**Between VP and NN: On the constructional types of German-er compounds**

# Between VP and NN

## On the constructional types of German *-er* compounds

Livio Gaeta and Amir Zeldes
Università di Torino / Georgetown University

This paper is concerned with the classification and analysis of different types of German synthetic compounds headed by deverbal agent nouns in *-er*, such as *Romanleser* 'novel-reader' or *Gedankenleser* 'mind-reader', where the non-head is seen to saturate an argument of the head lexeme while adhering to the semantic interpretation found in corresponding VPs (e.g. the distinct senses of *read* in the previous examples). In contrast to several previous approaches which attempt to explain the relationship between VPs and compounds using a unified mechanism of incorporation or derivation, we argue that different compounding patterns require different analyses and that the respective constructions are to some extent independent of each other. While some such compounds are modelled after frequent, familiar VPs and take account of the usage profile of syntactic phrases, other productive sets of compounds extend independently lexicalized schemas with fixed compound heads. To support our analysis we undertake the largest empirical survey of these formations to date, using a broad coverage Web corpus. We suggest several categories of verb-object lexeme pairs to account for our data and formulate an analysis of the facts within the framework of Construction Morphology.

**Keywords:** word-formation, synthetic compound, agent noun, incorporation, back-derivation, syntax, productivity, Construction Morphology, lexicon, corpus linguistics

### 1. Introduction

Synthetic compounds (SCs) are compounds in which the non-head saturates an argument of the head, a situation most commonly occurring in nominal

*Constructions and Frames* 9:1 (2017), 1–40. DOI 10.1075/cf.9.1.01gae
ISSN 1876–1933 / E-ISSN 1876–1941 

compounds with a deverbal head, as in English *car driver* or German *Autofahrer*.[1] These compounds are of particular interest to the discussion of the relationship between syntax and morphology, since they seem to require a word-level formation process to have access to "higher level" syntactic information for the head's base stem (the arguments of *drive*). Conversely, they require syntax to be aware of the saturation of these arguments at the word-level, thus making syntactic phrases such as *Autofahrer eines Porsche* 'car driver of a Porsche' infelicitous on account of the double realization of the argument.[2] At the same time, the semantic relationship of SC modifiers with syntactic arguments also needs to be accounted for: beyond the parallels in the subcategorization frames of verbs and nominalizations, where an accusative object lexeme is found filling a similar syntactic role in a compound, there is the question of its thematic role, and more precisely the semantic interpretation of the argument as it relates to the head. For instance, the sense of reading in *mind-reader* is the same as in *read [one's] mind*, but distinct from *novel-reader*, implying a tight semantic relationship between arguments in both constructions, which in our view requires some explanation.

Although there is probably a wide consensus on the necessity of a relationship between synthetic compounds and their verbal heads with regard to the inheritance of verbal properties like argument structure and semantic interpretation, it is not clear to what extent such a relation must be represented in grammar, whether one mechanism of derivation applies to all cases, and what the implications would be for expected empirical data. Is a homogeneous account of SCs sufficient to cover the empirical data on their usage? Are there different types of SCs requiring qualitatively different explanations, or also quantitatively different degrees of syntactic involvement within one account? If so, what kind of theoretical framework do we need to represent these? In this paper we will be taking a Construction Morphology approach to deal with German synthetic compounds, which exhibit diverse constructional types.

For example in German, some pairs of head and argument lexemes appear to be well-attested syntactically in VPs and morphologically in compounds with

1. As we will concentrate on deverbal SC heads, in this paper we refer primarily to the relationship between VPs and compounds, though some aspects of the discussion should also be relevant to deadjectival SCs or compounds with relational noun heads and their relationship with NPs. The latter case will be touched upon briefly further below.

2. One anonymous reviewer observes that the infelicity of *Autofahrer eines Porsche* may be due to pragmatic reasons as the expression is partially redundant and hence dispreferred since *Fahrer eines Porsche* already encodes the information that a car is being driven. We certainly agree and view the syntactic, semantic and pragmatic aspects of infelicitous multiple argument filling as different sides of the same coin.

the expected correlation in meaning, e.g. *Vogelbeobachter* 'bird-watcher' and *Vögel beobachten* 'watch birds'. Other cases are restricted to one construction or the other. For example, compounds with specific heads can be systematically paralleled by a different VP head, forming a sort of systematic suppletion; e.g. *Englischlehrer* 'English teacher' with the head *Lehrer* 'teacher', but the phrase *Englisch unterrichten* 'teach English' using a different verbal lexeme (**Englischunterrichter* 'English teacher') is ruled out, and systematically, *Lehrer* replaces the conceivable but blocked head (**Unterrichter*). Conversely, some lexemes available in VPs are either extremely rare or not attested at all in SCs, such as light verb constructions, e.g. *Gebrauch machen* 'make use' but not **Gebrauchmacher* 'use-maker'. In this paper we will therefore try to seek out mismatches and imbalances in the attestation of lexeme pairs in SCs and VPs and classify the sorts of incongruences that violate the one-to-one relationship between morphological and syntactic argument selection. The analysis of usage data will show that a monolithic approach to the phenomenon misses out on relevant subcategories that have intuitively interpretable reflexes in corpus data.

The paper is structured as follows: In the next section, we will briefly discuss previous approaches to SCs couched in different theoretical frameworks. In Section 3, we will present the corpus used in this study and describe the method used to automatically extract SCs and provide the spectrum of nominal arguments selected by each predicate. In Section 4, the extracted data will be used to identify groups of SCs based on their relationship with syntactic patterns. Section 5 discusses the productive behavior and the formation of novel SCs by examining rare and unique SC types (*hapax legomena*). Section 6 discusses the relevance of our findings for theoretical models and suggests a usage-based account of several types of German deverbal SCs in *-er* within the framework of Construction Morphology (Booij 2010).

## 2. Previous approaches

In previous work, three main directions modelling the relationship between VPs and compounds can be identified (either via transformations or other relations between constructions), which will be taken as reference points for our analysis. These range between the poles of a purely syntactic treatment to a more lexical one (cf. Gaeta 2010 for a critical survey):

i. Incorporation, i.e. morphological derivation via suffixation of a verb, as in *to read a novel* > *to novel-read* > *novel-reader*;

ii. Morphological derivation and subsequent compounding, as in *read* > *reader*; *novel* + *reader* > *novel-reader*;
iii. Morphological derivation via suffixation of a word group, as in *read+novel+er* > *novel-reader*.

In the first approach (cf. Lieber 1981; Baker 1988; Siebert 1999), a clear syntactic motivation is assumed via the mechanism of noun incorporation and on this basis an agent noun is subsequently derived. While it is not entirely clear whether noun incorporation should be considered as a form of compounding in this context or as a purely syntactic phenomenon (cf. Aikhenvald 2007), this approach suggests a direct identity between verb-argument relations in VPs and SCs, and leads to the question about the mechanisms constraining incorporation (why are verb forms like *novel-read* not generated in English in practice?) and its subsequent use in compounding (is any VP eligible for realization as a SC?). Even if these problems remain open for some cases, it is clear that some compounds lend themselves to such an analysis, especially in cases where incorporated forms are attested in VPs (including German *Autofahrer* 'car driver' as well as *autofahren*, lit.'car-drive').[3]

The second approach combines a morphological derivation of the agent noun with the inheritance of syntactic properties from the base verb, i.e. its argument structure. In this regard, there have been several proposals in the literature: the lexicalist view, represented in Booij (1988), sees argument inheritance in agent nouns like *reader* as the result of the combination of the semantic properties of the base *read* and the suffix *-er* via a binding of the verb's agent argument by the suffix. Later work by Booij in the framework of Construction Morphology (e.g. 2010: 49–50) explicitly allows an incorporated verb construction of the type $[[N][V]]_V$ to be unified with the constructional schema of synthetic compounds, even if the corresponding incorporated verb is not attested in isolation, as in **novel-read* above (see Section 6 for discussion of this proposal).

The Distributed Morphology approach of Alexiadou & Schäfer (2010), by contrast, makes the presence of a vP – in which v is assumed to be a 'verbalizer', i.e. the head that transforms a root into a verb – below the word level responsible for the verbal properties of the noun phrase. The DM approach comes especially close

3. The latter can be treated as a back-formation from *Autofahrer* 'car driver'. For an overview see also Wurzel (1998). For some evidence that synthetic status sometimes interacts with compounding markers such as linking elements in German, see Nübling & Szczepaniak (2011, 2013). Earlier generative approaches positing a fully-fledged syntactic derivation (e.g. Lees (1960) and Kürschner (1974) for German, also echoed in more recent work – cf. Roeper 2005: 141) will not be discussed here; however, they also lead to similar expectations with regard to the full correspondence of VP and SC argument behaviour (see also Botha (1984) for criticism and ten Hacken (2009) for a historical survey of such approaches).

to the incorporation approach in so far as it suggests that a syntactic structure, namely vP, is "incorporated" much like a stem below the word level. In both these approaches, the formation of the agent noun is always presupposed, so that the potential question regarding so-called *Zusammenbildungen*, i.e. compounds with no corresponding head noun attested independently, such as *Appetithemmer* 'appetite suppressant' (but ?*Hemmer* 'suppressant'), are difficult to motivate (see Leser (1990) for many examples and an in-depth critical discussion of the phenomenon).

Finally, the third approach suggests that, at least in the cases where the head noun is unattested, the morphological derivation takes as input a syntactic unit which has become stabilized in the lexicon. Thus, SCs like German *Totsagung* 'declaration of death' (lit. 'death-saying') cannot be formed on the basis of **Sagung* 'saying' because this form does not occur in German. This approach suggests that the sequence *totsagen* 'declare dead' has become a lexical unit (or a phrasal lexeme, cf. Masini 2009), and therefore it can form the input to the action noun-suffix *-ung*. While this approach solves the overgeneration resulting from the second approach, the mechanism it postulates is problematic for application to open-ended data, because it is difficult to foresee a priori whether a syntactic sequence has become a phrasal lexeme.

In this paper, we suggest that these approaches illuminate different types of SCs, which form a group of heterogeneous but closely related constructional schemas. These will be shown to have different empirical reflexes in the distribution of VP and corresponding SC pairs of head lexemes and their argument spectrums, independently from the theoretical model postulated. To this end, we will present a large scale corpus study of German SCs based on deverbal agent nouns in *-er* as in the examples of *Fahrer* 'driver' and *Beobachter* 'watcher' above, a compounding type which has stood at the heart of the discussion.[4] We will be concerned both with the structure of the lexicalized vocabulary of well-attested compounds and the behaviour of rare, non-lexicalized or novel compounds. For the former, we investigate whether there is a correlation between realized verbal argument types and the realized modifiers in SCs and whether verbs with more objects correspond to compounds with more modifiers. We will suggest that a correlation between these quantities indicates a joint, or interrelated usage profile which must be stored as such. For the latter, we aim to find out whether novel SCs correspond to established verb-object constructions or whether they have an independent vocabulary

4. The choice of German rather than English is further motivated by the well-known high productivity of compounding and especially of SCs in the former (cf. Schlücker 2012) as well as by the convenient word level marking of compounds (which are written as one word without spaces), making German the ideal testing grounds for empirical studies of SC formation (see the next Section for more details).

extended along the lines of existing patterns of heads and non-heads (cf. the positional families of de Jong et al. (2002) and the formulations of compounding schemas in Booij (2010)). If so, this would be evidence for at least some independence for each construction. Finally, we use the answers to these questions to see if there are different types of SCs, which in turn point to distinct mechanisms of compound formation and therefore warrant separate theoretical accounts.

## 3. Method

### 3.1 Data and motivation

In order to find sufficient examples of SCs and get a good idea about possible direct objects of verbs corresponding to their heads, a very large corpus is required. For this reason we use the DeWaC corpus (Deutsch: Web as Corpus, collected by the Web as Corpus WaCky initiative; Baroni et al. 2009) with approx. 1.63 billion automatically part-of-speech tagged and lemmatized tokens of Web data, sentence segmentation and data on source URLs for every page. This corpus is not balanced for genres or text types (see Sharoff 2010 for an analysis of text types in the corpus), but since SCs and transitive VPs are ubiquitous phenomena and both are searched for in a massive, very varied collection of the same texts, we expect good recall (i.e. we see no reason to assume the relationship between VPs and SCs will be skewed and we expect all relevant phenomena to be represented, even though errors will inevitably appear as well and require careful examination).[5]

It should be noted that, surprisingly, no comparable investigations of SCs on this scale have been carried out to date.[6] As a matter of fact, the data on which the proposals in the previous section are based come either from speakers' intuition or from dictionaries. Both sources are however difficult to evaluate. On the one hand, speakers' intuition may vary considerably depending on education, normative

5. In expecting the range of SC phenomena to be covered in such a large Web corpus we concur with Franz Rainer's statement: "One of the notorious disadvantages of corpora is the fact that they contain no negative evidence. Now, the Internet is so huge a corpus that one might suspect that the absence of certain instantiations of at least highly productive patterns should turn out to be a reliable indicator of their low acceptability." (Rainer 2003: 131–132). Nonetheless, comparability and reproducibility of results demand using a local copy of the data as a corpus, even if harvested from the Web, over using the services of a search provider such as Google, which cannot be reproduced (see Lüdeling et al. 2007).

6. A partial exception is Kohvakka & Lenk (2007) comparing the valency of German and Finnish agent nouns. However, although large corpora were used in that study, only ten specific heads were examined as examples.

attitude, etc., and perhaps mostly depending on context. On the other hand, dictionaries are not reliable either, because they do not record many compounds out of the huge number of potential compounds which are normally created in a language like German (cf. Barz 1995: 16). More generally, dictionaries usually discard completely regular and transparent formations, because "dictionary-users need not check those words whose meaning is entirely predictable from its [sic] elements, which by definition is the case with productive formations" (Plag 1999: 96). For SCs this is massively the case, since they are in most cases completely regular and transparent.[7] In this light, we claim that a large-scale corpus approach is essential to our question, if the relationship between actually occurring SCs and VPs is to be examined empirically.

The choice of German as a case study for SCs is motivated by several factors, such as the prevalence of compounds in general and SCs in particular in language use as well as the established literature on the description of the phenomenon in previous studies. A further facilitating circumstance is the fact that German compounds are easy to identify automatically in contrast, for instance, to English SCs, because they usually constitute a graphematic unit. Moreover, we decided to focus our attention on those SCs which are headed by an agent noun with the suffix *-er*, since they have figured prominently in the discussion of SC/VP relations. These have a further advantage in the relative ease of identifying potential compounds because of the homogeneous ending, as well as in the high proportion of compounds in which the modifier encodes a direct object argument with respect to the deverbal head. Agentive SCs also allow us to gather data relating to the argument structure of a verb much more reliably than for instance SCs headed by an action noun, especially those suffixed with the highly productive *-ung*. For the latter, the subject/agent role also comes into play as a possible modifier (e.g. *Politikerentscheidung* 'a politician decision'), which complicates the investigation.

### 3.2 Identifying and segmenting synthetic compounds

As a first step in detecting SCs with nominalized deverbal heads in *-er*, a regular expression search was carried out using the Corpus Workbench (Christ 1994) for common nouns (with the German STTS tag NN)[8] ending in *-er*, *-ers* or *-ern* (the last two being the genitive singular and dative plural forms of the *-er* ending). The optional *n* or *-s* were then removed, if found, and results were aggregated and

7. This does not exclude cases which are partially interpreted with the help of additional information retrieved contextually or on the basis of world knowledge. We return to this issue in Section 4 below.

8. For the STTS part-of-speech tag-set for German see Schiller et al. (1999).

counted. This preliminary result set of over 1.8 million types and over 25 million tokens was then classified according to four criteria:

The form

1. contains no non-alphabetic characters except hyphenation (numerals and other symbols were excluded)
2. begins with a capital letter (obeying standard German noun orthography)
3. exhibits two further canonical German syllables before the suffix *-er* (to ensure at least a stem vowel for each of the verbal head stem and the non-head stem)
4. does not have a vowel <i> immediately before <er>, except in cases of the diphthong written as <ei>, to account for heads like *-befreier* 'liberator', *-schreier* 'screamer' (otherwise the final ending <ier> is realized as a long vowel /i:/, which is incompatible with agent nouns in *-er*).

These criteria reduced the list to some 90,000 types representing 14 million tokens. For these candidates, a head/ non-head segmentation was attempted. In order to achieve this, a list of all verb forms ending in *-en* (the most reliable verbal marker) or lemmatized with a standard verbal lemma (ending in *-en*) in the corpus was retrieved (some 1.4 million types) and subjected to similar canonicity constraints (a minimal valid syllable structure and orthography, no hyphenation allowed, between 5 and 15 characters long), leaving over 180,000 types.

For each compound candidate, an attempt was made to match its right side to the longest possible verb, swapping *-en* (or *-n*, e.g. after *-r-*) for *-er*. Thus the compound *Präsidentenherausforderer* 'president challenger' was matched with the longest possibility *herausfordern* 'challenge' and not with its substring *fordern* 'demand'. Since *-er* nominalization may alter the verbal stem through umlaut or schwa metathesis around a liquid, alternative stem rules were used for cases liable to undergo such changes, e.g. one rule allows the alteration of *tragen* 'carry' into the non-existent *trägen* to match the *-er* noun form *Träger* 'carrier', and likewise *verlängern* 'extend' through *verlängeren* to *Verlängerer* 'extender' and *sammeln* 'collect' over *sammlen* to *Sammler* 'collector'.

The remaining material preceding the extracted verb stems was then tagged with TreeTagger (Schmid 1994) under two different conditions: "as is", and with certain possible linking elements (German *Fugenelemente*) removed if they were present, such as *-s*, *-e*, *-es*, *-n*, *-en*, *-ens* and *-er*, e.g. a linking *-s* after feminine non-heads in *-ung* as in *Abteilungs leiter* 'department manager' (from *Abteilung* 'department' and *Leiter* 'manager'). Obvious classes of non-head errors, such as individual letters tagged as a noun (e.g. the letter 'A'), and head-errors (through manual inspection of all verbs under 6 characters, many of which were erroneous) were removed.

Candidates for which the non-head was successfully tagged as a noun with a known lemma were then finally marked as SCs of the corresponding deverbal nominal head and non-head noun. In a few cases several forms were plausible nouns, in which case the longer form was heuristically preferred: e.g. *Eisenbrecher* 'iron breaker' has the non-head *Eisen* 'iron', although *Eis* 'ice' is also a possible non-head noun, if it were to take the possible linking element *-en*. Thus it is always assumed that the shortest possible string should be assigned to linking elements.

### 3.3 Identifying verbal objects

Given the size of the corpus used in the extraction of relevant data and the consequent risk of noise from errors, we decided to focus on precision rather than recall (for a similar strategy cf. Kawahara & Kurohashi (2005) on extraction of large datasets for PP-attachment resolution using highly selective patterns). Since word order in main clauses may have the object either before or after the finite verb, making it often indistinguishable from the subject, a part-of-speech regular expression was devised to target subordinate clauses, which are much more robustly SOV. The pattern allowed clauses of the form:

> CONJ (REFL) NP1 (ADJUNCTS) NP2 VVFIN
> Wherein:
> 1. CONJ is any subordinating conjunction (e.g. *dass* 'that', *weil* 'because', *wenn* 'if')
> 2. REFL is a possible reflexive pronoun
> 3. NP1 is a subject-capable NP (pronominal or nominal, with possible attributes and determiners compatible with nominative case, but not the pronoun *es* 'it', which may be a preposed object)
> 4. the ADJUNCTS area may contain any tokens except verbs or punctuation
> 5. NP2 is an object-capable NP (like NP1, but with appropriate accusative-capable articles or attributes, e.g. *einen* and *eine*, but not dative *einem*), which may not be preceded by a preposition
> 6. VVFIN is a finite lexical verb (non auxiliary)

By collecting the head of NP2 and the verb, this search resulted in over 900,000 potential verb-object lemma pair types, from almost 2 million token pairs. These verbs were then filtered using the non-verb table created for the segmentation of the compound heads, and pairs were compared to the extracted compound member pairs from the compound data for analysis, bearing in mind the recovered alternative stem forms (umlaut, metathesis etc.).

Though we are aware that this method may be flawed due to subordinate clauses displaying a restricted variety of syntactic patterns as compared to main clauses, we expect these limitations to be homogeneously distributed across all

verb patterns, so that their impact can only affect the number and frequency of the VP patterns without introducing any systematic distortions into the general results. This means that our database may be skewed to the extent that attestation in main clause OV pairs, as opposed to those in subordinate clauses, interacts with lexeme pair realization in SCs. For the remainder of this article, we will assume that OV pairs in subordinate clause VPs adequately represent OV pairs in VPs in general.

### 3.4 Classifying verbal and compounding lexeme pairs

A comparison of compound head + non-head pairs to the extracted OV pairs from VPs can in principle yield three groups of lexemes:

G1. OV pairs with no corresponding compound: [+ OV, – SC]
G2. compounds corresponding to OV: [+ OV, + SC]
G3. compounds with no corresponding OV: [– OV, + SC]

The relationship between these categories is illustrated in Figure 1 (the OV group is assumed to be much larger than the compound group):

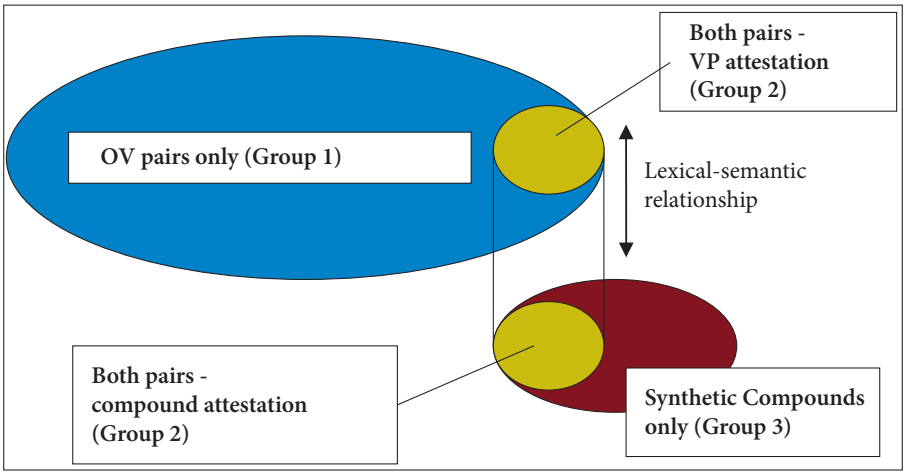


**Figure 1.** Overlapping groups of lexeme pairs attested as SCs and/or OV pairs

The axis of lexical-semantic relationship between the morphological and syntactic attestation of Group 2 lexemes (G2) is meant to represent the lexical identity of nominal and verbal stems as well as the semantic similarity in compound and VP interpretations mentioned in Section 1. Although compounds are semantically underspecified and can potentially have a limitless range of interpretations,

compounds paralleling VPs are mostly (but not always, see below) interpreted along the same lines as the corresponding syntactic constructions, e.g. a *Brunnenvergifter* 'spring poisoner' is someone who poisons springs, though of course another root compound reading is also conceivable, e.g. 'the poisoner from the spring' or any other relationship between 'spring' and 'poisoner' (cf. Downing's (1977) famous example *apple-juice seat* 'a seat in front of which a glass of apple-juice had been placed' or Heringer's (1984: 2) *Fischfrau* 'fish woman/wife', with at least 11 distinct senses). We will return to a discussion of the nature of the lexical-semantic relationship in Section 6.

## 4. Results

### 4.1 Group 1

Unsurprisingly, G1 forms the overwhelming majority of the data, with well over 900,000 types, of which some 40,000 appear well attested at over 5 times each. By contrast, we find under 5,000 well attested SCs in *-er* ($f > 5$) in the sets G2+G3, despite there being good reason to believe the compound search was much more exhaustive than the OV search (since we only considered a rather particular subset of transitive clauses using the high-accuracy subordinate clause pattern in Section 3.3). Less frequent SCs are discussed in Section 5 separately, though reliable estimates of their overall frequency are difficult due to high error rates. Based on the ratio of 40:5 in well attested types with subordinate clauses alone, the proportion of overlap within these, and an estimate of the sparsely attested types (see Section 5), we feel it is reasonable to assume that no more than 10% of syntactic objects, and probably much less, have a corresponding SC in *-er* in our data. A naïve expectation might be that the usage of SCs relies on the usage of corresponding OV pairs in VPs (i.e. SCs are formed by analogy to familiar VPs). If this were the case, then we should expect three predictions to hold true:

i. G3 should be very small
ii. G2 should contain all the heads and non-heads that are very common OV pairs[9]
iii. Those common OV pairs should also form the most frequent SCs.

**9.** An anonymous reviewer has remarked that these OV pairs, and especially the more common ones, can also be lexicalized at the syntactic level. We certainly agree, though it seems clear that a corresponding SC does not require this; we do not believe that we can identify the presence of lexicalization by means of SC attestation (in fact, the existence of G3 would provide a counter-example).

The first prediction is borne out by the data: only some 700 well attested SCs do not correspond to OV pairs in VPs. This is not very surprising, since SCs are a comparatively rare phenomenon whereas object filling is very common, though the identity of these compounds will merit an examination below. The second prediction would result if frequent OVs should have a higher chance of bringing about compounds, all other things being equal, since they correspond to more frequently used semantic concepts (i.e. the infrequency of a verb like *sift* might predict the infrequency of compounds with the head *sifter*). Using a random bag of words model, if we pull out a pair of head and non-head lexemes from the OV bag to construct an SC, then regardless of what the ratio G1 : G2 is, the most frequent type pairs in G1 are more likely to appear (as predicted in (ii) above) and they are likely to appear most often (as predicted in (iii)). These predictions can be tested in the data (see Table 1).

**Table 1.** The 50 most common VP pairs and corresponding SC attestation

| **VP attestation** | | | **SC attestation** | |
|---|---|---|---|---|
| O lemma | V lemma | *f* | SC lemma | *f* |
| *Rolle* 'role' | *spielen* 'play' | 5088 | *Rollenspieler* 'role player' | 780 |
| *Frage* 'question' | *stellen* 'put' | 2290 | *Fragensteller* 'inquirer' | 83 |
| *Gebrauch* 'use' | *machen* 'make' | 2134 | | |
| *Fehler* 'mistake' | *machen* 'make' | 1895 | | |
| *Entscheidung* 'decision' | *treffen* 'meet' | 1738 | | |
| *Spaß* 'fun' | *machen* 'make' | 1669 | *Spaßmacher* 'fun maker' | 400 |
| *Ziel* 'goal' | *erreichen* 'reach' | 1544 | | |
| *Geld* 'money' | *verdienen* 'earn' | 1461 | *Geldverdiener* 'money earner' | 27 |
| *Gedanke* 'thought' | *machen* 'make' | 1341 | | |
| *Rechnung* 'bill' | *tragen* 'carry' | 1267 | | |
| *Erfahrung* 'experience' | *machen* 'make' | 1246 | | |
| *Kommentar* 'comment' | *abgeben* 'submit' | 1237 | | |
| *Zeit* 'time' | *nehmen* 'take' | 1121 | *Zeitnehmer* 'time taker' | 272 |
| *Buch* 'book' | *lesen* 'read' | 1096 | *Buchleser* 'book reader' | 44 |
| *Wahrheit* 'truth' | *sagen* 'tell' | 1046 | *Wahrheitssager* 'truth sayer' | 7 |
| *Hilfe* 'help' | *brauchen* 'need' | 1028 | | |
| *Mühe* 'effort' | *machen* 'make' | 963 | | |
| *Anwendung* 'application' | *finden* 'find' | 925 | | |
| *Mühe* 'effort' | *geben* 'give' | 896 | | |
| *Sinn* 'sense' | *machen* 'make' | 893 | | |
| *Leben* 'life' | *führen* 'lead' | 865 | | |

*continued*

**Table 1.** (*continued*)

| VP attestation | | SC attestation | | |
|---|---|---|---|---|
| *Glaube* 'belief' | *schenken* 'bestow' | 827 | | |
| *Weg* 'way' | *finden* 'find' | 816 | | |
| *Kind* 'child' | *bekommen* 'get' | 806 | | |
| *Frage* 'question' | *beantworten* 'answer' | 803 | *Fragenbeantworter* 'question answerer' | 8 |
| *Problem* 'problem' | *lösen* 'solve' | 794 | *Problemlöser* 'problem solver' | 500 |
| *Geschichte* 'story' | *erzählen* 'tell' | 780 | *Geschichtenerzähler* 'story teller' | 1060 |
| *Möglichkeit* 'possibility' | *geben* 'give' | 756 | | |
| *Sorge* 'worry' | *machen* 'make' | 743 | | |
| *Leistung* 'service' | *erbringen* 'render' | 725 | *Leistungserbringer* 'service renderer' | 4269 |
| *Ziel* 'goal' | *verfolgen* 'pursue' | 691 | | |
| *Antwort* 'answer' | *geben* 'give' | 669 | *Antwortgeber* 'answer giver' | 32 |
| *Kenntnis* 'skill' | *erlangen* 'acquire' | 658 | | |
| *Verantwortung* 'responsibility' | *übernehmen* 'take over' | 650 | | |
| *Schaden* 'damage' | *nehmen* 'taker' | 647 | | |
| *Aufgabe* 'task' | *erfüllen* 'fulfill' | 644 | | |
| *Fortschritt* 'progress' | *machen* 'make' | 639 | | |
| *Schaden* 'damage' | *zufügen* 'inflict' | 631 | | |
| *Stimme* 'voice' | *hören* 'hear' | 624 | *Stimmenhörer* 'voice hearer' | 19 |
| *Einfluß* 'influence' | *nehmen* 'take' | 605 | | |
| *Tür* 'door' | *öffnen* 'open' | 605 | | |
| *Urlaub* 'vacation' | *machen* 'make' | 600 | *Urlaubsmacher* 'vacation organizer, travel agent' | 6 |
| *Möglichkeit* 'possibility' | *bieten* 'offer' | 592 | | |
| *Chance* 'chance' | *geben* 'give' | 584 | | |
| *Geld* 'money' | *bekommen* 'get' | 575 | | |
| *Bild* 'picture' | *machen* 'make' | 558 | *Bildermacher* 'picture maker' | 49 |
| *Krieg* 'war' | *führen* 'lead' | 553 | *Kriegführer / Kriegsführer* 'war leader' | 7/68 |
| *Lösung* 'solution' | *finden* 'find' | 549 | | |
| *Film* 'film' | *sehen* 'see' | 548 | *Filmseher* 'film seer' | 6 |
| *Grenze* 'border' | *setzen* 'set' | 548 | | |

Looking at the top 50 most common OV pairs, we find that less than half (only 17 pairs) have corresponding compounds attested, and what is more, the frequency of attested SCs does not correlate with VP frequencies, clearly contradicting predictions (ii) and (iii).

Analysis of entries not attested as SCs reveals several possible reasons. Most immediately we notice that *bekommen* 'get, receive' does not generally form the required *-er* nominalization (?*Bekommer* 'getter'), thus making compounds like ?*Kinderbekommer* 'children getter' unavailable.[10] This seems to be the case with some other, less frequent verbs, such as *erreichen* 'reach' (cf. ?*Erreicher* 'reacher') and *beantragen* 'apply for something' (cf. ?*Beantrager* 'applier'). The latter is probably lexically blocked by the compound *Antragsteller* 'applicant' (lit. 'application putter'). Another possibility is that the nominalization already exists, but does not form the required agent noun reading, as in *treffen* 'meet' with lexicalized *Treffer* 'hit, goal' but not **Treffer* 'meeter'. Also interesting is *bieten* 'offer (v.t.), bid (v.i.)' with *Bieter* 'bidder' representing only the intransitive reading, while **Bieter* 'offerer' appears to be ungrammatical (as opposed to the alternative near synonym particle-verb derivation *Anbieter* 'offerer'). From these examples it becomes clear that the corresponding compounds to these VPs are lexically blocked, since the prerequisite *-er* derivation cannot take place or is inappropriate. It is important to observe that this characterizes precisely frequent items, contrary to the predictions in (ii-iii) above, probably because frequent items are more prone to developing the entrenchment which leads to blocking (cf. Gaeta 2015 for a recent overview).

Another reason for the absence of an SC seems to be the idiomaticity of the VP which is incompatible with the base sense of the *-er* nominalization, especially

10. This statement must be qualified somewhat: while ?*Bekommer* is judged by many speakers to be completely unacceptable, sporadic examples of it can be found in compounds on the Web. Our corpus contains no isolated occurrences of the noun and exactly one example of a compound with this head, in the sentence:

– *Nun kriegen alle Null-Karten-Bekommer eine neue Chance*
– 'Now all zero-ticket-getters get another chance' (in reference to a sold out performance).

We therefore mark *Bekommer* and similar cases with ? rather than *. Uncontroversial, possible neologisms that are coincidentally not attested in our corpus will be marked with a °. Nevertheless, in the face of the extremely frequent *bekommen* 'get', the rarity of *Bekommer* (one in 1.63 billion tokens) is a clear indication that it is a highly marked, possibly conscious formation (cf. Bauer 2001: 66), and we take the absence of the expected wide range of compounds based on the arguments of the verb, both in this case and others like it, to be a significant fact of usage.

where the OV pair forms a light verb construction.[11] This seems to be the case for *machen* 'do, make', where *(sich) Gedanken machen* 'consider (lit. make (oneself) thoughts)', *Gebrauch machen* 'make use', *Sinn machen* 'make sense', *Fehler machen* 'make a mistake', *Erfahrung machen* 'gain an experience' and *Fortschritt machen* 'make progress' all do not exhibit the expected SCs: ?*Gedankenmacher* 'thought maker', ?*Gebrauchmacher* 'use maker' etc. Similarly, *Mühe geben* 'make an effort (lit. give exertion)' shows no corresponding ?*Mühegeber* 'effort maker', though *(eine) Antwort geben* 'give (an) answer' does parallel *Antwortgeber* 'answer giver'. It therefore seems that at least in some cases a nominalization in *er* does not accommodate light verb arguments.

Despite this, *Macher* 'maker' and *Geber* 'giver' are very common SC heads, with such frequent compounds as *Filmemacher* 'filmmaker' or *Geldgeber* 'money giver, financer' where the sense of the head resembles 'make' in the sense 'manufacture' and 'give' in a general sense, respectively (the behavior of the former is mirrored in English *carmaker* but not the idiomatic **use maker* or **sense maker*). Thus, the pattern licensing an SC with *Macher* requires a certain reading of *machen*, and likewise for *Geber*, which demonstrates that not every VP is likely to have a corresponding SC in usage. At the same time it must be noted that we do find some idiomatic combinations like *Urlaub machen* 'go on vacation' (lit. 'make vacation') next to *Urlaubsmacher* 'travel agent' (lit. 'vacation maker'), which do not seem to be blocked, although the interpretation of the two pairs is different (the compound follows the 'object creation' pattern of e.g. *Filmemacher* 'filmmaker', against the expectation of a consistent correspondence in semantics). OV idiomaticity therefore seems to be a definite hindrance to the formation of SCs with the same lexemes, but not as an absolute barrier. In any case, selectivity is strongly in evidence here, both quantitatively (some of the SCs are very rare despite frequent OVs) and in some cases almost categorically, i.e. with no attestation in this large corpus, which might not necessarily imply ungrammaticality *strictu sensu*.

These results show that in addition to a very wide variety of rare OV pairs which may coincidentally not be attested in SCs yet, we must recognize groups of pairs which either form no suitable nominalization or else only form a subset

**11.** A word of warning must be added here against confusing idiomatization and lexicalization. We carefully distinguish between what is idiomatized, i.e. opaque or non-compositional at a certain level of (morpho-phonological, semantic) analysis, and what is lexicalized, i.e. entrenched in the lexicon. The latter usually results from the high frequency of a pattern and/or the absence of a syntactic pattern to be referred to. Notice that while what is lexicalized is not necessarily idiomatized, the opposite is necessarily true. Thus, syntactic idiomatized patterns also have to be lexicalized, such as the aforementioned *(sich) Gedanken machen* 'think about, lit. make thoughts', which qualifies as an idiom. In the rest of this article, we will make reference to lexicalization in the specific sense of lexical entrenchment (cf. Bybee 2010, 2013).

of possible SCs, notably using certain senses of the nominalization. These may at the same time be well attested with multiple non-heads. The sub-classification of Group 1 is summarized in Table 2.

**Table 2.** Summary of OV non-attested SC types

| | Type | OV example | SC example |
|---|---|---|---|
| G1.a. | idiomatized, esp. light verbs | *Gedanken machen* 'make thoughts' | ?*Gedankenmacher* 'thought maker' |
| G1.b.i. | resists nominalization | *Kind bekommen* 'get children' | ?*Kinderbekommer* 'children getter' |
| G1.b.ii. | unsuitable nominalization | *Möglichkeit bieten* 'offer a possibility' | *Bieter* 'bidder' but: ?*Möglichkeitsbieter* 'possibility offerer' |
| G1.c. | available (but unattested) | *Polizeiwache stürmen* 'storm a police station' | °*Polizeiwachenstürmer* 'police station stormer' |

Thus these groups show some input restrictions on SC formation evidenced in usage: some lexicalized units like *Kinder bekommen* 'get children' or *(einen) Fehler machen* 'make (a) mistake' are not available for suffixation (as suggested in the third account in Section 2), just as some, even incorporated VPs (e.g. the incorporated light-verb constructions like *Gebrauch machen* 'make use', where *Gebrauch* 'use' occurs without an article), do not exhibit corresponding SCs, which might be expected to be possible in the incorporation approach discussed in Section 2.

### 4.2 Group 2

Turning to Group 2, we find a small but well-attested set of some 4,500 lexeme pair types, which can be identified with high accuracy by cross-referencing VP attestation with all compounds in *-er*. A possible expectation that SC frequency should correspond to VP frequencies is not met here either (see Figure 2): Statistically speaking there is no significant correlation between the frequencies at all (Spearman's $r^2 = 0.01897$, $p > 0.05$).

Thus some compounds are much better attested than the corresponding VPs and vice versa, while some are more balanced between the two. In order to identify these groups automatically, we divide the VP frequency by the compound frequency, producing a ranking from proportionally 'most verbally attested' to 'most nominally attested'. Table 3 shows an excerpt from the top, middle and bottom of the result set (the position of the top and bottom items in the Table has also been highlighted in Figure 2):


**Figure 2.** SC vs. VP frequencies for well-attested lexeme pairs (logarithmic)

Items from the middle of the Table, with frequency ratios of about 10:1 in favor of VP realization, form the core of compositionally interpretable transparent SCs often discussed in the literature. Here we find compounds like *Vogelbeobachter* 'bird watcher', *Metallbearbeiter* 'metal processor' or *Wachstumsbeschleuniger* 'growth accelerator', where the sense of the head is virtually identical to the corresponding verb, but a nominal interpretation of it independently of the verb is not ruled out. The top of the Table shows items which intuitively suggest a phrasal prototype as the model for the compound lexeme pair, such as *Wahrheitssager* 'truth teller' (from *(die) Wahrheit sagen* 'tell the truth') or *Fragenbeantworter* 'question answerer' (from *(eine) Frage beantworten* 'answer (a) question') where the sense of the compound corresponds to the syntagm more closely and the realization of the head without its argument would be odd or have a different sense (e.g. 'sayer, teller' or 'answerer'). The prevalence of the syntactic realization suggests that some of these VPs may also be collocations, which also fits with the tendency to keep them in their familiar, entrenched form.

Finally, at the bottom of the Table we find more 'morphological' cases, where the compound has a very entrenched lexicalized meaning that may also be idiomatic, i.e. not immediately or completely derivable from the VP. Rather, a verbal realization is back-derived from or interpreted by analogy to the lexicalized compound if it should occur (cf. Wurzel's 1998 'reverbalizations'): the syntagm *Arbeit nehmen* in the sense 'take work, be employed' is clearly derived from *Arbeitnehmer* 'employee (lit. work taker)', which the staggering frequency difference should make clear (*Arbeit nehmen* is attested only 15 times, of which the

**Table 3.** Pairs with mostly VP realization, balanced attestation, and mostly compound realization

| Compound | Object | Verb | *f*(VP) | *f*(SC) | *f*(VP)/*f*(SC) |
|---|---|---|---|---|---|
| *Wahrheitssager* 'truth sayer' | *Wahrheit* 'truth' | *sagen* 'say' | 1046 | 7 | 149.4286 |
| *Fragenbeantworter* 'question answerer' | *Frage* 'question' | *beantworten* 'answer' | 803 | 8 | 100.375 |
| *Urlaubsmacher* 'vacation organizer' | *Urlaub* 'vacation' | *machen* 'make' | 600 | 6 | 100 |
| … | | | | | |
| *Vogelbeobachter* 'bird watcher' | *Vogel* 'bird' | *beobachten* 'watch' | 11 | 77 | 0.142857 |
| *Kartenbenutzer* 'card user' | *Karte* 'card' | *benutzen* 'use' | 3 | 21 | 0.142857 |
| *Wachstumsbeschleuniger* 'growth accelerator' | *Wachstum* 'growth' | *beschleunigen* 'accelerate' | 3 | 21 | 0.142857 |
| *Pflegedienstleister* 'nursing service provider' | *Pflegedienst* 'nursing service' | *leisten* 'provide' | 2 | 14 | 0.142857 |
| *Metallbearbeiter* 'metal processor' | *Metall* 'metal' | *bearbeiten* 'process' | 2 | 14 | 0.142857 |
| *Videoverleiher* 'video lender' | *Video* 'video' | *verleihen* 'lend' | 2 | 14 | 0.142857 |
| … | | | | | |
| *Wahlleiter* 'election supervisor' | *Wahl* 'election' | *leiten* 'lead, head' | 1 | 2724 | 0.000367 |
| *Themenstarter* 'discussion initiator' | *Thema* 'topic' | *starten* 'start' | 1 | 2747 | 0.000364 |
| *Sternengucker* 'star gazer' | *Stern* 'star' | *gucken* 'look' | 1 | 4126 | 0.000242 |
| *Arbeitnehmer* 'employee' | *Arbeit* 'work' | *nehmen* 'take' | 15 | 135183 | 0.000111 |

majority do not even exhibit the idiomatic sense 'be employed', whereas the SC is attested 135,183 times). The unusual ability to interpret 'taking work' as referring to becoming an employee (note this is not strictly speaking compositionally predictable) is related to the 'worker' sense of the idiomatic compound. Similarly *Wahlleiter* 'election supervisor', *Themenstarter* 'discussion initiator (lit. topic starter)' and *Sternengucker* 'stargazer' are fairly frequent, but the corresponding VPs occur only once (although for the latter case a somewhat frequent PP realization exists: *in die Sterne gucken* 'gaze into the stars', which is however nowhere nearly as common as the SC). Importantly, although their semantics may seem more

transparent, some similar subtleties in meaning can be observed, e.g. *(eine) Wahl leiten* lit. 'lead (an) election' can mean 'work as an election supervisor', where again the sense of professional work typical for deverbal agent nouns in *-er* is reflected in an interpretation of the VP.

From the point of view of frequency and entrenchment, we take the very rare items (VPs and SCs), and particularly those occurring only once in our corpus (also called *hapax legomena*, from Classical Greek 'said once') to signal different types of non-lexicalized, productive formations (cf. Baayen 2009 and Gaeta & Ricca 2015 for an overview). This would mean that the verbal realization in the latter cases is also productively formed, whereas the frequent SCs are probably familiar to most speakers, and thus serve as the basis for the 'back-derivation' of these syntagms. The importance of hapax legomena in predicting the behavior of unseen novel cases will be discussed in more depth in Section 5 below.

In line with the description above, we can divide G2 into three subgroups, though clearly the transitions between the groups are gradual since they express the scalar dominance of compounding or syntactic realization for lexeme pairs. Table 4 reviews these non-sharply defined subgroups.

**Table 4.** Summary of pair types attested as both VP and SC

| | Type | VP example | SC example |
|---|---|---|---|
| G2.a | mostly OV (possibly collocation) | *(die) Wahrheit sagen* 'tell (the) truth' | *Wahrheitssager* 'truth sayer' |
| G2.b | balanced (possibly collocation/lexicalization) | *Vögel beobachten* 'watch birds' | *Vogelbeobachter* 'bird watcher' |
| G2.c | mostly SC (possibly back-derivation/ lexicalization) | *Arbeit nehmen* 'take work' | *Arbeitnehmer* 'employee' |

Thus verbal collocations[12] and lexicalized SCs form two sides of the same coin, but the lexemes in question behave independently, with neither construction necessarily implying the other should be as established in the mental lexicon.

**12.** The term 'collocations' as used for VPs (as opposed to SCs) is meant to designate the more flexible nature of lexicalized phrasal combinations that nevertheless allow some variation in terms of word order, definiteness, number and modification of objects, etc.; on the other hand, a lexicalized SC at the word level assumes the same internal form in each case (cf. also Booij's (2010: 20) use of the term 'constructional idiom', and see below on modifiability as a motivation for SC-modeled VP derivation).

### 4.3 Group 3

The evaluation of Group 3 (SC-only pairs) is most difficult, since the fact that no VP cross-reference can be found for its members means accuracy is rather low in the automatic search (the existence of a corresponding VP pair is what corroborates the compound's synthetic status in the previous category). A large variety of items that appear formally identical to synthetic compounds are thus in fact semantically interpreted as root compounds with oblique semantic relations. Some examples to illustrate the problem are:

(1) *Radarbeobachter* 'radar observer' is not someone who observes radars, but rather someone who observes (e.g. air traffic) using a radar.

(2) *Nischenhersteller* 'niche manufacturer' is not someone who manufactures niches, but a manufacturer occupying a market niche.

Note that these cases could conceivably be read as SCs, i.e. 'observing a radar (screen)' and 'creating a niche' (*herstellen* can mean both 'manufacture' and 'create'), but these were not the intended corpus senses. Still, some real examples of SCs not attested as VPs can be found, usually as a result of strong lexicalization, especially of the head noun. Table 5 shows some of the most frequent cases where the relationship between head and modifier is not opaque after manual filtering.[13]

*Krankheitserreger* 'pathogen' is a lexicalized combination, though the formation is quite transparent from *Krankheit* 'disease' and *erregen* 'excite, provoke'; a corresponding syntactic combination is unattested in the corpus despite a frequency of well over 5000 cases for the SC. We do not contest the fact that Google examples for such VPs and others on the list can be actually found. However we take their extreme rarity as an indication that in these cases it is likely that different forces are at work from those operating in G2 (see Section 6 for discussion).

There are also many grades of partial lexicalization affecting head nouns with more than one particular non-head, especially in the lower frequencies. Thus *-Vertreter* 'sales representative' in *Staubsaugervertreter* 'vacuum cleaner salesman' preserves a reading of the verbal *vertreten* 'to sell as a salesperson (esp.

**13.** We have rejected from our analysis completely opaque or idiomatized candidates, such as *Schriftsteller* 'writer' (but literally a 'writing stander'). Though morphologically analyzable as a synthetic compound, this example is semantically completely opaque (it is hard to see what reading of the verb *stellen* 'to stand (sth.)' could be relevant), and therefore stands in no relation to the verb. Some other doubtful candidates were rejected if the argument was not strictly accusative, e.g. *Reiseführer* 'travel guidebook' (but literally 'trip guider'). In this case, which is certainly a lexicalized term but not completely opaque, the relationship is not clearly accusative: the book in question 'guides on a trip' but it is arguable if it actually 'guides the trip' in the sense of a direct object.

**Table 5.** SC lexemes unattested in VPs and their corresponding constituent lexemes

| SC | Object | Verb | f(SC) |
|---|---|---|---|
| *Krankheitserreger* 'pathogen' | *Krankheit* 'disease' | *erregen* 'provoke' | 5481 |
| *Wirtschaftsprüfer* 'financial auditor' | *Wirtschaft* 'economy' | *prüfen* 'check' | 5347 |
| *Mobilfunkbetreiber* 'mobile communications operator' | *Mobilfunk* 'mobile communications' | *betreiben* 'operate' | 3207 |
| *Handelsvertreter* 'trade representative' | *Handel* 'trade' | *vertreten* 'represent' | 3009 |
| *Automobilhersteller* 'automobile manufacturer' | *Automobil* 'automobile' | *herstellen* 'manufacture' | 2923 |
| *Reiseleiter* 'tour guide' | *Reise* 'tour' | *leiten* 'lead, head' | 2584 |
| *Medienvertreter* 'media representative' | *Medien* 'media' | *vertreten* 'represent' | 2506 |
| *Konkursverwalter* 'liquidator' | *Konkurs* 'bankruptcy' | *verwalten* 'administrate' | 2146 |
| *Staubsaugervertreter* 'vacuum cleaner salesman' | *Staubsauger* 'vacuum cleaner' | *vertreten* 'sell (door to door)' | 116 |

door to door)', which is no longer attested with this object and is probably well on the way to becoming obsolete (although numerous compounds are found, e.g. *Versicherungsvertreter* 'insurance salesman', *Teppichvertreter* 'carpet salesman'). The verb is currently used overwhelmingly with the sense 'represent'. The same head is found with precisely this more common sense higher up in Table 5, where it signifies 'representatives' of various commercial functions, such as *Handelsvertreter* 'trade representative' or *Medienvertreter* 'media representative'. Here the question is why we do not find any attestation for verbally representing trade or the media, despite the remarkable frequency of these nouns. A possible answer is that *-Vertreter* as a head has become a prototypical pattern for deriving a certain class of professions (representatives), which bears the morphological appearance of a syntactically motivated compound, but does not require a non-head which is semantically compatible with the object role of the corresponding verb. Indeed, both 'trade' and 'media' are inanimate and somewhat non-concrete conglomerations of organizations, whereas the typical objects of *vertreten* in our corpus seem to be mainly ideas or opinions if they are abstract (e.g. *Meinung* 'opinion', *Auffassung* 'view', *Position* 'position') or else concrete humans (e.g. *Rechtsanwalt* 'lawyer', *Mitglied* 'member', *Mensch* 'person') or groups of humans (e.g. *Unternehmen* '(a) business', *Gruppe* 'group'). It is therefore possible that the difficulty with these non-heads as objects lies in the fact that it is not entirely clear who is being represented, which is more compatible with the underspecified semantics of compounds, but not as compatible with the VP construction, even though the latter might be conceivable and grammatical in principle.

Another verbal argument incompatibility may explain *Reiseleiter* 'tour guide'. Here the problem may be that one does not really guide a tour, but rather the people in the tour (see also Section 4.3 for a discussion of such compounds as cases of metonymy). It is worth mentioning that non-heads containing *Reise* 'trip, tour' are also hardly attested for *leiten*, with only 2 hits for *Reisegruppe leiten* 'lead a tour group'. This shows that beyond the better compatibility for groups of people with the verb, there is either less need for the verbal realization with these lexemes, or else the entrenchment of the frequent *Reiseleiter* leads speakers to choose the compound when they need to express the corresponding meaning. We therefore feel that it is likely that a syntagm like *(eine) Reise leiten* 'guide a tour' is derived under the influence of the entrenched word *Reiseleiter* (attested over 2,500 times), rather than the other way around.

Similar problems may explain *Mobilfunkbetreiber* 'mobile communications operator' or *Konkursverwalter* 'bankruptcy administrator, liquidator', unparalleled in the corpus by a verbal *Mobilfunk betreiben* or *Konkurs verwalten*. An explanation of these cases as non-synthetic or simply as root compounds is not satisfactory for several reasons: firstly on semantic grounds, since the heads do require an argument (a *Betreiber* 'operator' implies something being operated) and the non-heads in fact fill this argument position (what is being operated/administered, etc.); secondly, on paradigmatic grounds, since we find analogous cases that are attested syntactically (e.g. *Veranstaltungstechnik betreiben* 'operate event-technology', *Erbe verwalten* 'administrate inheritance, bequest'); and finally, since sporadic attestation of the syntagms in question can be found outside the corpus using a Google search, in documents almost certainly produced by native speakers of German:

(3) *Alle anderen sollen bleiben wo sie sind und ihren hausgemachten Konkurs verwalten*
'Everybody else should stay where they are and administrate their homemade bankruptcy' (http://www.nachtwelten.de/vB/history/topic/40937-1.html, accessed July 6th, 2012)[14]

Thus, even if syntagms are not categorically ruled out but are simply very strongly dispreferred, such data does not lend itself to the incorporation or deverbal derivation scenarios introduced in Section 2. A more plausible explanation, especially in the context of a usage-based model, would be that syntagms like (3) are generated through a verbalization of the SC, i.e. by a conflating activation or unification of both a lexically specified SC and a lexically unspecified VP pattern. We therefore

14. The choice of the VP in this case may be facilitated by the speaker's desire to modify 'bankruptcy' in the NP 'homemade bankruptcy'. This modification would not have been possible within the compound. See Section 6 on a schema unification analysis accounting for such cases.

interpret the results from Table 5 as showing compounds that are independent from syntactic realization, and the rare syntactic realizations that may be found to correspond to them (often only as hapax or dis legomena) as secondarily derived from the lexical entries of the compounds. The mechanism responsible for these 'asyntactic' synthetic compounds in the first place will be discussed in Section 6.

To sum up, G3 contains either SCs that are strongly lexicalized as a lexeme pair so that a nominal realization is preferred (e.g. *Krankheitserreger* 'pathogen'), or else SCs with heads that have a special semantics or preserve an idiosyncratic sense independently of an underlying verb (e.g. *-Vertreter* 'salesman', where the corresponding verb is nearly obsolete).[15] In less extreme cases, the head is lexicalized to the extent that it becomes divergently compatible with arguments that seem to be avoided with the verb, though compositionality is still in evidence and the senses conform with those found inVPs.

### 4.4 Interim conclusion

In sum, we reach the classification of SCs in Table 6.

It is worth noting that the classes in Table 6 can to a large extent be extracted automatically on empirical criteria: G1.a-b are found by looking for frequent VP lexemes with no corresponding SC, while a. can be distinguished from b. by checking if the required head nominalization is attested with other non-heads. G1.c consists mostly of rare VP pairs with no SC. The range between G2.a-c is formed by ranking entries according to the ratio of SC frequency to VP frequency, and G3.a-b is the remainder of SCs with no VP attestation. Here we expect all the items in G3.c to be frequent (indeed *Krankheitserreger* heads the list), but a clear cutoff point is unlikely since lexicalization does not correspond to a precise frequency threshold. The distinction between G3a and b requires recourse to semantics and would be difficult to establish automatically (perhaps using distributional semantic methods; cf. Wulff 2008; Erk 2012). As these criteria show, the classes are not all categorical and clearly defined, but rather gradual (especially the subclasses of G2), which is to be expected if we keep in mind that levels of lexicalization and idiomatization can differ.

**15.** Indeed the case of *Schriftsteller* 'writer' in footnote 13 above goes one step further, where complete opacity leads to a formation which looks outwardly like a synthetic compound but is completely idiomatized and no longer perceived as containing an argument relation.

**Table 6.** Synopsis of SC classes

| | Type | OV example | SC example |
|---|---|---|---|
| G1.a | idiomatized | *Gedanken machen*<br>'make thoughts' | ?*Gedankenmacher*<br>'thought maker' |
| G1.b.i | no nominalization | *Kind bekommen*<br>'get children' | ?*Kinderbekommer*<br>'children getter' |
| G1.b.ii | no suitable nominalization | *Möglichkeit bieten*<br>'offer a possibility' | *Bieter* 'bidder' but:<br>?*Möglichkeitsbieter*<br>'possibility offerer' |
| G1.c | possible (unattested) | *Polizeiwache stürmen*<br>'storm a police station' | °*Polizeiwachenstürmer*<br>'police station stormer' |
| G2.a | mostly OV (possibly collocation) | *(die) Wahrheit sagen*<br>'tell (the) truth' | *Wahrheitssager*<br>'truth teller' |
| G2.b | balanced | *Vögel beobachten*<br>'watch birds' | *Vogelbeobachter*<br>'bird watcher' |
| G2.c | mostly SC (possibly lexicalized) | *Arbeit nehmen*<br>'take work' | *Arbeitnehmer*<br>'employee' |
| G3.a | idiosyncratic argument (non-head avoided with V) | *#Konkurs verwalten*<br>'administrate bankruptcy' | *Konkursverwalter*<br>'bankruptcy administrator' |
| G3.b | idiosyncratically lexicalized head | (Obs.) *Staubsauger vertreten*<br>'sell vacuum cleaners' | *Staubsaugervertreter*<br>'vacuum cleaner salesman' |
| G3.c | lexicalized compound | *#(eine) Krankheit erregen*<br>*'cause a disease'* | *Krankheitserreger*<br>'pathogen' |

## 5. Synthetic compounds and productivity

In this section, we will attempt to answer some questions about the nature of the different types of synthetic compounding as productive processes: are productively formed, non-lexicalized SCs usually syntactically motivated (i.e. is there a syntactic parallel available on which to model the compound, and is it found in the data)? Can we identify productive series of heads or non-heads that act independently of their phrasal counterparts? Do more productive SC members also correspond to more productive syntactic elements (i.e. do heads corresponding to verbs with more varied objects also have a wider variety of compound non-heads)? If not, when, how often and why?

### 5.1 Identifying and measuring productivity

Though long considered to be one of the "unclearest terms in linguistics" (Mayerthaler 1981: 124), productivity has enjoyed increasing attention in

recent years, especially in the area of word formation (for an overview see Bauer 2001; Gaeta & Ricca 2015). Very coarsely defined, the study of productivity relates to the establishment of criteria to determine whether a certain linguistic process, such as word formation with a certain affix (e.g. nouns in *ness*), is productive or not, or else to what extent it is productive. These criteria are then used to assign processes to a certain productivity grade: in algebraic theories, this will be a binary decision or one of a number of discrete categories (fully productive, semi-productive, unproductive etc.), whereas usage-based approaches typically attempt to assign a numeric value on a scale, or at least to compare processes ordinally (e.g. noun formation in *ness* is more productive than the formation in *ity*).

Here we adopt the scalar approach of Baayen (1993, 2001, 2009), who uses vocabulary size (type counts, designated by *V* for vocabulary) and counts of hapax legomena to estimate productivity rates using corpus data. The idea behind using such counts to assess productivity can be understood intuitively if we consider that the attested vocabulary size of a certain process corresponds to how productive it has been up until now. Thus a process with more types has a higher 'realized productivity', in Baayen's terms, than one with fewer types (see also Barðdal 2008; Zeldes 2012). On the other hand, to assess how prone a process is to forming neologisms (regardless of whether it is used often or rarely), we may want to know what the proportion of neologisms is in its output – a process with mostly neologisms is very productive, whereas a repetitive process, with few neologisms, has little 'potential productivity', no matter how large its realized vocabulary so far. As mentioned above, hapax legomena, as a natural superset of neologisms (assuming every neologism is unique, which is not always true), are taken to estimate the rate of neologism formation in the pattern in question.[16]

With these concepts at hand, we now turn to examine first the realized vocabulary of different SCs as compared to the argument vocabulary of corresponding VPs, and then the behavior of SC hapax legomena.

### 5.2 Do more VP objects mean more SC non-heads?

If productively formed SCs come primarily from the prototypical class found in G2.b above, i.e. cases like *Vogelbeobachter* 'bird watcher' where the semantic interpretation of the SC and its corresponding VP is very similar, then a verb with a wide variety of objects (high realized productivity in Baayen's terms) might

**16.** For reasons of space, the issue has been simplified considerably. Some mathematical complications arising from this also cannot be discussed here – see Evert & Lüdeling 2001; Gaeta & Ricca 2006; Säily 2011. For a discussion of productivity in Italian and English compounds see also Gaeta & Ricca 2009.

naively be expected to correspond to a head with many non-heads (if one watches many things, then there might be watchers of many things). This hypothesis can be tested in our dataset: Figure 3 plots the number of non-heads for an SC head against the number of distinct object lemmas attested for the corresponding verb.

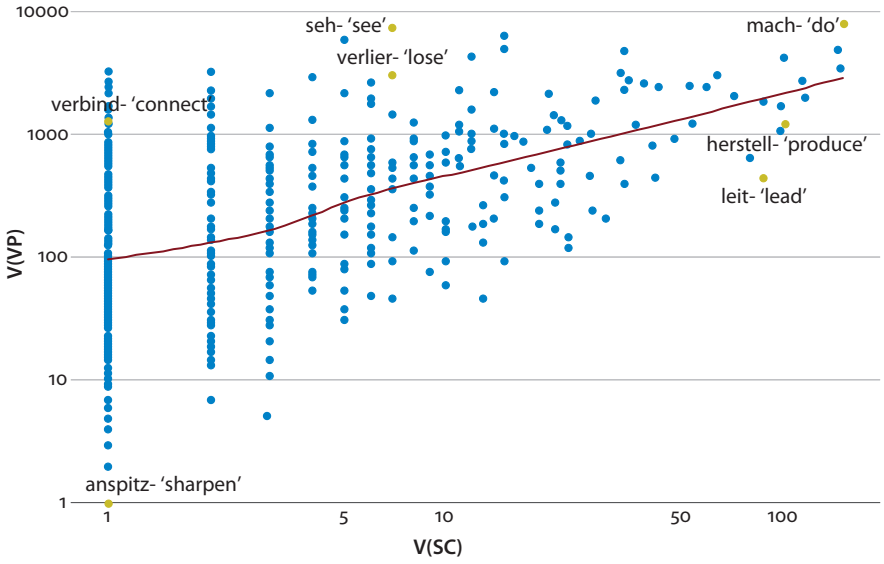


**Figure 3.** Weak but significant correlation for number of VP objects and SC non-heads for each head lexeme. Locally weighted scatterplot smoothing (the diagonal curve) suggests a log linear relationship.

The data shows that the vocabulary of verbal objects *V(VP)* is significantly, though rather weakly, correlated ($p < 0.001$, Spearman's $r^2 = 0.2161$) with the vocabulary of compound non-heads *V(SC)*.[17] A locally weighted scatterplot smoothing in the Figure above suggests the relationship is more or less loglinear, but clearly most points are quite far from the regression line. This results from the fact that most SC heads are rare and have very few modifiers in the data regardless of how many objects the corresponding verb has (the first three vertical clusters of points on the left, which are very populous for verbs with anything between 5 and 3,000 objects). At the same time, we find some outliers at the edges of the data with either

17. The coefficient of determination $r^2 = 0.2161$ means that only about a fifth of the variance for one construction can be predicted or explained by the data of the other. This is rather surprising, as one would expect a considerable correlation even if only on semantic grounds: verbs with very few arguments should correspond to few SCs if at all, and the reverse should hold for verbs with very many arguments. That this is not overwhelmingly the case is a measure of the at least partly independent nature of SC and VP formation processes, likely owing in large part to the semantics of nominalization discussed below.

extraordinarily many objects and still few SC types (the left hand side of the topmost data points, e.g. *seh-* 'see', *verlier-* 'lose'), or many non-heads but comparatively few objects (bottom of the right edge, e.g. *leit-* 'lead, head', *herstell-* 'produce'). These results indicate different realized productivity tendencies: some lexemes are more prone to SC formation than others, and this is understandable in view of the examples in the data. The verb *sehen* 'see', for instance, has over 7,400 objects compared to only 7 SC non heads. It accumulates a very wide variety of arguments while describing an atelic 'seeing' action (anything one can see), but the semantics of the corresponding agent noun *Seher* 'seer' imply a description of someone who sees something either habitually or as a profession, which is a fairly unusual situation (some exceptions are *Geisterseher* 'ghost seer', attested partly as Friedrich Schiller's unfinished novel of that title, and *Sternseher* 'star seer', an antiquated term for 'astrologist' but also a species of fish, Lat. *Uranoscopus scaber*). From a formal semantic point of view, 'seer' could of course also designate the experiencer of a punctual seeing action, but this type of usage is not found in our data.

Conversely, while it is quite possible to verbally head or lead a group of people or an institution, the variety of bodies that have heads, leaders or managers referred to in an SC with the head *Leiter* is disproportionally large against the expectation based on the number of objects found for the verb. It therefore appears that the meaning of the morpheme LEIT- common to *leiten* and *Leiter* is more suitable for the SC context than the morpheme SEH-. Finally, there are also cases which are very productive in both constructions, for example *machen* / *Macher* 'do, make / doer, maker', which has the highest number of both verbal arguments and SC non-heads in the data. This is understandable on account of the very general meaning of the verb, since one often refers to the action carried out by a practitioner of that action; still, the overwhelming ratio of objects to non-heads tells us that most 'doing' actions do not lead to a nominalization in the form of such a practitioner, which makes sense intuitively. With examples of both disproportionate SC and VP prevalence and only a modest correlation, the hypothesis connecting verbal and compounding realized productivity should be rejected: separate realized productivity is in evidence in many cases. Given this detachment of the two phenomena, in the next Section we will try to focus only on novel SCs in order to find out which conditions lead to productive SC generation and how they relate to a possible VP realization of SC lexemes.

## 5.3 How are novel SCs generated?

Looking at Figure 3 above, it is not surprising that some heads do not appear in any novel SCs in our data: some items head very few compounds in general, and the semantics of a verbal head can be more or less appropriate for the

nominalization (e.g. *Seher* 'seer', where both of these factors are evident, produces no hapax legomena at all in our data). Items in the right half of the diagram, by contrast, contain more than just a relative abundance of familiar, entrenched SCs – they also produce novel items. In keeping with Baayen's methodology outlined above, we will look for the patterns with the most hapax legomena in order to find the most productive SC heads. Do novel SCs largely correspond to already established, better attested VP pairs? Since most of the frequent SCs correspond to VP pairs (recall that G2 is substantially larger than G3) and VPs are in general much more frequent, it might be reasonable to expect at least a majority of cases where a corresponding VP is already attested for a hapax SC. Table 7 lists top SC heads ordered by their hapax count, along with the proportion of these hapax legomena also attested as VPs.

**Table 7.** Top hapax SC heads and the proportion of OV attestation

| SC head | SC hapax types | Attested as VP | VP/SC |
|---|---|---|---|
| *Hersteller* 'manufacturer' | 1130 | 92 | 0.081416 |
| *Leiter* 'head, leader, manager' | 1057 | 51 | 0.04825 |
| *Führer* 'head, leader, manager' | 867 | 147 | 0.16955 |
| *Besitzer* 'owner, possessor' | 802 | 178 | 0.221945 |
| *Anbieter* 'provider, offerer' | 716 | 136 | 0.189944 |
| *Vertreter* 'representative' | 664 | 71 | 0.106928 |
| *Manager* 'manager' | 644 | 10 | 0.015528 |
| *Macher* 'maker, doer' | 629 | 240 | 0.381558 |
| *Betreiber* 'operator' | 568 | 57 | 0.100352 |
| *Lehrer* 'teacher' | 392 | 30 | 0.076531 |
| *Bewohner* 'inhabitant' | 381 | 12 | 0.031496 |
| *Sender* 'sender, transmitter' | 366 | 21 | 0.057377 |
| *Sammler* 'collector' | 344 | 1 | 0.002907 |

As we can see, nowhere near half of these SC heads' hapax legomena are attested in corresponding VPs. Here we would like to suggest that the ratio R of VP/SC type attestation gives an empirical assessment of the intuitive syntactic motivation for compounds with that particular head. As expected, heads typical for the agent noun sense and signifying occupations, like *Leiter* 'head, leader, manager' or *Lehrer* 'teacher' have extremely low ratios (well under 10%, despite VPs being the much more common construction). It therefore appears, for instance, that one teaches far fewer things with the verb *lehren* 'teach' than one can be a *Lehrer* 'teacher' of. It should be noted again that the verbal sense for many of these possible

objects is supplied by the near synonym *unterrichten* 'teach', which usually avoids the agent noun formation in *– er* (?*Unterrichter*). However, the most extreme case in the Table, *Sammler* 'collector' has under 0.3% hapax SCs attested in VP types, even though no corresponding verbal substitute to the verb *sammeln* 'collect' comes to mind. It is thus plausible that novel types of collectors are modeled after established formations with the head 'collector', without necessarily building on the speaker's experience with VPs involving the verb *sammeln* 'collect', which are much more limited in our corpus.

For all verbs in Table 7, syntactic lexical behavior seems to be a bad predictor of productive compounding vocabulary, indicating that low values of the ratio R correspond to low syntactic motivation. For these cases, syntactic attestation does not appear to be a chief motivator in the selection of lexemes realized in new SCs. But what are the reasons for the behavior of these hapax legomena? Let us take a closer look at some novelties within this group of SCs for a few representative heads. For these items, we wish to know whether the lack of a VP realization is accidental or whether it is excluded for some reason. Latter cases would obviously support a separate mechanism of derivation not involving VP usage info, though at the same time they would challenge the definition of (all) SCs as potentially corresponding to verbal patterns.

Number two in Table 7, which was also the most morphologically oriented head lexeme in Figure 3 above, is *Leiter*, making both its realized and its potential productivity chiefly compound related. Only 51 of its over 1,000 hapax legomena are syntactically attested, with a few being variants or errors for otherwise well-attested pairs (e.g. *Stationleiter* for *Stationsleiter* '(medical) department chief'; cf. Nübling & Szczepaniak 2011, 2013 on variation in linking elements in German compounds); the remaining approx. 950 lexemes also contain a few errors, but the large majority of cases are valid compounds. On careful examination, it is possible to find a few items which seem implausible for VP realization. These generally hinge on a sort of metonymy or ellipsis, whereby the non-head indirectly refers to the intended object extension, e.g. *Besuchsleiter* 'visit guide' or *Betreuungsleiter* 'support/care leader'. We have (unsystematically) received conflicting speaker judgments as to the acceptability of *einen Besuch leiten* 'guide a visit' and *eine Betreuung leiten* 'lead a support', but judgments agree on the preferability of *Besuchergruppe leiten* 'guide (a) visitor group' and *Betreuungsgruppe leiten* 'lead (a) support group'.

Another head showing metonymic cases is *Hersteller* 'manufacturer', with most hapax SCs in total. Here we find *Nahrungsergänzungshersteller* 'food-supplementation manufacturer' or *Erotikhersteller* 'erotica manufacturer', which do not have verbal correspondences, and may be considered either metonymic or simply truncated versions of the also attested *Nahrungsergänzungsmittelhersteller* 'food-supplementation-substance manufacturer' and *Erotikartikelhersteller* 'erotic item

manufacturer'.[18] If we compare these compounds to similar compounds with non-deverbal heads, we find that this kind of metonymy may actually be characteristic of compounding in general, e.g. *Erotikladen* 'erotica shop' for *Erotikartikelladen* 'erotic product shop'.[19] It therefore appears that the increased possibility of metonymy is not a specific feature of SCs, though it may still set SC argument spectrums apart from those of corresponding VPs.

A different subtype of VP infelicity results from a lexical mismatch which seems to be enforced only for the verbal variant. The head *Besitzer* 'owner, possessor' can take some hapax non-heads for which speakers verbally choose *haben* 'have' instead of *besitzen* 'possess': *Ausbildungsplatzbesitzer* '(lit.) apprenticeship place possessor' but *einen Ausbildungsplatz haben* 'have an apprenticeship place', and preferably not ?*einen Ausbildungsplatz besitzen* 'possess an apprenticeship place', and similarly for the near synonym compound *Lehrstellenbesitzer*. The fact that *haben* 'have' does not form an agent noun in *-er* may also play a role here, motivating a sort of suppletion of *Besitzer* for **Haber* 'haver',[20] but the absence of *besitzen* with these objects is still conspicuous. Similar cases with *-Besitzer* occur with objects which are not prototypical concrete possessions, but for which it is relevant to establish categories of people having them or not, e.g. *Breitbandbesitzer* 'broadband owner' (but in our corpus ?*Breitband besitzen* 'possess broadband' is unattested). VP variants have been judged to be more awkward by informants and are conspicuously missing despite the high frequency of the same object lexemes with other verbs, though a controlled grammaticality study remains to be carried out.

These cases show that many novel SCs are formed without necessarily referring to lexicalized, or even rarely attested VPs. This applies not only to hapax legomena: a similiar majority of syntactically unattested lexeme pairs for these heads can be found in dis legomena, tris legomena, etc. Since items that are so infrequent in a corpus of the size used here cannot be expected to be lexicalized, the formation of these compounds must be seen as spontaneous and the mechanisms licensing their generation require an explanation. This leads us back to the question whether or not a uniform explanation of all SCs (or at least the ones with agent nouns in *-er* examined here) is possible. Here we claim that different derivations, each with several subtypes, must be responsible both for more syntactically motivated cases,

**18.** Here we should note that unlike English erotica, Erotik cannot mean concrete things; thus if Erotik herstellen existed, it would mean something like 'create an erotic atmosphere'. Similarly, Nahrungsergänzung means 'supplementing food', while the supplements themselves are normally referred to as Nahrungsergänzungsmittel '(lit.) nutrition supplementation means'.

**19.** We thank Anke Lüdeling for this observation.

**20.** Some exceptional lexicalized compounds do actually take the head -Haber, such as Machthaber 'power-haver, someone in power', but the pattern is not in productive use.

where not only relative VP/SC frequency but also semantic similarity and overlapping argument selection show a close connection to VPs, and for more morphologically motivated cases, where novel compounds seem to take other lexicalized compounds as a model.

That there should be morphologically motivated novel compounds should come as no surprise. In fact, it would be rather surprising if speakers were able to create novel compounds without consulting VPs for non-deverbal relational nouns, e.g. *Mafiaboss*, but not for deverbal heads like *Leiter* 'head, leader, manager', as in the hapax legomenon *Sondereinsatzkommandoleiter* 'special deployment commando leader' (notwithstanding the possibility of a syntactic °*Sondereinsatzkommando leiten* 'lead (a) special deployment commando'). If the agent noun *Leiter* is lexicalized (as both its frequency and its being listed in dictionaries would lead us to believe), it should be subject to the same relational noun compounding as its near synonym *Boss*, which may be argued to possess a similar argument structure. With these results in mind we now turn to an analysis of the different types of compound constructions within the framework of Construction Morphology. Our primary goal is to reconcile the relationship between VPs and SCs with the more morphologically motivated cases, offering mechanisms that could better explain the empirical data by using a series of constructional schema unifications.

## 6. Synthetic compounds and phrases as interrelated schemas

In the previous sections, we have shown that although German *-er* compounds often have a close relationship with related verbs and their selection of arguments, they do not form a homogeneous group, particularly with respect to this relationship. By comparing actual realization patterns of SCs and VPs in large quantities of data, it is possible to divide compounds into more or less syntactically motivated cases, where disparities are found owing to different factors (Table 6). These categories, despite showing gradual transitions, show marked differences between extremes on a scale of syntactic motivation and, upon closer examination, reveal some semantic groups which offer an intuitive interpretation of this situation (especially the affinity of the class of 'professions' to the SC preference for agent nouns in *er*). It also appears that (de)verbal heads can be prolific in forming compounds, VPs, or both but, importantly, no necessary correlation holds between these facts. Finally, some compounds which appear to follow the SC pattern are conspicuously absent from VP attestation, for the reasons we have explored above. A grammar covering productive derivation of SCs as well as modeling their usage will need multiple mechanisms to predict these preferences adequately.

In order to meet this requirement, we suggest an analysis based on Construction Morphology (CM, Booij 2010). As a form of Construction Grammar (CxG, see Goldberg 2006, 2013), CM assumes that the basic units of grammar are constructions, pairings of form and meaning to which usage information, such as degrees of entrenchment, may be attached. The mental lexicon in CxG is seen as holding not only fully specified lexemes but also partially specified schemas with open slots or variables, such as *jog [one's] memory*, or even lexically unspecified constructions with no phonological material, such as the English ditransitive construction. These then account for the occurrence of productive combinations of smaller units or constructions. The mental lexicon is thus organized hierarchically, with constructions building on existing constructions. In Booij's application of CxG concepts to the domain of morphology, productive word formation is seen as relying on such schematic constructions or 'schemas', in which underspecified formations of the form $[[\text{Prefix-X}]_X$ represent e.g. prefixation in general, while specific prefixes are each represented in a separate construction, e.g. $[\textit{un}\text{-A}]_A$ for adjective negation with *un-*. Thus the prefix *un-* exists only as part of a construction. Crucially, schemas can be *unified*, meaning that multiple compatible schemas may apply simultaneously, e.g.:

$$[\textit{un}\text{-A}]_A + [\text{V-}\textit{able}]_A = [\textit{un}\text{-}[\text{V-}\textit{able}]_A]_A \qquad \text{(Booij 2010: 42)}$$

Schema unification of this sort helps explain why we find many negative potential adjectives in English without a corresponding positive base for the derivation (e.g. lexicalized *unbeatable*, listed in many dictionaries, but only very rarely *beatable*, which is usually not listed).

In our view, the different types of *-er* compounds we have seen above are generated via different constructional unification processes. The general schema unification for productively formed, syntactically motivated *-er* compounds is given in Figure 4, roughly following Booij (2010: 49–50). Lines signify constructional inheritance from top to bottom and single braces span the relevant parts for unification where ambiguity might result.

As a type of compound, *-er* SCs naturally instantiate the general compounding schema and its more specific subschema, NN compounding. At the same time, the head of the compound is a deverbal agent noun which is licensed by a particular type of suffixal derivation schema. Finally, since the SC contains both verb and noun lexemes, and since the verbal lexeme semantically supports an argument structure, the schema is unified with a verbal compound schema of the type $[\text{N-V}]_V$, as assumed by Booij, despite the fact that a corresponding incorporated verb is usually not attested outside of the agentive SC schema (e.g. **vogelbeobachten* 'to bird-watch' is not formed in German, though it is in English). A subtle difference in our analysis of these cases as compared to Booij's is that we assume an explicit

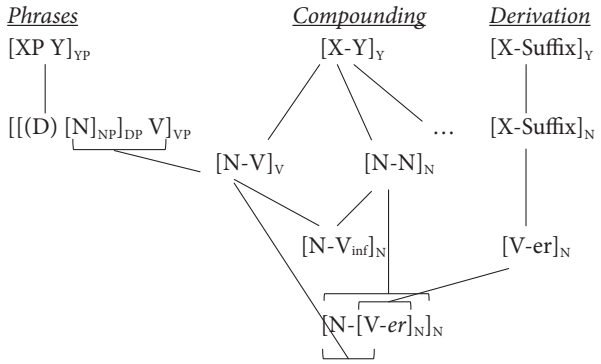


**Figure 4.** General unification schema for syntactically motivated, unlexicalized *-er* compounds.

link between compounds and syntactic VP's with full NP arguments. This type of relationship is necessary to account for the identical interpretation in VPs corresponding to syntactically motivated SCs and the usage similarities in selectional preferences: the information about usage of the verb is employed in conventional VP argument selection and SC argument selection. The lexical noun and verb are unified with the [N-V] component of the SC schema. In support of this assumption, we note the occurrence of NN-compounds which are headed by a nominalized infinitive reflecting the VN-sequence as represented in Figure 4: $[N\text{-}V_{inf}]_N$. In fact, *Vogelbeobachten* is commonly attested as a compound, e.g. *In Deutschland ist das Vogelbeobachten zugegebenermaßen relativ langweilig* 'Admittedly, in Germany bird-watching is relatively boring'.[21]

On the other hand, the (indirect) link with the OV-sequence does not jeopardize the essentially morphological nature of the SCs, which is expressed by the basic $[N\text{-}[V\text{-}er]_N]_N$ schema assumed in Figure 4. That the latter has to be preferred over the incorporating schema $[[N\text{-}V]_V\text{-}er]_N$ is shown – besides reverbalizations like *Arbeit nehmen* discussed below – by the occurrence of linking elements in SCs such as *Abteilungsleiter* 'department manager' discussed in Section 3.2 above, which clearly refer to a paradigmatic dimension normally found with NN-compounds. Notice that compounds headed by a nominalized infinitive display similar linking elements: [*N*]*ur Hermann Weis … ist wohl durch das Abteilungsleiten zu sehr ab-*

**21.** Cf. https://www.travel-to-nature.de/blog-detail/auf-der-suche-nach-dem-goettervogel-und-warum-man-in-costa-rica-voegel-beobachten-sollte/, accessed September 17th, 2016.

*gelenkt* 'Only Hermann Weis is too distracted certainly because of the department management'.[22]

Moving on to morphologically motivated cases, we assume that novel *-er* compounds can be formed by analogy to existing compounds with the same head.[23] The schema unification scenario is given in Figure 5 for a novel compound headed by *Leiter* 'head, manager, leader'.

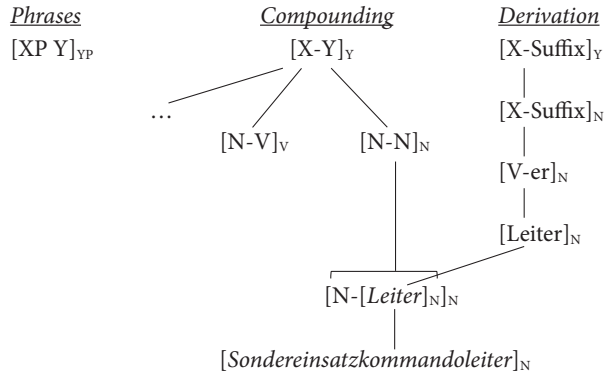


**Figure 5.** Unification schema for novel compounds headed by lexicalized *-Leiter*.

The head *-Leiter* can still be analyzed as a deverbal agent noun derived with the suffix *-er*, which also motivates its inherited argument structure. However, a key difference is the fact that compounds in *-Leiter* already form an entrenched schema, meaning that specific kinds of *-Leiter* can now be derived without any reference to VP semantics or usage preferences. In this way, *Leiter* behaves much like a relational noun of the type *Boss* and the schema can form, for example, *Sondereinsatzkommandoleiter* 'special deployment commando leader'. In extreme cases, the meaning of the lexicalized head may drift away from the base verb's meaning as is the case with *-Vertreter* 'salesman', whose underlying verb sense is on the verge of becoming obsolete; or the head may even be limited to individual fully specified cases, as in the idiomatic *Schriftsteller* 'writer, lit. writing stander', or the lexicalized but transparent *Krankheitserreger* 'Pathogen, lit. disease exciter'. The dissociation or gradual lack of unification with the VP construction makes it

**22.** Cf. https://tt-freimann.de/archiv/fr_saisonabschluss2001_5.php, accessed September 17th, 2016. We leave it open whether reverbalizations and paradigmatic relations are treated more adequately with the help of second order schemas as suggested by Booij (2015).

**23.** The same can be said for compounds with a shared modifier motivating new compounds with that modifier (cf. de Jong et al. 2002; Baayen et al. 2010 i.a.), though that is not our focus here.

possible for the head to have different preferences and even different semantics from the corresponding verb.

Finally, in cases where a novel VP is generated on the model of a lexeme pair that is entrenched as an SC, the direction of inheritance in the phrasal domain on the left is reversed, as shown in Figure 6, for the phrase *(einen) Konkurs verwalten* 'administrate (a) bankruptcy'.

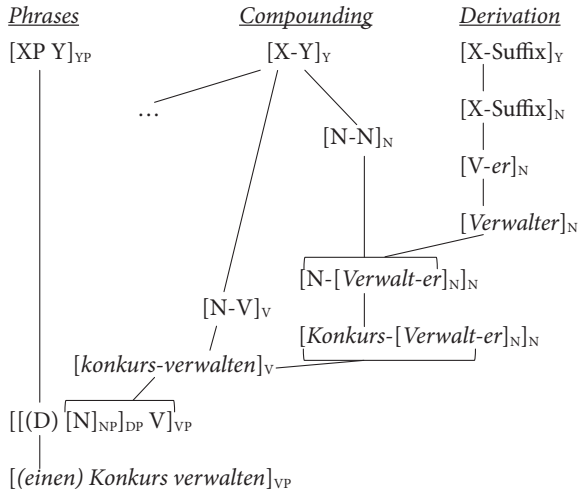


**Figure 6.** Unification schema for the backformation of *(einen) Konkurs verwalten* 'administrate bankruptcy'

In this case, the motivating lexeme is fully specified, i.e. the lexicalized compound *Konkursverwalter* 'bankruptcy administrator' mediates between the more general SC schema and the VP realization. It is an open question whether an unattested verb of the form **konkursverwalten* 'bankruptcy-administrate' should be assumed as part of the derivation, but if we accept Booij's position that a schema can be productive only within another schema, then we may consider the $[N\text{-}V]_V$ schema to be a participant in the derivation of *(einen) Konkurs verwalten* 'administrate (a) bankruptcy', which otherwise results from unmediated conflation of the lexeme *Konkursverwalter* and the VP schema, without reference to the status of the constituent lexemes of the compound as nominal and (de)verbal. This profiles a backderivation or a reverbalization similar to that found in *Arbeitnehmer* 'employee' / *Arbeit nehmen* 'take work, be employed' above. Note that the resulting phrasal construction can of course be unified with a variety of other schemas and need not assume only this exact form. In fact, the motivation for producing the back derivation may be closely related to this possibility: in Example (3) in Section 4.3 above, the adjectivally modified phrase *ihren hausgemachten Konkurs verwalten*

'administrate their homemade bankruptcy' is incompatible with the compound schema (cf. **hausgemachte Konkursverwaltung*), quite possibly motivating the formation of the VP construction for unification with the adjective that is required by communicative needs.[24]

Cases of suppletion such as *haben* 'have': *Besitzer* 'owner' and *unterrichten* 'teach': *Lehrer* 'teacher' can also be accommodated if we assume conventionalized links between these constructions within the structured lexicon. It seems likely that established 'have' phrases can give rise to novel 'owner' compounds and vice versa, and the absence of the expected compounds in '-haver', compensated for with '-owner', seems to confirm this idea. This process requires unification of partially lexicalized constructions with non-identical items, but would appear to be inevitable in general if partially specified suppletive constructions are to be accounted for within the framework of Construction Morphology (the authors are as yet not familiar with work addressing this issue, which would therefore be a point for further study).

## 7. Conclusion

Using the corpus data explored in the previous sections and the succession of unification steps sketched out above, we hope to have shown that there are indeed different types of German synthetic compounds in *-er*, which can be more morphologically or syntactically motivated, and which can and should be analyzed in distinct ways. All types of compounds can be derived with the same machinery and from the same basic inventory of schemas within the framework of Construction Morphology; however, the combinations, their order and the direction in which existing prototypes are generalized to novel cases can be different in ways that explain semantic differences and senses carried over between domains (e.g. denominal VPs which can mean 'work as Xer' and not just 'do X', and compounds that retain collocational meanings). At the same time, the suggested analysis offers motivation for the observed preferences of strongly lexicalized compounds to avoid or block the occurrence of corresponding VPs, and of verbal collocations with specific senses (especially in the case of light verbs) to resist realization as deverbal compounds.

---

**24.** Notice that it is not generally impossible to have bracketing paradoxes in German compounds, where an adjective modifies only the modifier of the compound as in [reitende Artillerie]kaserne 'horse artillery barracks'. This is however only possible when the adjective and the modifier form a phrasal lexeme as in the case of reitende Artillerie 'horse artillery' (cf. Gaeta & Ricca 2009: 36–37 for a discussion).

In a sense, it appears that all of the approaches introduced in Section 2 capture important aspects of the behavior of synthetic compounds found in this survey. Yet, a key advantage of the constructional approach that has not been taken advantage of to date is that we are under no obligation to derive all compounds that are superficially similar in the same way. It can be hoped that further work within this approach will lead us to a better understanding of the transitions between the prototypical compounding types listed in Table 6, including the way in which lexicalized compounds change their nature gradually over time and how conventionalized schemas arise. Here, we also see a challenge in extending the discrete analyses suggested here to a cognitively plausible gradient model of unified schema activation, which remains to be formalized in Construction Morphology.

## Acknowledgements

Parts of this paper were presented at the Humboldt-Universität Berlin (2009) and at the *8th Mediterranean Morphology Meeting (MMM8)* held in Cagliari, 15–16.9.2011. We thank all people present on these occasions as well as Geert Booij and one anonymous reviewer for valuable suggestions and remarks. Needless to say, we are solely responsible for views expressed and mistakes.

## References

Aikhenvald, A. Y.) (2007). Typological distinctions in word-formation. In T. Shopen (Ed.), *Language typology and syntactic description*. 2nd edition. Vol. 3 (pp. 1–65). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511618437.001

Alexiadou, A., & Schäfer, F. (2010). On the syntax of episodic vs. dispositional *-er* nominals. In A. Alexiadou & M. Rathert (Eds.), *The syntax of nominalizations across languages and frameworks* (Interface Explorations 23) (pp. 9–38). Berlin & New York: De Gruyter Mouton. doi: 10.1515/9783110245875.9

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology 1992* (pp. 181–208). Dordrecht: Kluwer. doi: 10.1007/978-94-017-3710-4_7

Baayen, R. H. (2001). *Word frequency distributions*. (Text, Speech and Language Technologies 18). Dordrecht, Boston & London: Kluwer. doi: 10.1007/978-94-010-0844-0

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook*, Vol. 2 (pp. 899–919). Berlin: Mouton de Gruyter. doi: 10.1515/9783110213881.2.899

Baayen, R. H., Kuperman, V., & Bertram, R. (2010). Frequency effects in compound processing. In S. Scalise & I. Vogel (Eds.), *Cross-disciplinary issues in compounding* (Current Issues in Linguistic Theory 311) (pp. 257–270). Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/cilt.311.20baa

Baker, M. (1988). *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic* (Constructional Approaches to Language 8). Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/cal.8

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3), 209–226. doi: 10.1007/s10579-009-9081-4

Barz, I. (1995). Komposita im Großwörterbuch Deutsch als Fremdsprache. In I. Pohl & H. Ehrhardt (Eds.), *Wort und Wortschatz. Beiträge zur Lexikographie* (pp. 13–24). Tübingen: Niemeyer.

Bauer, L. (2001). *Morphological productivity* (Cambridge Studies in Linguistics 95). Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511486210

Booij, G. E. (1988). The relation between inheritance and argument structure: Deverbal *-er*-nouns in Dutch. In M. Everaert, A. Evers, R. Huybregts, & M. Trommelen (Eds.), *Morphology and modularity. In honour of Henk Schultink* (pp. 57–74). Dordrecht: Foris Publications.

Booij, G. E. (2010). *Construction Morphology*. Oxford: Oxford University Press.

Booij, G. E. (2015). The nominalization of Dutch particle verbs: Schema unification and second order schemas. *Nederlandse Taalkunde*, 20, 285–314.

Botha, R. P. (1984). *Morphological mechanisms: Lexicalist analyses of synthetic compounding*. Oxford: Pergamon Press.

Bybee, J. L. (2010). *Language, usage and cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511750526

Bybee, J. L. (2013). Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (Eds.), *The Oxford handbook of Construction Grammar* (pp. 49–69). Oxford: Oxford University Press.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. *Proceedings of Complex*, 94, 23–32. Budapest.

de Jong, N. H., Feldmand, L. B., Schreuder, R., Pastizzo, M., & Baayen, R. H. (2002). The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, 81(1–3), 555–567. doi: 10.1006/brln.2001.2547

Downing, P. A. (1977). On the creation and use of English compound nouns. *Language*, 53(4), 810–842. doi: 10.2307/412913

Erk, K. (2012). Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10), 635–653. doi: 10.1002/lnco.362

Evert, S., & Lüdeling, A. (2001). Measuring morphological productivity: Is automatic preprocessing sufficient? In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of Corpus Linguistics 2001* (pp. 167–175). Lancaster.

Gaeta, L. (2010). Synthetic compounds with special reference to German. In S. Scalise & I. Vogel (Eds.), *Cross-disciplinary issues in compounding* (pp. 219–235). Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/cilt.311.17gae

Gaeta, L. (2015). Restrictions in word formation. In P. O. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation. An international handbook of the languages of Europe*, Vol. 2 (pp. 858–874). Berlin & New York: Mouton de Gruyter. doi: 10.1515/9783110246278-004

Gaeta, L., & Ricca, D. (2006). Productivity in Italian word formation: A variable-corpus approach. *Linguistics*, 44(1), 57–89. doi: 10.1515/LING.2006.003

Gaeta, L., & Ricca, D. (2009). Composita solvantur: Compounds as lexical units or morphological objects? *Italian Journal of Linguistics / Rivista di Linguistica*, 21(1), 35–70.

Gaeta, L., & Ricca, D. (2015). Productivity. In P. O. Müller, I. Ohnheiser, S. Olsen, & F. Rainer (Eds.), *Word-formation. An international handbook of the languages of Europe*, Vol. 2 (pp. 841–858). Berlin & New York: Mouton de Gruyter. doi: 10.1515/9783110246278-003

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldberg, A. E. (2013). Constructionist approaches to language. In Th. Hoffmann & Gr. Trousdale (Eds.), *The Oxford handbook of Construction Grammar* (pp. 15–31). Oxford: Oxford University Press.

Heringer, H. J. (1984). Wortbildung: Sinn aus dem Chaos. *Deutsche Sprache*, 12, 1–13.

Kawahara, D., & Kurohashi, S. (2005). PP-attachment disambiguation boosted by a gigantic volume of unambiguous examples. In R. Dale, K. -F. Wong, J. Su, & O. Y. Kwong (Eds.), *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)* (pp. 188–198). Berlin & Heidelberg: Springer.

Kohvakka, H., & Lenk, H. (2007). 'Streiter für Gerechtigkeit' und 'Teilnehmer am Meinungsstreit'? Zur Valenz von Nomina agentis im Deutschen und Finnischen. In H. Lenk, & M. Walter (Eds.), *Wahlverwandtschaften. Valenzen – Verben – Varietäten. Festschrift für Klaus Welke zum 70. Geburtstag* (pp. 195–218). Hildesheim, Zurich & New York: Georg Olms.

Kürschner, W. (1974). *Zur syntaktischen Beschreibung deutscher Nominalkomposita. Auf der Grundlage generativer Transformationsgrammatiken*. (Linguistische Arbeiten 18). Tübingen: Niemeyer. doi: 10.1515/9783111635729

Lees, R. B. (1960). *The grammar of English nominalizations*. The Hague: Mouton de Gruyter.

Leser, M. (1990). *Das Problem der 'Zusammenbildungen': eine Lexikalistische Studie*. Trier: WVT Wissenschaftlicher Verlag.

Lieber, R. (1981). *On the organization of the lexicon*. PhD Thesis, University of New Hamsphire.

Lüdeling, A., Evert, S., & Baroni, M. (2007). Using web data for linguistic purposes. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus linguistics and the web*. (Language and Computers-Studies in Practical Linguistics 59) (pp. 7–24). Amsterdam & New York: Rodopi. doi: 10.1163/9789401203791_003

Masini, F. (2009). Phrasal lexemes, compounds and phrases: A constructionist perspective. *Word Structure*, 2(2), 254–271. doi: 10.3366/E1750124509000440

Mayerthaler, W. (1981). *Morphologische Natürlichkeit*. Wiesbaden: Athenaion.

Nübling, D., & Szczepaniak, R. (2011). *Markmal(s?)analyse, Seminar(s?)arbeit* und *Essen(s?) ausgabe*: Zweifelsfälle der Verfugung als Indikatoren für Sprachwandel. *Zeitschrift für Sprachwissenschaft*, 30(1), 45–73. doi: 10.1515/zfsw.2011.002

Nübling, D., & Szczepaniak, R. (2013). Linking elements in German origin, change, functionalization. *Morphology*, 23, 67–89. doi: 10.1007/s11525-013-9213-9

Plag, I. (1999). *Morphological productivity. Structural constraints in English derivation* (Topics in English Linguistics 28). Berlin & New York: Mouton de Gruyter. doi: 10.1515/9783110802863

Rainer, F. (2003). Studying restrictions on patterns of word-formation by means of the Internet. *Italian Journal of Linguistics / Rivista di Linguistica*, 15(1), 131–139.

Roeper, T. (2005). Chomsky's *remarks* and the transformationalist hypothesis. In P. Štekauer & R. Lieber (Eds.), *The handbook of word-formation* (pp. 125–146). Dordrecht: Springer. doi: 10.1007/1-4020-3596-9_6

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141. doi: 10.1515/cllt.2011.006

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Technical report, Universität Stuttgart, Institut für maschinelle Sprachverarbeitung & Universität Tübingen, Seminar für Sprachwissenschaft.

Schlücker, B. (2012). Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In L. Gaeta & B. Schlücker (Eds.), *Das deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte* (Linguistik – Impulse & Tendenzen 46) (pp. 1–25). Berlin: Mouton De Gruyter. doi: 10.1515/9783110278439.1

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the Conference on New Methods in Language Processing* (pp. 44–49). Manchester, UK.

Sharoff, S. (2010). In the garden and in the jungle. Comparing genres in the BNC and Internet. In *Genres on the web. Computational models and empirical studies* (pp. 149–166). Springer.

Siebert, S. (1999). *Wortbildung und Grammatik. Syntaktische Restriktionen in der Struktur komplexer Wörter* (Linguitische Arbeiten 408). Tübingen: Niemeyer. doi: 10.1515/9783110915921

ten Hacken, P. (2009). Early generative approaches. In R. Lieber & P. Štekauer (Eds.), *The Oxford handbook of compounding* (Oxford Handbooks in Linguistics) (pp. 54–77). Oxford: Oxford University Press.

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach*. London/New York: Continuum.

Wurzel, W. U. (1998). On the development of incorporating structures in German. In R. M. Hogg & L. van Bergen (Eds.), *Historical linguistics 1995*, Vol. 2: *Germanic linguistics* (pp. 331–344). Amsterdam & Philadelphia: John Benjamins. doi: 10.1075/cilt.162.24wur

Zeldes, A. (2012). *Productivity in argument selection. From morphology to syntax* (Trends in Linguistics: Studies and Monographs 260). Berlin & Boston: Mouton De Gruyter. doi: 10.1515/9783110303919

## *Authors' addresses*

Livio Gaeta
Lingua e linguistica tedesca Dipartimento di Studi Umanistici
Università di Torino
via S. Ottavio 20
I-10124 Torino
Italy

livio.gaeta@unito.it

Amir Zeldes
Department of Linguistics
Georgetown University
Poulton Hall, Room 243
1421 37th St. NW, DC 20057
Washington

amir.zeldes@georgetown.edu