Section 4 will briefly discuss the implications for theoretical and computational linguistics.

2 A statistical treatment of syntactic regularity

Baayen (2001) suggests measuring productivity of morphological rules according to three simple criteria: Given a sample of word size N, V(N), the cardinality of the vocabulary gives the number of different types in the vocabulary. V(1, N) gives the number of hapax types in the vocabulary. Finally, P(N), defined as

\[
E \left[ \frac{V(1,N)}{N} \right]
\]

where E[V(1, N)] stands for the expectation of V(1, N) – offers the probability that a new type will show up after N tokens have been sampled. The basic idea is that a productive process shows a steady increase in V(N) and V(1, N), and, as a consequence of the latter, the probability of a new type occurring will be sufficient even if N gets very large. This can best be illustrated by using the opposite as an example.

\[4\] As every literate speaker of German is aware, orthographic rules are conventional. As a consequence of the recent reformulation of German orthography, ‘noun-like’ elements are now almost always capitalized, and as a consequence, the first half of Helbig and Buscha’s criterion cannot be applied any longer. For the same reasons, orthographic criteria are not taken seriously in recent discussions of the definition of word (cf. Bauer 2000).
Baayen (2001) makes use of a six-sided die, and I will follow his explanation. What is \( E[V(N)] \) for sufficiently large \( N \) given a six-sided die? It will be pretty close to 6 for \( N \) and it will remain there if we extend \( N \). Assuming a sample of 1000, we can be sure that the vocabulary size \( E[V(1000)] \) asymptotically approximates 6. Similarly, \( E[V(1, 1000)] \) will approximate 0, because we can expect that after throwing a six-sided die a thousand times, every number will have shown up at least two times. In the initial stages of the sample, say at \( N = 6 \), \( V(6) \) will be smaller than 6 and \( V(1, 6) \) will be higher than 0. Figure (1) shows the expected curves if a small and finite set is assumed, as \( V(1, N) \) approximates 0, \( P(N) \) approximates 0 as well, since the likelihood that a new type will be encountered after \( N \) samples is obviously 0 if the number of types is finite.

While Baayen (2001) has applied these measures to morphological productivity, I think that they can also be used to determine whether a given construction is syntactically regular in the sense that we have to expect an increase of the total number of types, as well as of the number of hapax legonema if \( N \) is increased. In such a situation, the vocabulary size \( V(N) \) will not approximate an asymptote as \( N \) gets larger, but will steadily increase. The same holds for \( V(1, N) \), and as a consequence, \( P(N) \) will be low but different from 0.

If \( P+Noun \) combinations are an instances of a rule of grammar, we predict a steady increase of \( V(N) \), as well as of \( V(1, N) \) for the simple reason that the class of nouns is infinite. We use \( R(N) \) instead of \( P(N) \) to determine the relationship between \( E[V(1, N)] \) and \( N \). For the present purposes, we set \( E[V(1, N)] = V(1, N) \). Here, \( R(N) \) stands for the likelihood of seeing a new instance of an application of the grammar rule which combines \( P+Noun \) after \( N \) samples. \( R(N) \) is calculated in the same way as \( P(N) \), and the change of the letter only indicates that we are talking no longer about morphological productivity but of syntactic rule application.

3 A empirical study

I have investigated the distribution of the \( P+Noun \) combination \( unter+Noun \) in four consecutive editions of the Neue Zürcher Zeitung, from 1995 to 1998, totally comprising 106 million words. The different types of \( unter+Noun \) have been classified according to their log likelihood (Dunning 1993), using Perl and Text::NSP,\(^5\) and have been extracted automatically from the larger set of all bigrams where \( unter \) is \( w_j \). Using manual filters, occurrences of genitive marked nouns, proper names, as well as titles, which are typically used as appositions after P and before proper nouns, have been extracted. In addition all occurrences of nominalized verbs after \( unter \) have been ignored. Nominalized verbs (VN) after \( unter \) usually denote a circumstance of the verb, which has been modified by \( unter+VN \). This construction has been excluded since nominalized verbs do not require a determiner according to rule 442 and hence \( unter+VN \) forms a proper PP.

Finally all occurrences of \( unter+Noun \) have been split according to the number marking of the noun. Just as combinations of \( unter+VN \), instances of \( P+Noun_{pl} \) are not considered an irregularity since plural nouns may form an NP without addition of a determiner.

Figure (2) shows the development of \( V(N) \) and \( V(1, N) \) from an initial subset of the corpus (approx. 6 million words) to the full corpus (106 million words). As the corpus gets larger, we observe an increase of both \( V(1) \) and \( V(1, N) \). As Figure (2) shows, neither \( V(N) \) nor \( V(1, N) \) seem to approximate an asymptotic value, and the steady increase of \( V(1, N) \) allows the conclusion that the construction is not finite in its extension. This conclusion is further corroborated by the values for \( R(N) \) in Figure (3).

Figure (3) show the values for \( R(N) \) as a function of empirical values for \( V(1, N) \). Recall that non-productive morphological processes show a \( P(N) \) curve which asymptotically approximates 0. While the probability to see a new instance of \( unter+Noun_{sg} \) is quite small according to Figure (3), it significantly differs from 0. What is more, the empirical value for \( R(N) \) exceeds the expected value even if the full corpus is considered, though it looks as if the two value would converge of further data are considered. In order to evaluate the probability \( R(N) \) itself, I have compared the values for \( unter+Noun_{sg} \) with the values for all bi-

\(^5\) http://search.cpan.org/src/TPEDERSE/Text-NSP-1.01
grams *unter+Noun*$_{pl}$. This is illustrated in Figure (4).

Figure (4) shows that R(N) for singular and plural complements of *unter* is quite similar. The values for plural complements are slightly higher than the values for singular complements, but the curves itself bear a close resemblance. It should be kept in mind that combinations of *unter+Noun*$_{pl}$ are undisputed instances of a rule of grammar, viz. the ordinary rule combining P+NP to PP.

In sum, the values for V(N) and V(1, N) suggest that the combination of *unter+Noun*$_{sg}$ should not be analyzed by listing them. The R(N) value suggests that the likelihood of new types of such combinations is high enough to assume that we are not dealing with a finite vocabulary. This assumption is further corroborated by comparing R(N) for *unter+Noun*$_{sg}$ with R(N) for *unter+Noun*$_{pl}$. The latter combination is undoubtedly non-finite, and given the close resemblance between the curves for singular and plural complements, the combinations of *unter+Noun*$_{sg}$ should be analyzed as one of the infinite set of instances of a rule of grammar.

4 Implications

The results of section 3 have far reaching consequences for both theoretical and computational linguistics. The empirical broadening of research in computational linguistics has led to the inclusion of a wide variety of empirical phenomena. Quite often, one finds references to constructions which are more than a word, but are not treated as phrases, as can e.g. be witnessed by the use of terms like *multiword lexeme* or *word group* in recent work. While we cannot say much about the notoriously problematic distinction between words and phrases in computational linguistics, the results of the present study suggest that before combinations are classified as fixed or listed, it should be investigated whether they are finite or not, even if the rule which is responsible for the construction cannot readily be formulated. With respect to the combination of P+Noun$_{sg}$, the present results suggest that not much is gained if the construction is treated as a multiword lexeme since it still requires a treatment in terms of a rule apparatus. The nature of the rule itself, however, casts doubt on the idea that competent speakers of a language have tacit knowledge, i.e. access to all rules making up the language. If speakers of German can neither form new P+Noun$_{sg}$ on the fly, nor are able to judge P+Noun$_{sg}$ without a given contexts, the conclusion might very well be that P+Noun$_{sg}$ are combined by a rule which does not form part of speakers’s competence.

5 Figures

![Non-productivity in terms of V(N), V(1, N), and P(N)](image)

Figure (1): Distribution of V(N), V(1, N) and P(N) in case of finite (6) instances
Figure (2): Regularity in terms of $V(N)$ and $V(1, N)$

Figure (3): Regularity on the basis of $V(1, N)$
6 References


