Huber, Elisabeth (2019) The Car Pet in the Carpet. On the Interaction of Computer-Linguistic Methodology and Manual Refinement in Researching Noun Compounds. Journal of Computer-Assisted Linguistic Research 3: 78-95.

The Car Pet in the Carpet. On the Interaction of Computer-Linguistic Methodology and Manual Refinement in Researching Noun Compounds

Elisabeth Huber
LMU, Germany
Corresponding author: huber.elisabeth@lmu.de

Received: 25 January 2019 / Accepted: 3 March 2019 / Published: 15 July 2019

Abstract
Why does *football* combine productively with further nouns to form more complex expressions like *football game*, whereas seemingly comparable compounds like *keyword* only infrequently expand to more complex sequences? This project explores why some two-noun compounds are more readily available for forming triconstituent constructions than others. I hypothesize that the productivity of a two-noun compound in the formation of triconstituent sequences depends on the degree of entrenchment of that two-noun compound, assuming that only compounds that are entrenched to a certain degree are productive in forming more complex constructions. In order to test this hypothesis, a list of three-noun compounds in the English language needed to be compiled. The obvious thing to do would be to search for sequences of three nouns in POS-tagged corpora. However, since such automatized searches on the one hand do not allow the recall of all required instances and, on the other hand, often create results that are not precise enough, this requires substantial manual screening. Furthermore, in order to operationalize the concepts of entrenchment and productivity, it was necessary to count the usage frequencies of noun constructions. For this work, as well, the automatic elicitation of the data needed to be complemented by further manual selection in order to obtain correct usage frequencies. Both the complex automatic and manual work processes in the elicitation of the data will be presented in detail to give an impression of the extent of such a project.

Keywords: compounds, complex words, nouns, entrenchment, productivity, computer-linguistics.

DOI: https://doi.org/10.4995/jclr.2019.11316
PUBLISHER: UNIVERSITAT POLITÈCNICA DE VALÈNCIA.

1. Introduction

The focus of this study are triconstituent English noun compounds, i.e. meaningful recurrent sequences of three nouns with a naming function (henceforth abbreviated as '3N'), e.g. *credit card number*, *trade union leader*, *day-care centre*, *candlelight dinner*, *football game* or *seafood salad*. In a larger scale project, I am examining why such complex constructions are used and produced. To provide an answer, the question that needs to be addressed first is: Which three-noun compounds exist in the English language? In theory, any sequence of three nouns could be an item of interest here. However, this project employs a usage-based approach and is thus interested in those three-noun sequences which seem to be lexicalized as compound lexemes and are used by native speakers of English. Researchers who have addressed this phenomenon have contented themselves with naming a few examples (e.g. Carstairs-McCarthy 2002, 76-84) as there is no inclusive list of three-noun compounds for the English language. Generalisations on the phenomenon of 3N-compounds, however, can only be drawn from a comprehensive collection, not from a hand-selected sample of compounds. In order to get an adequate impression of this phenomenon it is necessary to work with big data sets. Therefore, a list of such constructions had to be compiled. Clearly, creating a list as complete as possible requires the use of computer-linguistic work. Section 2.1 will explain how this task was approached with the help of a POS-tagged data set.

However, such automatized searches show a poor performance in terms of both precision and recall, which is why substantial manual screening is required. With regard to recall, there is one type of three-noun compound that is quite problematic. For this reason, in research on such complex compounds often only the type *trade union leader* is addressed, i.e. constructions where the three nouns are orthographically separated by spaces, as this subtype is easily found in corpora. The type represented by *football game* or *family network*, however, in which two of the nouns form an orthographic unit, is hard to detect because the combination of the two nouns is tagged as one noun in corpora: *football*$_N$ *game*$_N$. Searching for them in a corpus poses a major problem, as they cannot be identified by a search for sequences of three nouns. Still, this type of three-noun compound is indispensable in the research of 3N as it is a very common and highly interesting type and thus needs to be included. How can a script find compounds within words? Section 2.2 will give an insight into the methodological process used to identify such constructions.

With regard to precision, the main challenge is to exclude a very large number of unwanted false positives. Automatic searches for 3N-sequences also identify constructions that are similar to the intended sequences form-wise but do not belong to the category of compounds, e.g. sequences of three nouns such as *tablespoon oil*. How can rather random sequences of nouns be distinguished from actual compounds? Since frequency of recurrence is a good but by no means robust criterion, manual work has to accompany automatized searching for constructions. Section 2.3 will give a detailed presentation of the manual refinement needed to narrow down the results of the automatized search to actual compound lexemes.

Section 3 will then demonstrate how a theory from Cognitive Linguistics was operationalized with the help of usage frequencies extracted from the corpus. It will be argued that in order to

serve as productive bases for three-noun compounds, two-noun compounds ('2N') need to be entrenched to a certain degree. Here, as well, the computer-based methodology that is applied to extract the usage frequencies cannot distinguish which occurrences are needed for each word to the necessary extent, which is why an additional individual selection of frequencies is necessary (cf. section 3.4.2).

2. Compiling a list of three-noun compounds

The project is based on the *Corpus of Contemporary American English* (COCA, Davies 1990). The COCA is a 560 million word corpus that has been compiled since 1990, with an even distribution of different genres, and containing both written and spoken data. As the online search interface of the COCA does not allow searches as complex as required, two sets of downloadable material were used: an n-gram set and an offline version of the corpus[1], both available for purchase. The n-gram sets were used for the compilation of the list of three-noun compounds (cf. section 2), while the offline version of the COCA was used in the operationalization of linguistic concepts (cf. section 3).

2.1. Selecting noun-sequences from the n-gram set

In order to create a list of three-noun compounds, the obvious thing to do would be to search for sequences of three nouns in POS-tagged corpora. As the n-gram sets include part-of-speech specifications, they were chosen as source material for the compilation of the 3N-list. The n-gram package contains 2-, 3-, and 4-grams which occur at least 3 times in that sequence. The lists of interest were, obviously, the 3-gram list, but also that of 2-grams, as this is where the problematic type of *football game* was expected to be found. The list of 2-grams contains 6.2 million types; the 3-gram list is even more extensive with 11.9 million types. These lists were uploaded to the online data management software MySQL on the servers of the CIS of LMU Munich.

Both in the sets of 2-grams and 3-grams the noun sequences were selected with the help of the CQP query syntax. Then, only those sequences were chosen in which each of the constituents consisted of at least three letters. This procedure was used to refine the n-gram sets to noun sequences that were likely to be words of the English language. Before these noun sets could be further examined for actual compounds, the 2-gram list needed to be further reduced to only those items that actually contained three nouns and not just two.

2.2. Finding three-noun compounds in two-noun sequences

As explained above, a three-noun sequence like *football game* is tagged just like a two-noun compound such as *board game*, which is why it is not found within the trigram list but among the bigrams. Thus, a way needed to be found to automatically identify 3N-constructions within the supposed two-noun sequences from the bigram dataset. The idea was to write a script to determine whether one of the components of the bigrams contained two English nouns.

[1] Licensed for the Department of English at LMU Munich.

In order to achieve this, first a list of English nouns had to be compiled. A frequency list was compiled of all words that were tagged as nouns. This list also had to be refined thoroughly, as it contained a lot of noise in the form of sequences that could not be considered nouns or even words (e.g. items containing regular expressions or numbers, or complete non-words like *acxb*, etc.). Through systematic searches and deletions (e.g. searching for words without vowels or containing regular expressions) this list was reduced to 17,988 noun types.

In a second step, a Python-script was written that compares the strings of the components of each bigram to the items in the noun-list in order to determine whether a matching string can be identified. If so, that part of the component is identified as a noun and, as a consequence, is split off from the rest of the word by breaking the word at the end of that string. In the next step, the code searches the remainder of the word to ascertain whether this fragment, as well, is identical to any other string in the noun-list. Only those instances where this is the case, i.e., where the code can identify two nouns of the noun-list in either the first or second constituent of a noun sequence, are marked as potential three-noun compounds. The following depiction illustrates the code mechanism.

sports car	→	| sport$_N$ | s | car$_N$ |	→ 2 items from noun list, no hit
football game	→	| foot$_N$ | ball$_N$ | game$_N$ |	→ 3 items from noun list, potential hit

In the case of *sports car*, the code will approximate the first constituent and after processing the first five letters, i.e. *sport*, it will find a matching noun in the noun-list. It will mark *sport* as a noun and split it off from the rest of the word, i.e. the *-s*. For this remainder, as well, it will check for entries in the noun list. This search will be unsuccessful. For *car*, too, the code will not be able to split off a rest that is a noun. For this reason, *sports car* will not qualify as a potential three-noun compound either.

In the case of *football game*, the code will process *foot* and find a matching entry in the noun list. Splitting off the rest of the word results in the remainder *ball*. A comparison of this rest to the items in the noun-list will yield a positive result, which is why *ball* will be marked as a noun as well. In the processing of the second constituent, the code will identify *game* as a noun in the noun-list. The "remainder", however, again will not be a part of the noun list, which is why, in total, three nouns will be identified, and, as a consequence, the code will mark *football game* as a potential three-noun compound.

Although this approach seemed a very promising solution to the task of finding three-noun compounds in a two-noun set, it created problems. The most crucial one was that the code delivered an extremely high number of false positives when words that were originally monomorphemic were split into two alleged nouns, thus producing what I call 'fake compounds'. The following list is just a small sample of interesting cases where one of the constituents was split by the code because it contained two items of the noun-list, but is not a compound of the English language:

- *break age*
- *lab oratory*
- *nap kin*
- *stag nation*
- *can vases*
- *car ton*
- *ball ads*
- *don key*

- *sin king*
- *can teen*
- *medal lions*
- *leg end*
- *pal ace*
- *man dates*
- *champ ion*
- *pronoun cement*
- *car pet*
- *account ant*

Reading these erroneously split words as actual compounds is almost a delight, especially when thinking about the potential meanings (e.g. 'nation of stags', 'cement made out of pronouns'). These interesting results did change the author's awareness of morphology in English.

However, as interesting as these results might be, they had to be sorted out. These cases are not malfunctions of the code as they are a correct implementation of the instructions. Thus, refinements could not be made to the code, but had to be done to the source material, i.e. the noun-list, or the result-list of potential 3N. The decision in each of these fake compounds was either to delete the problematic noun from the noun-list or to leave it in and correct the mistakes it caused in the result-list. Here, the considerations were rather practically based on the balance of expected benefits and costs for precision and recall. If a noun from the noun-list was the cause of a high number of fake compounds, it was deleted from the noun-list, even though that meant compounds with that noun would be lost. Examples are words like *ion* or *ant*. Taking them off the noun-list prevented the detection of a high number of false positives, thus increasing precision; but as the amount of real compounds containing the word *ion* is intuitively considered to be rather low and costs in terms of missed recall were therefore low too, the loss of such potential compounds seemed justified in order to prevent noise in the form of fake compounds. In other cases, where one of the problematic nouns was categorized as a common component of a compound, like *age*, it could not be excluded from the noun-list, but the results were - if detected - deleted from the list of potential three-noun compounds.

With the help of this process, 22,457 different types of three-noun sequences could be detected within the set of bigrams. These, however, just as well as the three-noun sequences in the set of trigrams were not all instances of actual compounds. In order to narrow down the POS-based lists to the sequences of interest in this project, considerable refinement was needed, which will be presented in the following sections.

2.3. Refining the lists

Are all sequences of nouns in a corpus real compounds? Quite clearly there are instances within the formal sequence of three nouns that are not parts of utterances that would be found in actual speech production. In order to take into account the differences between rather random sequences of nouns and meaningful complex constructions, the computer-linguistic work presented in the previous section was accompanied by manual sighting of the data.

In a first step the data were sighted with the help of different sorting mechanisms (e.g. by frequency, by alphabet, by word-length, etc.) to recognize false positives. These were deleted manually and used to optimize the query. Searching for words consisting of more than 13 letters, for example, helped to identify email-addresses (e.g. *paul@olivetree.com*) and names of online websites (e.g. *www.myamericanartist.com*), the detection of which led to further searches for regular expressions, etc.

As sighting of the data still indicated a high number of non-words, for example *xnw*, systematic searches for instances containing numbers, a repetition of the same letter, less common letters of the English language, or sequences not containing vowels helped to eliminate these unwanted items. This kind of noise in the data proved to be pervasive. Although all of these strings of letters had been tagged as English nouns, this step identified and deleted more than 6,000 instances of non-words.

Even at a stage when the pool of items had been reduced to sequences of three nouns, these were not necessarily compounds. Especially names, appositions and quantifiers are not of the type targeted in this project. Quantifiers occurred especially in measurements such as *cup olive oil* or *teaspoon sugar*. Similar sequences were those containing the words *kilo*, *meter*, *bag*, *cup*, etc., e.g. *kilo steel items*. With regard to names (although proper names had already been excluded with the help of tags), there were many sequences where the first noun was a title or description like *president*, *leader* or *democrat* followed by a name. Likewise, the names of places were excluded. Sequences where the third noun was a word like *place*, *square*, *resort*, *valley*, *bay* or *hotel* were searched for and deleted in those cases where they denoted a location or hotel.

In these examples, it was comparatively easy to identity items that did not belong to the intended category. However, there were many instances in which the decision was much harder. In some sequences it was not immediately clear whether all constituents of a sequence were actually nouns. The first nouns in 3N-sequences can, for example, resemble adjectives, especially in the denotation of material, cf. *plastic bottle disposal*, *silver earring*, *gold medal winner*, or colours such as *blackwater area*. When there were similar constructions in which these constituents had an analogous meaning but were not adjectives (cf. e.g. *feather earring - silver earring*), these items were not deleted from the list.

Also, due to the high incidence of multiple class-membership of certain verbs and nouns in the English language, many components within compounds could be both nouns and verbs, e.g. *play* in *playground*. In these cases the Oxford English Dictionary (OED) was consulted and compounds in which a component was marked as a verb were excluded. When information on the word class of the compound constituents was not available, the constructions were paraphrased. When paraphrasing suggested that a constituent was rather a verb than a noun, the sequence was excluded from the list.

To understand how time-consuming the work can be for even one word, consider the word *back*, which can be a noun but also a verb, adjective or adverb. It occurs quite commonly in compounds, in which it is not always an easy task to tell whether it is an instantiation of a noun or not. Compare, for example, *backyard*, *background*, *cornerback*, *paperback*, *backbone*. In some cases the *back*-constituent can be quite clearly identified as a noun (e.g. in *paperback*). In cases like *backbone* or *backyard*, however, it could just as well be an adjective or adverb. Here again, the OED was helpful only in some cases. The decision whether to include or exclude a compound with a *back*-component was even harder when it was not clear if the other constituent in the compound was a noun or could just as easily be a verb, cf. e.g. *playback*, *feedback*, *takeback*, *flashback*, *drawback*, *pullback*. In these cases, as well, the OED was consulted. When the

compounds were not listed, those in which paraphrasing clearly identified the corresponding constituent as a verb were deleted.

Another problem involved components ending in *-ing*, such as *dancing girl bar* or *fund raising event*, as such constructions could be different in nature. Although tagged as three-noun sequences, not all of them actually contain three nouns. They can be a noun sequence modified by an adjective in participle form, as in *a moving love story*, or a verb in the gerund form followed by a direct object, as in *building playgrounds*. These are clearly not three-noun sequences and thus had to be excluded. Here again, there was no way of automatizing this selection. Individual decisions had to be made for each sequence as the constructions appear too similar form-wise (compare *building playgrounds* vs. *building security manager*).

These were only some of the cases in which the decision to include or delete a sequence had to be taken individually in order to do justice to the difference between forms that might go back to different functions. This section has outlined why the linguistic research in this project needed to be much more fine-grained than the selection that could be made by a strictly rule-based script. Relying on automatized searching would have resulted in the inclusion of items that do not fall in the category of three-noun compounds at all.

2.4. Lemmatization

Both the 2N-list and the 3N-list had been reduced to the intended noun sequences, with the 3N-list resulting in 41,750 different types of three-noun compounds and the 2-N list encompassing 21,450 types of 3N. These lists needed to be united to obtain an encompassing collection of 3N-compounds. The major difficulty here was to find and unite the tokens that belong to the same type, i.e. orthographic variants as well as singular and plural forms of the same type, and add up the frequencies.

The first serious hurdle was summarizing the different orthographic realizations of the same type under the most frequent form. For triconstituent compounds, the possible combinations of hyphens, spaces and solid spelling (i.e., written as one orthographic word) theoretically add up to a maximum of nine different orthographical realizations that need to be taken into account for each sequence[2]. In no case are there actually nine different spellings, but it is quite common for a three-noun sequence to appear in more than two different forms (cf., e.g., *credit card company*, *credit-card-company*, *credit-card company* and *creditcard company*). A script was written that searches for all possible spelling variants of each word and adds up the frequencies of the tokens for all types.

The second step was to combine singular and plural forms under the more frequent form and add up the frequencies so they are counted as tokens of the same type. This was tricky especially in cases where the plural was not formed regularly by adding an *-s* to the singular forms. The code had problems identifying plural forms correctly and needed a lot of refinement, for example, in order to match plural forms ending in *-ies* with the singular forms ending in *-y*, or recognizing the singular form of *busses* as *bus*, but that of *houses* as *house*.

[2] i.e., N-N-N, N N N, NN N, N-N N, N N-N, N NN, NN-N, N-NN, NNN; 'N' = Noun.

2.5. Result

The sequence of steps described so far produced a list of 57,753 different types of triconstituent noun constructions of the English language, manifested by 824,329 tokens. These enormous numbers indicate that the sequence Noun+Noun+Noun is much more frequent than claimed in previous literature and underlines the need to explore this kind of construction in more detail.

This list still contains noise, i.e. three-noun sequences that are outside the scope of interest of this project, as the majority of noise was detected through random inspection. This, however, is a by-product of working with big data sets.

The collection and analysis of 3N-compounds described in this section is not an end in itself, but stands in the service of pursuing the larger goal of understanding the representation and productivity of 3N-compounds. As we will see, this requires further corpus-based computational methods.

3. Is entrenchment a prerequisite for productivity?

I will now proceed to discuss the question of whether entrenchment is a condition for two-noun compounds to be productive in forming three-noun compounds. The first sections will explain the linguistic concepts in more detail. Section 3.4 will then describe the methodology used for the extraction of usage frequencies. The corpus work that was necessary to operationalize the concepts of entrenchment and productivity will be described in detail.

3.1. Research question and hypothesis

In a first analysis of the most frequent 3N-constructions it was quite striking that some two-noun compounds appeared repeatedly as embedded compounds. Words like *football*, *weekend* or *newspaper* were common bases for 3N-constructions. Other words like *world heritage* or *body mass* appeared less commonly in three-noun compounds. Based on this observation, the focus of the next part of the project was on finding the commonalities of the most productive two-noun compounds.

In word-formation, it is commonly agreed that complex formations are always based on the principle of binarity (cf. Schmid 2016, 96; Carstairs-McCarthy 2002, 78). That means that even formations that consist of more than two components have a hierarchical structure in which two components are always combined at a time (with the exception of synthetic compounds, cf. Schmid 2016, 134-136). Three-noun root compounds are thus always extensions of binary compounds, i.e. 3N = [[N+N]+N] or [N+[N+N]], but not N+N+N. This means that a 3N like *trade union leader* is based on the embedded compound *trade union*, which then, as a unit, combines with *leader*. For this reason, the solution to the question which three-noun compounds are formed lies in the step from the binary compound to the triconstituent one.

In theory, any two-noun compound can be combined with a third noun. However, 2N differ considerably regarding the extent to which they take third nouns. *Football*, for example, is quite commonly combined with third nouns, compare e.g. *football game*, *football player*, *football coach* or *football match*. *Body mass*, in comparison, seems to be much less productive in forming 3N, as it mainly combines with *index* and only a small number of other nouns. It is quite typical of

constructions to be only partially productive (cf. Goldberg 2016, 369). The question of interest is thus: which two-noun compounds are commonly extended for use with new words and which ones are not?

Based on the inspection of the data collected, I formed the impression that the productivity of a two-noun compound in forming three-noun compounds depends on the degree of entrenchment of that two-noun compound. The concepts of entrenchment and productivity will be explained in more detail in the following sections.

3.2. Entrenchment of two-noun compounds

Entrenchment is a cognitive concept that tries to account for the representation of linguistic constructions in the individual's mind. More precisely, it describes the degree to which the formation, storage and activation of a construction is automatized (cf. Langacker 1987; 2008; Schmid 2007; 2017). Words can be entrenched to a stronger or lesser degree, with the main factor of influence being usage frequency.

One type of entrenchment that is particularly interesting for this project is syntagmatic entrenchment (cf. Schmid 2017, 11). It describes the strength with which two words are associated. If two words are commonly used together, for example *foot* and *ball*, the syntagmatic entrenchment is strengthened, i.e. the words are associated with each other quite strongly and the sequence is processed in an automatized way with no online computation necessary anymore. Words that are not frequently used together, like *foot* and *doctor*, are weakly linked to each other and are thus rather likely to be computed online when they co-occur.

How can this concept be operationalized? Entrenchment cannot be measured directly. It is understood as a function of frequency of exposure to linguistic items, i.e. the more often a speaker is exposed to a word, the more strongly it will be represented in their mind. Consequently, it is usually approached through frequency measures in corpora, which provide an approximate estimate of how strongly entrenched an item is in comparison to others.

When studying the syntagmatic entrenchment of combinations, raw token frequencies are not suitable but for these kinds of links more complex statistical measures need to be employed (cf. Stefanowitsch and Flach 2017). There are different measures for the operationalization of association strength (e.g. pointwise mutual information); the reader is referred to Thanopolous et al. (2002) for a summary of the advantages and disadvantages of these measures. For the calculation of the syntagmatic entrenchment within 2N-constructions in this project the G-test, a loglikelihood-ratio statistical significance test, is employed. It calculates how many times a sequence of two words occurs (e.g. *football*), how many times the first word occurs in isolation (e.g. *foot* by itself), how many times the second word occurs without the first one (e.g. *ball* by itself), and sets this into relation to the size of the corpus. This way it gives an impression of the degree to which the co-occurrence of the two words is significant (cf. Stefanowitsch and Flach 2017, 115-119). The following formula was used to calculate the entrenchment values for the embedded 2N-compounds of the items in the study:

$$G = 2\sum_{i=1}^{m} N_i \cdot \ln\left(\frac{N_i}{n_{0i}}\right)$$

3.3. Productivity for forming tripartite constructions

The second variable to be operationalized was productivity. Productivity in this project is understood as the degree to which a two-noun compound is available for combination with further nouns to form three-noun compounds. This concept is thus not considered an either-or category, but assumed to be a continuum from more productive to less productive.

The graphs in Figure 1 give an impression of the concept of productivity for two-noun compounds, i.e. the degree to which they are used to form concrete instantiations of 3N. They display the types and tokens of two exemplary 2N-constructions for the nouns they combine with. On the horizontal axis the different types of three-noun compounds formed by that specific two-noun compound are listed, with the corresponding token frequencies for each three-noun compound on the vertical axis.

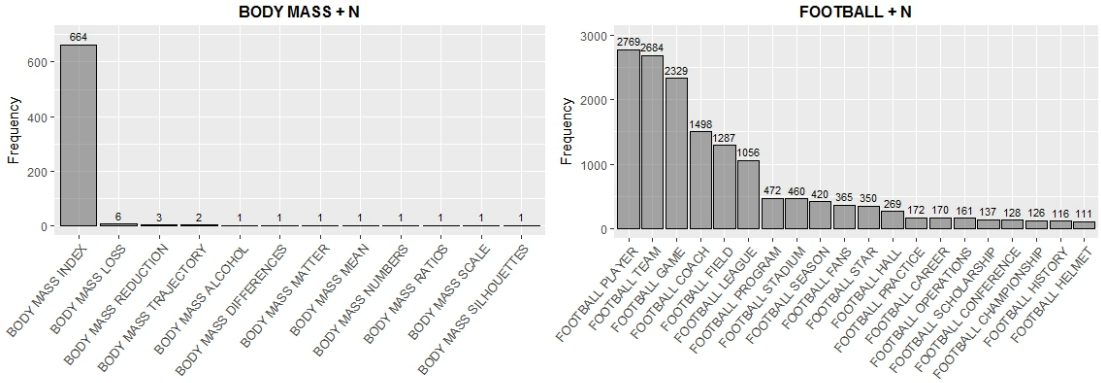

FIGURE 1. PRODUCTIVITY OF EXEMPLARY TWO-NOUN COMPOUNDS IN FORMING THREE-NOUN COMPOUNDS.

For 'body mass + N' all types and tokens are displayed in the table, whereas for 'football + N' only the first 20 types could be reproduced here for reasons of legibility. The two 2N-constructions show considerable differences in productivity: while *football* combines with a high number of different nouns with a strongly conventionalized paradigm of nouns for its third slot, *body mass* barely forms any triconstituent construction other than *body mass index*. The two-noun compound *football* is thus shown to be more productive in forming 3N than *body mass*.

Different measures have been discussed in the literature to operationalize the productivity of a word-formation process. Baayen (1993), for example, has established some corpus-based measures, using different components of productivity such as type frequency, token frequency and the occurrence of low-frequency instantiations, e.g., hapaxes. An attempt to measure productivity in terms of hapax legomena did not turn out effective with my data, as these low-frequency items mostly turned out to be noise. I agree with Baayen that both type and token frequency are relevant factors for productivity, which is why a combined measurement was chosen. The productivity of a given two-noun compound for forming three-noun compounds was operationalized as the product of types and tokens of the attested slot-fillers for the third

noun a 2N combines with, subtracting the most frequent type. This type is not included as productivity is a function of forming more than just the one conventionalized 3N.

This measure yields relative values for productivity that seem to reflect the actual degree to which a compound is available for the formation of triconstituent constructions. Strictly speaking, what is being measured is the past productivity of a pattern. However, frequent usage of a pattern entrenches that pattern in the minds of language users, which is why its past productivity has an influence on a pattern's present and future availability for combination. For this reason, it seems justified to use the existing types and tokens as measurement for the potential current productivity of a pattern.

3.4. Methodology

The question to be answered is whether two-noun compounds that are entrenched to a certain degree are more productive in forming three-noun compounds. The items for which both entrenchment and productivity were calculated were taken from the list of three-noun compounds compiled in section 2. Of this list, 200 constructions were chosen systematically through stratified sampling. This part of the project was limited to triconstituent constructions with a complex modifier, i.e. constructions of the structure [[NN] N] (e.g. *newspaper editor*), referred to as *left-branching* (Warren 1978, 11), as opposed to right-branching constructions, which have a complex head. This limitation was necessary in order to allow subsequent automatic processing. Schmid (2016, 208) compares the occurrence of these morphological structures for compounds within compounds and demonstrates that left-branching is, by far, more common, which is why a limitation to this type seemed justified at this point. For this sample, the entrenchment of the embedded two-noun constructions as well as their productivity in forming tri-partite sequences were to be measured in the ways described in the next section.

3.4.1. Acquiring usage frequencies

This section reveals the search details for the usage frequencies of the 200 test items which were extracted from the offline-version of the COCA with the help of a code.

The frequencies needed for the calculation of entrenchment include the following:

- frequency of noun1 (e.g. *foot*)
- frequency of noun2 (e.g. *ball*)
- frequency of bigram (noun1 + noun2, e.g. *football*)
- rest of corpus, i.e. all words that do not contain noun1, noun2, or the bigram

In order to operationalize the concept of productivity, the usage frequencies required from the corpus are the following:

- number of types of third nouns that a certain 2N combines with
- number of tokens of third nouns that a certain 2N combines with
- number of tokens of the most frequent 3N a certain 2N forms

The input for the code was the list of 200 selected 3N-compounds in singular and with separate spelling, i.e. the form N N N, e.g. foot ball game. This format was necessary to reduce the time for processing, since for this large data set computation is quite time-consuming. The code was run with Python 3.5 and accessed the 'text' directory as well as the 'word_lemma_PoS' directory from the offline version of the COCA corpus. The output file is a text document listing the required usage frequencies per three-noun sequence. Before the values given by the code could be used for the calculation of entrenchment and productivity, however, extensive manual refinement of the data was necessary due to different kinds of problems.

3.4.2. Problems and refinements

This section gives an insight into the problems of extracting usage frequencies by just presenting the difficulties that occurred in searching for the single words frequencies that were needed for the operationalization of entrenchment.

In counting the occurrences of a particular noun a script does not take into account potential erroneous tagging for word classes. The search for specific nouns can result in a high number of false positives when the items counted by the code are not instances of that specific noun. This is especially true for nouns in which the plural looks like a third person singular verb (e.g. *talks*, *cares*, *holds*, *changes*, *bombs*, *plans*, *states*, *notes*, *ends*, etc.). Although the code only searches for nouns, it is quite likely that a large number of words were incorrectly tagged as nouns in the corpus although they were verbs. This became quite clear in sequences like *Prime Minister talks* or *chess game ends*. However, by searching specifically for typical converted verbs, the majority of such cases could be deleted.

Furthermore, in an automatized search for words a code needs to either search for words or strings. When searching for the nouns of the compounds in isolation, the code is supposed to look for the single word frequencies, i.e. the occurrences of the first or second word by themselves. How many times, for example, does *basket* occur in the corpus? A code that is written to search for the items as actual words will search for them with blanks before and after and will thus not be able to find such words in more complex expressions. When looking for the word *basket*, for example, it will only find the occurrences of *basket* in isolation, ignoring the occurrences of *basket* in solid written compounds like *basketball*. The frequency of *basket* would thus not include its occurrences in compounds. The alternative is to have the code search for strings of letters instead of words. This approach was used in order to also find the items within compounds and affixations. This, in turn, creates the problem that the code will list all forms that contain a particular string as a potential instance of that string. For the word *record*, for example, the code also suggests counting the occurrences of *recording* or *recorder*, although these words are not desired as they are not instances of the noun in question. For this reason, all potential occurrences of a noun found by the code have to be checked manually. For some words, such as *entitlement*, all results found by the code as potential instances could easily be identified as clear instances of the noun: items found by the code were for example *entitlement*, *entitlement-conferring*, *self-entitlement*, *non-entitlement*, *anti-entitlement*, *nonentitlement*, *entitlements*, *health-entitlement*, *dis-entitlement*. All of the frequencies found could be counted as occurrences of the noun in question. This, however, was a minority of cases.

As the code searches for strings of letters instead of lemmas, there is a high number of false positives for words that happen to contain the string being sought, without having anything to do with the searched word. When searching for occurrences of the noun *line*, for example, the code yields an extreme number of results, more than 500 types, including words that are clearly not related to *line* but only happen to contain that string of letters, like *adrenaline, vaseline, holiness, manliness, feline, millinery, deadlines, discipline, linen, neighborliness, trampoline, liveliness, masculine, recline*. All of these incorrect instances had to be sorted out manually.

Even in cases where the instances found by the code actually contain the noun searched for, it can be hard to decide which compounds or pre- and suffixations that contain a certain noun should be accepted as belonging to this lexeme and thus constitute a factor for its entrenchment. To stay with the example of *line*, compare the following complex words in which the word *line* was found: *guideline, airline, gasoline, headline, outline, hotline, deadline*. Which of these compounds are transparent enough for the occurrence of that construction to contribute to the entrenchment of the word *line*? In these cases, individual decisions have to be made for all of the suggested occurrences of each word.

Words with rather abstract or vague meanings are also problematic. Consider, for example, the word *base*. Is the participle form *based* still related strongly enough to the lexeme *base* to contribute to its entrenchment? Are the occurrences of *base* in *database* or *basement* related closely enough semantically to the word *base* to strengthen its entrenchment?

In all of these cases, individual decisions had to be made for each of the potential occurrences of each of the two embedded noun-constituents of all 200 three-noun sequences from the sample.

3.5. Results

This section shows the results of the methodology presented in the previous section. First, the results for entrenchment will be presented, followed by the outcome of the productivity measures. Finally, the relation between these two variables will be examined.

3.5.1. Entrenchment of two-noun compounds

The entrenchment of two-noun compounds within triconstituent constructions was approximated through the G-Test. It delivered concrete values for each of the constructions, which were taken as a relative measure to compare the constructions with regard to their entrenchment, i.e. the strength with which these syntagmatic combinations of two nouns are represented in the mind of an average member of the English speech community. The scores as such are not interpretable, but the differences between them are.

The results for some samples of two-noun compounds are displayed in Table 2, ordered by the result of the G-Test, i.e. degree of entrenchment, in decreasing order. High values indicate a higher degree of entrenchment of the construction in the mind of an average language user, while lower values are a signal for a rather weak mental representation in this combination. To ensure visibility only a small sample is displayed, namely ten of the compounds which had high entrenchment values (items 1-10), followed by ten of the lowest ones (items 11-20).

	2N-COMPOUND	ENTRENCHMENT VALUE
1	*network*	464323
2	*healthcare*	379530
3	*newspaper*	342194
4	*weekend*	260239
5	*baseball*	242312
6	*classroom*	230054
7	*football*	210044
8	*bedroom*	152410
9	*credit card*	102104
10	*birthday*	100754
11	*health food*	1481
12	*money market*	1221
13	*air defense*	1181
14	*home loan*	1118
15	*career counselling*	1113
16	*teacher certification*	1014
17	*growth curve*	633
18	*wartime*	547
19	*energy storage*	517
20	*crisis pregnancy*	304

TABLE 1. ENTRENCHMENT VALUES FOR EXEMPLARY TWO-NOUN CONSTRUCTIONS IN TRI-PARTITE SEQUENCES.

Among the test items the complex words *network*, *health care* and *newspaper* can be expected to be quite strongly entrenched constructions in an average language user's mind, whereas sequences like *crisis pregnancy* or *energy storage* presumably have a rather weak mental representation in this combination.

3.5.2. Productivity of two-noun compounds

Table 2 displays the differences in productivity for exemplary two-noun compounds, more precisely some of the most productive items in the study (items 1-10), as well as some of the most unproductive ones (items 11-20). It can be seen that the test items vary greatly with regard to their availability for forming more complex constructions. Sequences like *healthcare*, *weekend* and *football*, for instance, are highly productive in forming tri-partite sequences, as opposed to compounds like *discourse analysis*, *newsmaker* or *work project* at the bottom end of the list, which have so far produced only few triconstituent compounds.

	2N-COMPOUND	PRODUCTIVITY VALUE
1	*healthcare*	20715742
2	*football*	13981815
3	*network*	9899315
4	*classroom*	8162253
5	*newspaper*	7722288
6	*weekend*	5582903
7	*railroad*	1979285
8	*bedroom*	1357953
9	*credit card*	1218130
10	*birthday*	972036
11	*population growth*	1833
12	*infant mortality*	1711
13	*fire protection*	1664
14	*internet service*	1064

15	*music festival*	980
16	*flu vaccine*	900
17	*industry research*	507
18	*growth curve*	333
19	*body mass*	308
20	*work project*	16

TABLE 2. PRODUCTIVITY VALUES FOR THE FORMATION OF TRICONSTITUENT CONSTRUCTIONS.

3.5.3. Relation between entrenchment and productivity

Figure 2 is a visual presentation of the relationship between the entrenchment and productivity of the test items, with logarithmic transformation of the data. As the data is not normally distributed, a linear regression model cannot be calculated. Purely descriptively, however, this is a strong indication that the variables tend to correlate: there is a tendency for more strongly entrenched compounds to be used more likely for further word-formation processes. Compounds with high values of entrenchment also feature a higher degree of productivity, whereas weakly entrenched compounds are not productive in forming triconstituent compounds.

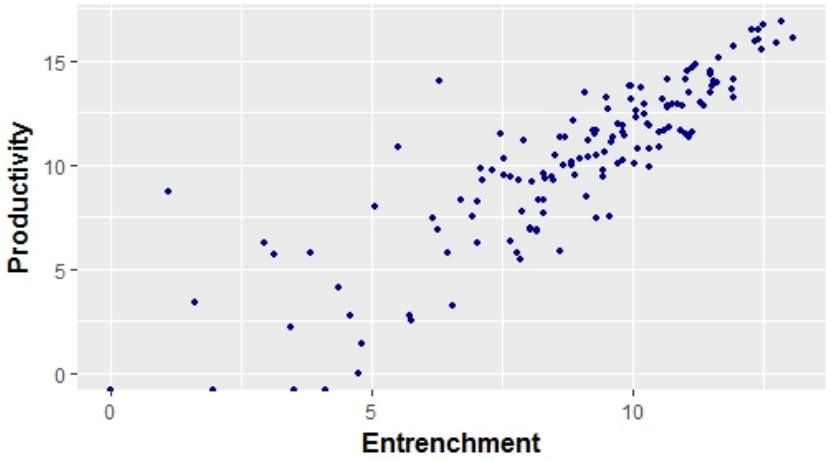

FIGURE 2. RELATION BETWEEN ENTRENCHMENT OF COMPOUNDS AND PRODUCTIVITY IN FORMING TRI-PARTITE SEQUENCES WITH LOGARITHMIC TRANSFORMATION OF THE DATA.

The next section will provide an explanation for this connection between the variables of entrenchment and productivity in the formation of three-noun compounds.

3.6. Explanation: Why are more strongly entrenched 2N more productive?

A possible explanation for the correlation between the entrenchment of a two-noun compound and its productivity in further word-formation processes can be found in the framework of entrenchment. The theoretical background presented before allows retracing the step that leads from a two-noun compound to a triconstituent construction: The syntagmatic associations for some two-noun compounds can be strengthened to such a degree that the constituents are not perceived as two separate, individual items anymore but, increasingly, as one chunk. This allows predictions about the cognitive effort needed to form more complex three-noun constructions based on a particular two-noun compound: once such a compound has reached a certain degree of entrenchment, the necessity of online processing is reduced to a minimum and

thus the cognitive effort of combining it with another noun can be assumed to be relatively low. A strongly entrenched two-noun compound would thus be more likely to be used in a more complex construction than a weakly entrenched compound, which should not be very likely to engage in further word-formation processes as it requires more cognitive work for assembling and then even for further combination.

4. Discussion

The study has shown that only compounds that are entrenched to a certain degree tend to be productively available for forming more complex compounds. This is considered an important result as it provides an insight into the nature and potential of word-formation. More precisely, it enables prediction of the products of word-formation processes and can answer the question of which compounds can serve as the basis for further, more complex constructions.

The relation found here is not true both ways. A productive 2N will always be entrenched, whereas the opposite is not necessarily true. There are 2N with a high degree of entrenchment which are still not productive in forming 3N, like *keyword*. The potential factors of influence that might be at work here still have to be explored.

In addition, entrenchment might not be the only condition for productivity as there are other factors that might have an influence, like length or semantics. It is, however, not within the scope of this project to explore this further. It might also be desirable to complement the results of this corpus study with psychological measurement.

Criticism can be levelled at the numbers used for calculation. Firstly, although a corpus is a sample of language use, there are admittedly potential differences between the language of an individual and the discourse within a community. Thus, all criticism regarding the operationalization of entrenchment through corpora (cf. Schmid 2010) also applies here. It is acknowledged that corpus frequencies can only be an approximation of the entrenchment of a certain structure in an average language user's mind. However, this method of operationalizing this linguistic concept seems to be the most efficient to date. Secondly, the concrete values for entrenchment and productivity are debatable because of the partly subjective and disputable decisions that had to be made, as described in section 3.

Tagging in corpora is, quite obviously, still prone to errors and requires much more training and precision to reduce noise in the data. Furthermore, semantic and morphological information needs to be added to corpora in order to enable researchers to conduct more precise searches. Despite manual revision of the data that supposedly only contained noun-sequences, the lists are quite likely to still contain a certain amount of noise that was not detected. This, however, is a common drawback of working with larger amounts of data.

Having said that, the aim of this project was to show that computer-linguistic methodology is clearly highly useful and indispensable in the implementation of linguistic projects working with big data on the one hand. On the other hand, however, such approach always needs to be complemented by manual refinement through a linguist, as scripts can (so far) only ever search for a category of constructions that share a certain form. This does not dispense with the

intuition of a linguist. All of the manual refinements of the computer-linguistic work and the individual single-word decisions that had to be made were indispensable in order to do justice to the semantic differences between similarly looking kinds of constructions.

References

Baayen, Harald. 1993. "On frequency, transparency and productivity." In *Yearbook of morphology*, edited by Geert Booij and Jaap van Marle, 181-208. Dordrecht, Boston, and London: Kluwer.

Blumenthal-Dramé, Alice. 2012. *Entrenchment in usage-based theories. What corpus data do and do not reveal about the mind*. Berlin: de Gruyter Mouton.

Carstairs-McCarthy, Andrew. 2002. *An introduction to English morphology. Words and their structure*. Edinburgh: Edinburgh University Press.

Davies, Marc. 1990-2011/2012. "The Corpus of Contemporary American English (COCA)". https://corpus.byu.edu/coca/.

Engelberg, Stefan, Henning Lobin, Kathrin Steyer, and Sascha Wolfer. 2018. *Wortschätze: Dynamik, Muster, Komplexität*. Berlin: De Gruyter.

Hilpert, Martin. 2018. "Wie viele Konstruktionen stecken in einem Wortbildungsmuster? Eine Problematisierung des Produktivitätsbegriffs aus konstruktionsgrammatischer Sicht." In *Wortschätze: Dynamik, Muster, Komplexität*, edited by Stefan Engelberg, Henning Lobin, Kathrin Steyer, and Sascha Wolfer, 91-106. Berlin: De Gruyter.

Langacker, Ronald. 1987. *Foundations of cognitive grammar. Vol. I: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langacker, Ronald. 2008. *Cognitive grammar*. *A basic introduction*. Oxford: OUP.

Schmid, Hans-Jörg. 2007. "Entrenchment, salience and basic levels." In *The Oxford Handbook of Cognitive Linguistics*, edited by Dirk Geeraerts and Hubert Cuyckens, 117-138. Oxford: OUP.

Schmid, Hans-Jörg. 2010 "Does frequency in text really instantiate entrenchment in the cognitive system?" In *Quantitative methods in cognitive semantics: Corpus-driven approaches*, edited by Dylan Glynn and Kerstin Fischer, 101-133. Berlin: Walter de Gruyter.

Schmid, Hans-Jörg. 2016. *English morphology and word-formation*. *An introduction*. Berlin: Schmidt.

Schmid, Hans-Jörg. 2017. *Entrenchment and the psychology of language learning. How we reorganize and adapt linguistic knowledge*. Boston: APA and Walter de Gruyter.

Stefanowitsch, Anatol, and Susanne Flach. 2017. "The corpus-based perspective on entrenchment." In *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge*, edited by Hans-Jörg Schmid, 101-127. Boston: APA and Walter de Gruyter.

Thanopoulos, Aristomenis, Nikos Fakotakis, and George K. Kokkinakis. 2002. "Comparative Evaluation of Collocation Extraction Metrics." *LREC*, 620-625.

Warren, Beatrice. 1978. *Semantic Patterns of Noun-Noun Compounds.* Göteburg: Acta Universitatis Gothoburgensis.