Originally published in: Koutsoukos, Nikos/Audring, Jenny/Masini, Francesca (Eds.): Morphological Variation: Synchrony and Diachrony. Online Proceedings of the Eleventh Mediterranean Morphology Meeting (MMM11) Nicosia, Cyprus, 22-25 June 2017. - Patras: Pasithee, Open Access Electronic Publications, 2018. Pp. 36-50.

Morphological variation: the case of productivity in German compound formation

Katrin Hein
Institute for the German Language
hein@ids-mannheim.de

Stefan Engelberg
Institute for the German Language
engelberg@ids-mannheim.de

1. Introduction

The development of very large corpora and their constant growth has changed our picture of the lexicon considerably. The empirical turn in linguistics that is driven by corpus-based methods enables us to uncover the dynamic nature of the lexicon, i.e., the processes of constant lexical change and the mechanisms that promote this change. Three features characterize the "dynamic lexicon" in particular (cf. Engelberg 2015a):

(i) ***Size***: Even a considerably small corpus of German with about a quarter billion running words contains almost 2 million different lexemes (Evert and Baroni 2005). Although it is difficult to extrapolate these numbers to very large corpora, we can expect at least more than 10 million lexemes in large corpora like the *Deutsches Referenzkorpus* (DeReKo) (cf. Institut für Deutsche Sprache 2017) with more than 30 billion running words. Large dictionaries of contemporary German consist of about 200.000 to 300.000 lemmata. Therefore, only a very small proportion of the lexemes occurring in corpora is lexicographically described.

(ii) ***Patterns***: In contrast to theories that conceptualize language as being based on lexical entities and rules that manipulate these entities, corpus-based research gives rise to a more pattern-based organization of language. In particular, in domains like idioms, argument structure, or complex words, semi-abstract and semi-regular linguistic patterns account for the variation and productivity observed.

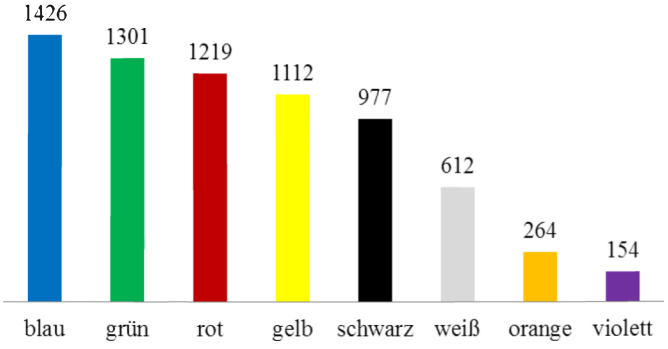
(iii) ***Distribution***: The quantitative distribution of entities in corpora allows us to reconstruct the nature of the dynamic processes in the lexicon. This comprises the changing frequencies of lexical items over time, the nature of Zipfian distributions, and the productivity of linguistic patterns.

From the perspective of lexicological theory as well as from the perspective of lexicographic language documentation, the question arises how the lexical wealth found in corpora can be adequately described and explained (cf. Engelberg 2014). Since compound formation is a dominant factor for the expansion of the German lexicon, the investigation of tendencies and idiosyncrasies in compound formation must play an important role in the investigation of the dynamic lexicon. The paper at hand discusses productivity in German compound formation. In a general way, we understand productivity as "the ease with which a linguistic process gives rise to new forms" (O'Donnell 2015: 3). We will look at compound formation from a lexeme-based synchronic perspective as a case of morphological variation. In particular, we focus on groups of compounds with semantically closely related head words, e.g., compounds with color words as heads. Our approach is characterized by a qualitative as well as a quantitative perspective on productivity. Taking the properties of the head lexeme as a starting point and applying corpus-based statistical methods, we try to gain new insights into

compound formation, especially into potential factors which govern their productivity. The approach presented here is one of the first attempts to apply the concept of productivity, which has been predominantly used in the domain of derivation, to compounding.

Our investigation starts with the observation that even semantically very similar words (e.g., *Angst* 'fear' vs. *Furcht* 'dread') or words within a semantic field (e.g., color words like *blau* 'blue' and *weiß* 'white') show strikingly different tendencies with respect to their occurrence as heads in compounds (cf. Fleischer and Barz 2012: 81f., 135). This observation is illustrated in Figure 1, which shows the simple type frequencies for German compounds whose head is a basic color word (cf. Engelberg 2015b).[1] It can be seen that, for example, *blau* 'blue' (as in *abendblau* 'evening-blue', *abgasblau* 'exhaust-blue', *acapulcoblau* 'acapulco-blue', etc.) gives rise to many more compounds than *weiß* 'white' (as in *alabasterweiß* 'alabaster-white', *albinoweiß* 'albino-white', *alaskaweiß* 'alaska-white', etc.).

Figure 1: Compounds with color words: Type count (Realized Productivity) (Engelberg 2015b)



Two questions guide our investigation:

(i) How can we measure the productivity of simplex words with respect to compound formation?
(ii) How can differences in compound productivity be explained? What are the principles that govern this variation?

2. Morphological Productivity

As "morphological productivity is one of the most contested areas in the study of word-formation" (Bauer 2001: i), this concept cannot be discussed here in full detail. We will sketch some qualitative and quantitative aspects of morphological productivity and its applicability to compounding (cf. Section 2.1). Our paper focuses on the question how empirically observable differences in compound productivity can be explained; in Section 2.2, we will discuss potential factors for productivity.

[1] The investigation of color compounds is based on a part of the German Reference Corpus (DeReKo) with a size of 5.405.723.269 running words. All words that ended in one of the ten basic color words and the respective inflectional forms of these words were extracted and stored with their token frequencies. The ten color words can be used both as adjectives and nouns.

2.1 Productivity in compound formation

Productivity in compound formation is a rather unexplored field of morphology. While it is beyond question that compounding in general is a productive process of German word formation (Olsen 2015: 364 f.), it is quite surprising that the productivity of compounding has not been investigated in more (empirical) depth, but cf. Tarasova (2013) and Roth (2014). While Roth focuses on the competition between collocations and compounds, Tarasova is interested in the productivity of compound constituents and, in particular, in the question "whether the productivity of a compound constituent on the morphological level coincides with the productivity of the semantic relation realized in the constituent family" (Tarasova 2013: iii).

Until now, the notion 'morphological productivity' has been predominantly applied to the domain of derivation (cf. Bauer 2005); cf. the (methodically similar) investigations of Gaeta and Ricca (2006, 2015) for Italian or Scherer's (2005) and Hartmann's (2016) diachronic operationalization of current productivity measures for German derivations. In what follows, we will demonstrate the fruitful applicability of the concept of morphological productivity to the domain of composition.

A question that is crucial in this context is: What does 'productive' mean? If we keep in mind that Aronoff (1976: 35) considered productivity to be "one of the central mysteries of derivational morphology", this is far from being a trivial question (cf. Bauer 2001, 2005 and Plag 1999 for a more detailed discussion). The complexity of the concept of productivity becomes evident when one looks at the six readings of productivity proposed by Rainer (Rainer 1987: 188–90, quoted from Gaeta and Ricca 2015: 843).

6 possible readings of the productivity of WFRs (word formation rules):

(i) the number of words formed with a certain WFR;
(ii) the number of new words coined with a certain WFR in a given time span;
(iii) the possibility of coining new words with a certain WFR;
(iv) the probability of coining new words with a certain WFR;
(v) the number of possible (or generatable by rule) words formed with a certain WFR;
(vi) the relation between occurring and possible words formed with a certain WFR.

Similarly, Barðdal (2008: xi) "found that not only there were different *definitions* of productivity figuring in the literature, but also that there were different *concepts* of productivity around". Correspondingly, she identifies 19 senses of "productive", more precisely adjectives that are used as synonyms for "productive" in the literature, e.g., "frequent", "rule-based", "having a wide coverage", "easily combinable", "occurring or existing", etc. (Barðdal 2008: 10 f.).

It is important to highlight that those synonyms – as well as the different readings proposed by Rainer – clearly display that productivity is in the tension between 'availability' and 'profitability', i.e., between the theoretical possibility of new coinages and the exploitation of this potential. Moreover, productivity can be considered a qualitative or a quantitative phenomenon (cf. Scherer 2005; Rainer 1987; Plag 1999: 11–35).

Our lexeme-based investigation of compounding in German proceeds from the following understanding of productivity: First, we perceive productivity as a gradual phenomenon. This means that we do not only differentiate between the two poles 'productive' vs. 'non-productive'. Second, productivity is considered to be a quantitative phenomenon (cf. Roth 2014: 167). The advantage of this view has already been formulated by Gaeta and Ricca

(2015: 484): "Different facets of this complex phenomenon may be reflected quantitatively by different statistical measures". Consequently, "statistical work on large corpora has contributed decisively to a deeper understanding of the notion of productivity and the disentanglement of its diverse components" (Gaeta and Ricca 2015: 848).

2.2 Measuring productivity

We compute the different types of productivity of compounds on the basis of current productivity measures (cf. Baayen 1992, 1993, 2001, 2009) and data from a large corpus of German (*Deutsches Referenzkorpus*, DeReKo). The three now almost classical productivity measures from Baayen (2009) are given below:

(i) Realized Productivity: V (C, N)
The number of different types V belonging to a word formation pattern C in a corpus of N running words.

(ii) Expanding productivity: V(1, C, N) / V(1, N)
The number of different types V with a frequency of 1 belonging to a word formation pattern C in a corpus of N running words divided by the number of all types in the corpus with the frequency of 1.

(iii) Potential productivity: V(1, C, N) / N(C)
The number of different types V with a frequency of 1 belonging to a word formation pattern C in a corpus of N running words divided by the number of all tokens in the corpus belonging to word-formation pattern C.

Applied to patterns of compounds ending in one of the color words ***blau*** 'blue', ***gelb*** 'yellow', ***grün*** 'green', ***orange*** 'orange', ***rot*** 'red', ***schwarz*** 'black', ***violett*** 'violet/purple', and ***weiß*** 'white', the three measures yield the results shown in Figures 1 to 3, based on the numbers shown in Table 1 in a part of the German Reference Corpus with a size of 5.405.723.269 running words.

Table 1: Frequencies of hapax legomena and tokens for compounds with color words

head word	hapax legomena	compound tokens
blau	767	40.884
gelb	630	31.396
grün	649	42.962
orange	131	3.646
rot	624	51.159
schwarz	557	23.883
violett	74	5.344
weiß	257	33.628

The measure of Realized Productivity (Figure 1, Section 1) shows a dominance of compound patterns formed on the basis of monosyllabic, inherited color words (in contrast to loanwords) referring to primary colors (plus *green*). However, the measure only counts instances of the pattern formed in the past. It does not give an idea of the current productivity, i.e., of the number of compounds we can expect in the near future. This idea is better captured by the measure of Expanding Productivity (Figure 2) that considers the numbers of hapax legomena, i.e., the number of words that occur only once in a certain corpus. Although not every hapax is necessarily a new word, every new word in the language necessarily starts with the

frequency of one.[2] Thus, measures taking the number of hapaxes into consideration might be a good approximation to 'newness' in the lexicon. However, the fact that the number of hapaxes is the decisive factor in determining the measure of Expanding Productivity is often seen as a shortcoming of Baayen's measures. While some of those problems can be rejected by following the argumentation of Gaeta and Ricca (2015: 847), a more practical one remains: the hapax dependency requires manually checked data: "For hapaxes to be a reliable tool, however, it is necessary that corpus data are carefully and time-consumingly checked by manual inspection: a fully automatic listing of items associated with a given ending in a corpus would indeed produce huge distortions" (Gaeta and Ricca 2015: 847)

Figure 2: Compounds with color words: Expanding Productivity (simplified[3]) (Engelberg 2015b)

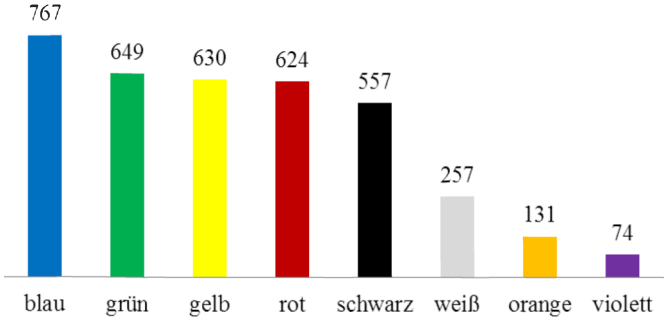
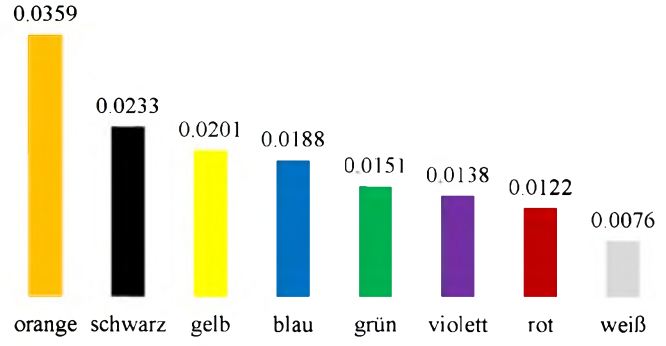


Figure 3: Compounds with color words: Potential Productivity (Engelberg 2015b)



In our example, the differences in results for Realized Productivity and Expanding Productivity are rather slight, indicating that the productivity of color compounds is not currently changing a lot. The third measure, Potential Productivity, however, yields very

[2] The chance that a hapax legomenon is indeed the first occurrence of a word depends of course on many factors like the size of the corpus, its textual composition, and its temporal resolution.

[3] Baayen's measure is simplified here. Since the number of all hapaxes in the corpus could not be computed, the mere number of hapaxes for each pattern is given. However, since the number of all hapaxes in the corpus is constant within the investigation on color compounds, the eight patterns can still be compared among each other.

different results. It is supposed to capture the degree of saturation of a word formation pattern (Baayen 2009: 902). As Figure 3 shows, in this sense of productivity, *orange* is by far the most productive color word. One might argue that since it has only produced a small number of types yet, it is indeed a less saturated pattern. However, the measure of Potential Productivity has been criticized for its token dependency: it relates the number of hapaxes of a category to the total number of tokens of that category. Gaeta and Ricca (2006: 62) argue that "the ratio h/N [hapax legomena/tokens] does not seem to give meaningful results if, in a given corpus, one compares the results obtained for affixes with very different token frequencies". Comparing categories with varying token size is problematic, because the lower the number of tokens, the higher is the value for productivity (P), as can be seen with *orange* in Figure 3. The reason for this is that for low numbers of tokens, the numbers of types increase more than with high token numbers. Consequently, the value P should only be calculated for categories with an identical or very similar number of tokens. Otherwise, an extrapolation of token numbers would become necessary. This procedure, however, also has problems when applied to actual token frequencies that are too distinct from each other (cf. Roth 2014: 169 f.).

Of course, the three measures proposed by Baayen do not exhaust the possibilities of the operationalization of different concepts of productivity. Other classical measures like the type-token-ratio can be applied to determine the lexical diversity of a category. As highly lexicalized types can distort the results, it can be revealing to know if a certain group of compounds (i.e., compounds with the head word *gelb* 'yellow') is dominated by a small group of lexicalized coinages or displays a high number of different types. Apart from that, measures of the productivity of compounds – in contrast to derivational morphology – should probably take into consideration the frequency of the head of the compound in its use as a simplex. We will not attempt to discuss these possibilities in this short article; we still aim at a deeper theoretical understanding of the different measures in terms of what facets of productivity they exactly capture.

3. How can differences in compound productivity be explained

Our brief look at the productivity of color compounds in the last section has not only shown how strongly the concept of productivity changes with its quantitative operationalizations, but it has also provided some first ideas which linguistic factors might influence productivity. Hypotheses emerging from the results in Figures 1 and 2 might be that monosyllabic headwords might be more productive than polysyllabic ones, that inherited headwords might be more productive than borrowed ones, that color words referring to primary colors might be more productive than color words referring to secondary and tertiary colors, etc. Even more interesting are tendencies that do not give rise to straightforward hypotheses. Under all measures we have tested so far, *weiß* 'white' is always less productive than *schwarz* 'black'. A central aim of our project, therefore, is to empirically carve out factors that determine the productivity of compound formation, or in other words: to empirically determine factors that govern the productivity of simplex words with regard to the formation of compounds. For this purpose, potential factors for productivity have to be outlined in a first step (Section 3.1). Subsequently, empirical evidence for these factors is determined on the basis of some pilot studies (Section 3.2).

3.1 Potential factors

We assume that productivity in compound formation might be influenced among others by the following factors. In this context, not only are the properties of the simplex in focus, but

also the properties of the unit with which the simplex is combined have to be taken into account.

(i) Morpho-phonological properties of the immediate constituents
- Syllable structure
- Properties of adjacent phonemes at the link between constituents

(ii) Morpho-syntactic properties of the immediate constituents
- Part of speech (For example, the composition of two nouns is considered to be the most productive type of composition, cf. Fleischer and Barz 2012: 81)
- Morphological complexity of constituents (While this factor influences the productivity of base words in derivation, Fleischer and Barz (2012: 81) call into question whether an increasing morphological complexity automatically is connected with a lower activity in compounding.)
- Valence properties of the head constituent, cf. Gaeta and Zeldes (2012): They investigated whether there is a strong correspondence between synthetic compounds and corresponding object-verb pairings; however, a statistically significant correlation could not be found.
- Position of constituents within the complex word: For example, Tarasova (2013: iii; cf. Fleischer and Barz 2012: 135 f.) demonstrates empirically "that a constituent is more productive in just one of the positions (modifier or head)"

(iii) Compound type (e.g., determinative compound vs. copulative compound; the former is considered to be more productive than the latter)

(iv) Semantic properties of the immediate constituents
- Meaning / semantic field
- Polysemy (According to Fleischer and Barz (2012: 82), the main reading of polysemous words is the most active with regard to word formation: For example, monomorphematic color words like *rot* ('red') or *grün* ('green') form only a few complex words in which a different reading than the reading 'color' is instantiated.)
- Aspects of taxonomy (Basic level categories in taxonomies like *mammal – dog – poodle* might be particularly productive.)
- Semantic proximity

(v) Semantic patterns of compounding
- The semantic relation between the constituents (cf. ten Hacken 2016; Hein 2015: 218–38) (In addition to computing productivity values for categories defined via the lexeme in head position, we also want to compute productivity values for semantic patterns of compounding, e.g., in color-compounds patterns like 'intensifying color compound' (*knallgelb* 'bang-yellow') versus 'comparative color compound' (*zitronengelb* 'lemon-yellow') versus color-color compound (*blaugelb* 'blue-yellow').)

(vi) Textual factors (genre, register)

(vii) Frequency and extra-linguistic relevance (This applies in particular to the simplex in head position. For example, central perception adjectives for the description of taste, like German *süß* ('sweet') or *sauer* ('sour'), show a higher activity in word formation than more peripheral adjectives like *herb* ('bitter/tart/harsh') (Fleischer and Barz 2012: 82).)

In the three pilot studies that we have conducted so far (Engelberg 2015; Hein 2016; Schneider 2016), we mainly concentrated on the evaluation of the two factors 'semantic proximity' and 'frequency of the head noun'.

3.2 Pilot studies

For all three studies, the following approach has been adopted: In the first step, we determined the productivity of the compounds with the help of different productivity measures (cf. Baayen 2009, 1992) on the basis of large corpora. In this context, we focused on groups of compounds with head words that are semantically similar or had a similar frequency as a simplex respectively.[4] In the second step, we tried to interpret and to explain the differences in productivity.[5]

3.2.1 Factor 'semantic properties of the head constituent'

We conducted two studies in which we investigated the influence of the factor 'semantic similarity' on productivity. In both cases, a part of the German Reference Corpus constituted the empirical basis. The question whether similar semantic properties of the head lexemes lead to comparable productivity values with regard to compound formation, was crucial in this context.

On the one hand, we studied compounds with a monomorphematic color word (e.g., *gelb* 'yellow') as head word, e.g., *neontextmarkergelb* 'neon-highlighter-yellow' as described in Section 2.2. On the other hand, we investigated compounds with a monomorphematic expression of a negative emotion independent of the position of the emotion word within the compound (Schneider 2016). Two pairs of semantically similar German words have been considered: *Angst* ('fear') vs. ***Furcht*** ('dread') and ***Wut*** ('anger') vs. *Zorn* ('wrath'). Moreover, we also included the nouns *Scham* ('shame') and *Hass* ('hatred').

Table 2: Frequencies of hapax legomena among compounds, compound tokens, and occurrence as a simplex for emotion words

head word	hapax legomena	compound tokens	occurrence as simplex
Angst 'fear'	2.842	141.001	748.975
Hass 'hatred'	1.276	70.219	86.730
Wut 'anger'	1.764	53.794	93.051
Furcht 'dread'	377	23.984	64.822
Zorn 'wrath'	545	20.393	74.564
Scham 'shame'	556	16.583	23.224

Regarding the relevance of the factor 'semantic proximity', both studies clearly indicate that semantic proximity between simplex words does not automatically lead to comparable productivity values with regard to the formation of compounds. As was foreshadowed in the introduction and in Section 2.2, color words like *weiß* 'white' versus *schwarz* 'black' show strikingly different tendencies to occur as a head word in compounds. The same holds for the

[4] In all groups of compounds that we focused on, we tried to control for general morpho-syntactic factors. For example, only simplex words have been considered.

[5] One of the main points of criticism in Baayen's approach, the problem of comparing productivity values for categories with a different number of tokens (cf. Section 2.2) holds at present for all three pilot studies.

semantically quite homogenous group of compounds that express emotions: The plot for their Realized Producitivity (Figure 4) shows clear differences a) between the six considered simplex words and b) within the two pairs of semantically very similar head words:

(i) *Angst* ('fear') is 7.7 times more productive in the formation of compounds than *Furcht* ('dread').
(ii) *Wut* ('anger') is 3.3 times more productive as a constituent in compounds than *Zorn* ('wrath').

Figure 4: Compounds with an expression of an emotion: Realized productivity (Schneider 2016)

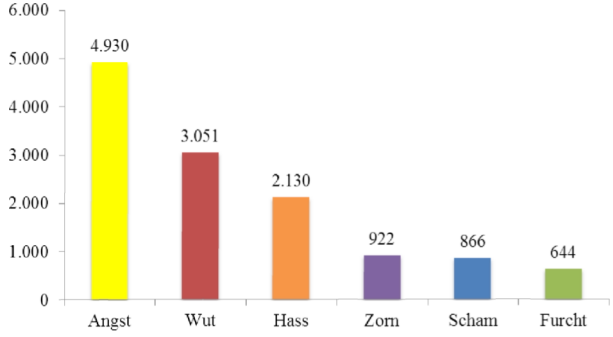


Figure 5: Compounds with an expression of an emotion: Type-Token Ratio (Schneider 2016)[6]

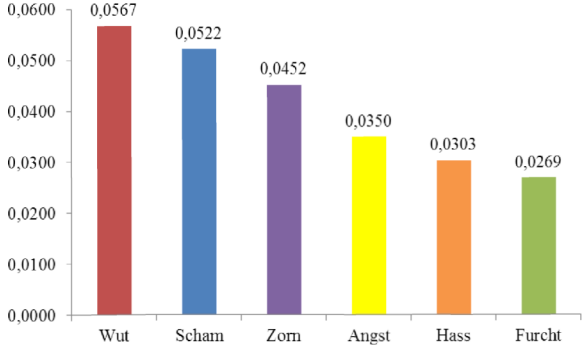


Figure 5 represents the ratio between the number of types and the number of tokens (TTR). Compared to the measure of Realized Productivity, it yields very different results.

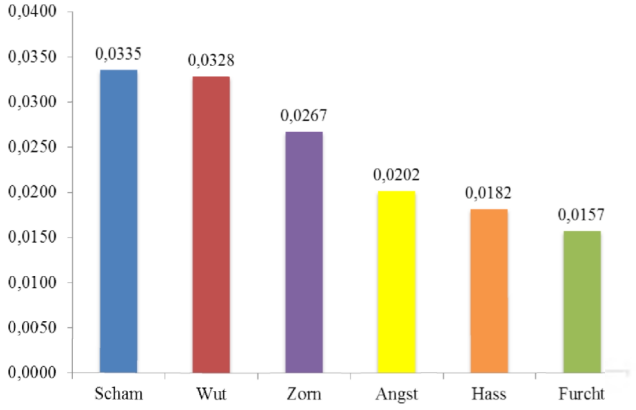
According to this measure, in contrast to the measure of Realized Productivity, *Angst* displays a rather low value, while *Wut* and *Scham* figure as the most productive expressions in the sense of providing the highest lexical diversity. Moreover, in Figure 5, the productivity

[6] It should be noted that the values for *Hass* are slightly distorted by the considerably high frequency of the proper name *Hassmann* and that the frequently occurring compound *Ehrfurcht* ('awe/reverence'), which recursively enters processes of compound formation, influences the results for *Furcht* (cf. Schneider 2016: 11).

values for the two considered pairs of semantically similar words are closer than in Figure 4 (*Angst*: TTR=0,0350 vs. *Furcht*: TTR=0,0269; *Wut*: TTR=0,0567 vs. *Zorn*: TTR=0,0452).

The results for Potential Productivity are plotted in Figure 6. On the one hand, just as in the case of color compounds (cf. Section 2.2), this measure yields very different results compared to the values of Realized Productivity (cf. Figure 4). In this reading of productivity, *Scham* ('shame') is the most productive simplex with regard to compound formation. While *Angst* ('fear') is ranked as the most productive word according to Realized Productivity, it displays a very low Potential Productivity. It should be noted that the results for Potential Productivity strongly resemble the results for TTR (cf. Schneider 2016: 23).

Figure 6: Compounds with an expression of an emotion: Potential productivity (Schneider 2016)



Instead of a clear connection between the semantic proximity of simplex words and their productivity in compound formation, both studies point at other potential connections: Semantic proximity seems to lead to comparable patterns of compounding. This holds for both studies in which the role of semantic proximity between simplex words as constituents in compounds was explicitly evaluated: The color compounds as well as the emotion compounds are dominated by a specific limited set of semantic patterns. For example, the compounds ending in *gelb* ('yellow') indicate that there are three patterns which seem to be characteristic for color compounds:

(1) Color-color compounds
rotgelb 'red-yellow'
bläulichgelb 'bluish-yellow'
rotweißschwarzgelb 'red-white-black-yellow'

(2) Intensifier compounds (intensity / tonality / shading)
knallgelb 'bang-yellow'
schrillgelb 'acute-yellow'
schreigelb 'screaming-yellow'

(3) Comparative compounds (comparison with the color of an object)
zitronengelb 'lemon-yellow'
saharagelb 'Sahara-yellow'
erdnussgelb 'peanut-yellow'

Our current work concentrates on computing productivity values for semantic patterns of this kind. This means that the category C (in Baayen's measures) is no longer defined via the lexeme in head position, but via the semantic pattern that is instantiated in the coinages within a certain group of compounds.

While semantic proximity between the head words probably leads to comparable patterns of compounding but not to comparable productivity values, the latter seem to be influenced by another factor: the frequency of a simplex in isolation. In other words, rather than assuming a connection between productivity and semantic properties, there seems to be one between the frequency of a simplex in isolation and its productivity in compound formation. It is evident from the token numbers in Table 2 and the Realized Productivity plotted in Figure 4 that the simplex with the highest frequency (in isolation), *Angst*, also produces the highest number of compound types – in this case, this not only holds for the number of compound types, but also for the number of compound tokens. The Realized Productivity values of compounds with *Wut* ('anger') and *Hass* ('hatred') confirm this observation: *Wut* and *Hass* are frequent simplex words in our investigation (ranks 2 and 3 in the frequency ranking) and also form the second highest, respectively third highest number of compound types. Nevertheless, there is no *clear* correlation between the number of simplex tokens and the number of corresponding compound types. For example, *Scham* and *Zorn* differ clearly in their occurence as simplex words but show approximately the same number of compound types.

With respect to the other productivity values, the assumed connection between the frequency of a simplex and its productivity in compound formation seems weaker: According to Potential Productivity and Type-Token-Ratio, the most frequent simplex word of our investigation, *Angst* 'fear', is one of the least productive simplex with regard to compound formation; the other way around, the least frequent simplex, *Scham* 'shame', turns out as the most productive simplex with regard to compound formation according to Potential Productivity (cf. the afore mentioned opposite results for Realized and Potential Productivity). However, Figure 6 also displays results pointing in the same direction as for Realized Productivity: The simplex *Wut* ('anger') is the second most frequent simplex of the investigation and is also the second most productive word with regard to the formation of compounds.

3.2.2 Factor 'frequency of the head constituent (in isolation)'

The influence of the factor 'frequency of a simplex' on its productivity in compound formation has been investigated in a separate study (Hein 2016). For this purpose, we have analysed binary compounds ending in simplex words from three different frequency layers:

(i) **Low** (e.g., *Ermächtigung* 'authorization'): more than 10, less than 50 occurences in our corpus[7]; extraction of 20 word-forms (by random sampling). Note that this definition of 'low' makes only sense in the context of the current study: If one considers the extreme Zipf-like distribution of word frequencies, 50 occurences have to be considered as a relative high frequency. However, for the purpose of this study it would not have been

[7] For the investigation at hand, we compiled a subcorpus consisting of 5.000 texts (9,25 million tokens) from our IDS corpora; we also included oral language (cf. DGD 2017). This, as well as the extraction of the compounds was done by our colleague Sascha Wolfer.

constructive to select only nouns with a frequency of 1 or 2 because it can be expected that such nouns produce only a very low number of compounds if any.

(ii) **Middle** (e.g., *Finger* 'finger'): more than 1.000, less than 2.000 occurences in our corpus; extraction of 20 word-forms (by random sampling).

(iii) **High** (e.g., *Kopf* 'head'): extraction of the 20 most frequent word forms from our corpus.

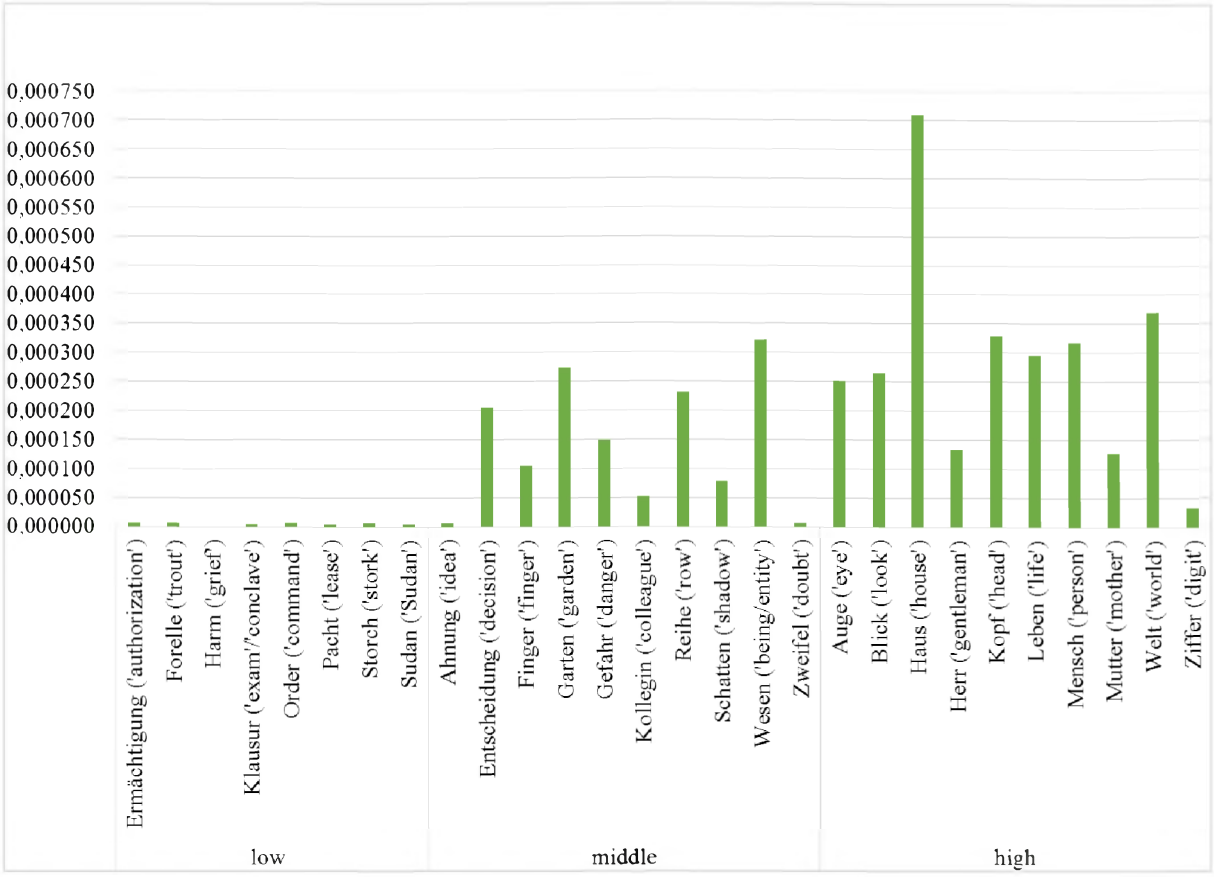
Out of this list of 60 word forms, we selected 28 lexemes by trying to consider a wide variety of different words, e.g., abstract vs. concrete nouns (e.g., *Ahnung* 'idea' vs. *Auge* 'eye') or derived (in a linguistic sence) vs. non-derived nouns (e.g., *Entscheidung* 'decision' vs. *Mensch* 'human'). In a next step, we extracted the corresponding compounds automatically, more precisely all compounds whose head word is formed by one of these 28 simplex words. The complete list of simplexes is displayed in Figure 7.

At first sight, the results seem to indicate that the parameter 'frequency of a simplex' influences the productivity in compound formation: Words that are more frequently used as a simplex are more productive in compound formation than infrequent simplex words. This is an expected finding: What is infrequent in isolation is not likely to be semantically modified by a non-head within a compound.

The connection between the frequency of a simplex and its productivity in compound formation becomes evident when one looks at the plot for Expanding Productivity in Figure 7. Expanding Productivity is supposed to give an answer to the question whether a morphological category is attracting new members, i.e., it tells us something about the near future. Expanding Productivity is the quotient of the number of hapaxes of a category C and the total number of hapaxes in a given corpus (cf. Section 2.2). The plot in Figure 7 shows the simplex words on the x-axis, grouped according to their frequency layer – and within each layer alphabetically. The values for Expanding Productivity are plotted on the y-axis.

According to this measure, the "winners" with regard to compound formation are simplex words with a middle or high frequency: *Haus* 'house' is the most productive simplex (e.g., *Barbiehaus* 'barbie house'; *Kaiserhaus* 'imperial house'; *Drei-Sterne-Haus* 'three-star house'), followed by 2) *Welt* 'world' (e.g., *Unterwelt* 'underworld'; *Vorstellungswelt* 'imaginary-world'), 3) *Kopf* 'head' (e.g., *Affenkopf* 'ape-head'; *Briefkopf* 'letterhead'; *Dickkopf* 'bullhead'; *Institutskopf* 'institution-head'), 4) *Wesen* 'being/entity' (e.g., *Bildungswesen*, lit. "entity of education" > 'education system'; *Einzelwesen* 'individual-being') and 5) *Mensch* 'person' (e.g., *Erfolgsmensch* 'success-person'; *Familienmensch* 'family-person').

Figure 7: Head words from three different frequency layers: Expanding Productivity (Hein 2016)



While the values for Realized Productivity point in the same direction – the most productive simplex words belong to the frequency layers 'high' and 'middle' (1. *Haus*; 2. *Kopf*; 3. *Welt*; 4. *Wesen*; 5. *Leben*) – the results for Potential Productivity are again the other way round (cf. Section 3.2.1).

In addition to the connection between frequency and compound productivity, the study with head words from three different frequency layers indicates the relevance of further parameters for productivity. Among others, the factor 'polysemy of the simplex in head position' seems to play a role here.[8]

This becomes clear when we look at the "winning head words" corresponding to Expanding Productivity again: *Haus* ('house'), *Welt* ('world'), *Kopf* ('head'), and *Wesen* ('being/entity') all have something in common: they can be understood as abstract nouns and as concrete nouns. Notice that the most productive head noun – *Haus* ('house') – is not the most frequent simplex of the investigation, but that it has many different readings. Among others, *Haus* can be understood as an abstract noun in the sence of 'dynasty' (cf. *Kaiserhaus*) as well as a concrete noun in the sence of 'building' (cf. *Barbiehaus*). Consequently, at first sight, head words that (in isolation) can be understood as both abstract nouns and concrete nouns seem to be more productive than head nouns that are not polysemous in that sense. However, a closer look at the corresponding compounds reveals that the basic meaning of the compounds ending in *house* is quite homogenous: they are clearly dominated by the main reading of house as 'building'. This puts into question the influence of the factor 'polysemy'

[8] However, it is known that there is a strong correlation between the frequency of a word and its polysemy, i.e., the two factors are interdependent to a certain degree.

and supports Fleischer and Barz (2012: 82) claim that the main reading of polysemous bases dominates their behavior as word formation units.

4. Outlook

At present, we are predominantly concerned with the following two issues: First, we are exploring automatic processes in the extraction and the processing of compounds. In particular, we are testing the applicability of morphologically parsing the extracted compound candidates. This should reduce the amount of manual annotation and facilitate the identification of more abstract patterns (e.g., N+N, A+N). Second, we are trying to gain a better understanding of the explanatory power of different possible measures for the productivity of compounds. Among other things, this requires us to have a better understanding of one of the problems of Baayen's productivity measures, namely, the dependency on the number of tokens, which makes it difficult to compare productivity values of categories with varying token size (cf. Section 2.2).

As was already mentioned in Section 3.1, in a next step, we will determine productivity values for semantic patterns of compounding, and we will investigate other potential factors for productivity (e.g., part of speech of the immediate constituents). In the long run, we also aim at gaining more general insights into the nature of composition with the help of the analysis of selected simplex words, semantic patterns, and their corresponding compounds.

References

Aronoff, M. (1976) *Word formation in generative grammar*. Cambridge: Massachusetts: MIT Press.

Baayen, H. R. (1992) Quantitative aspects of morphological productivity. In: G. Booij & J. van Marle. (Eds.), *Yearbook of morphology* (1991). Dordrecht: Kluwer Academic Publishers, 109-149.

Baayen, H. R. (1993) On frequency, transparency and productivity. In: G. Booij & J. van Marle (Eds.), *Yearbook of morphology* (1992). Dordrecht: Kluwer Academic Publishers, 181–208.

Baayen, H. R. (2001) *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Baayen, H. R. (2009) Corpus linguistics in morphology: morphological productivity. In: A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An international handbook.* Berlin: Mouton De Gruyter, 900–919.

Barðdal, J. (2008) *Productivity. Evidence from case and argument structure in Icelandic*. Amsterdam/ Philadelphia: Benjamins.

Bauer, L. (2001) *Morphological productivity*. Cambridge: Cambridge University Press.

Bauer, L. (2005) Productivity: Theories. In: P. Štekauer & R. Lieber (Eds.), *Handbook of word-formation.* Dordrecht: Springer, 315–334.

DGD (2017) Datenbank für gesprochenes Deutsch. Mannheim: Institut für Deutsche Sprache. http://dgd.ids-mannheim.de.

Engelberg, S. (2014) Gegenwart und Zukunft der Abteilung Lexik am IDS: Plädoyer für eine Lexikographie der Sprachdynamik. In: Institut für Deutsche Sprache (Ed.), *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache.* Mannheim: Institut für Deutsche Sprache, 243-253.

Engelberg, S. (2015a) Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons. In: L. M. Eichinger (Ed.), *Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. Jahrbuch 2014 des IDS.* Tübingen: Narr, 205-230.

Engelberg, S. (2015b) *Wortbildung - ganz dynamisch*. IDS Mannheim: unpublished manuscript.

Evert, S. & M. Baroni (2005) Testing the extrapolation quality of word frequency models. In: P. Danielsson & M. Wagenmakers (Eds.), *Proceedings of ‚Corpus Linguistics'*. Birmingham, July 14-17, 2005.

Fleischer, W. & I. Barz (2012) *Wortbildung der deutschen Gegenwartssprache*. Berlin/Boston: de Gruyter.

Gaeta, L. & D. Ricca (2006) Productivity in Italian word formation: a variable-corpus approach. *Linguistics* 44: 57–89.

Gaeta, L. & D. Ricca (2015) Productivity In: P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (Eds.), *Word-formation. An International Handbook of the Languages of Europe. Volume 2, IV: Rules and restrictions in word-formation I: General aspects.* Berlin/Boston: De Gruyter Mouton, 842–858.

Gaeta, L. & A. Zeldes (2012) Deutsche Komposita zwischen Syntax und Morphologie. Ein korpusbasierter Ansatz. In: L. Gaeta & B. Schlücker, *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte.* Berlin/New York: de Gruyter, 197–217.

Hacken, P. ten (ed.) (2016) *The semantics of compounding*. Cambridge: Cambridge University Press.

Hartmann, S. (2016) *Wortbildungswandel. Eine diachrone Studie zu deutschen Nominalisierungsmustern*. Berlin: de Gruyter.

Hein, K. (2015) *Phrasenkomposita im Deutschen. Empirische Untersuchung und konstruktionsgrammatische Modellierung*. Tübingen: Narr.

Hein, K. (2016) *Simplex-Frequenz-Studie. Zur Kompositionsaktivität von Simplizia aus drei Frequenzschichten.* IDS Mannheim: unpublished manuscript.

Institut für Deutsche Sprache (2017) *Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2017-II* (Release vom 01.10.2017). Mannheim: Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo.

O'Donnell, T. J. (2015) *Productivity and reuse in language. A theory of linguistic computation and storage*. Cambridge: Massachusetts: MIT Press.

Olsen, S. (2015) Composition. In: P. O. Müller, I. Ohnheiser, S. Olsen & F. Rainer (Eds.), *Word-formation. An International Handbook of the Languages of Europe. Volume 1, II: Units and processes in word-formation I: General aspects*. Berlin/Boston: De Gruyter Mouton, 364–386.

Plag, I. (1999) *Morphological productivity. Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Rainer, F. (1987) Produktivitätsbegriffe in der Wortbildungstheorie. In: Wolf Dietrich & Hans-Martin Gauger (Eds.), *Grammatik und Wortbildung romanischer Sprachen: Beiträge zum Deutschen Romanistentag in Siegen, 30.9.-3.10.1985.* Tübingen: Narr, 187–202.

Roth, T. (2014) *Wortverbindungen und Verbindungen von Wörtern. Lexikografische und distributionelle Aspekte kombinatorischer Begriffsbildung zwischen Syntax und Morphologie*. Tübingen: Francke.

Scherer, C. (2005) *Wortbildungswandel und Produktivität: eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Schneider, A.S. (2016) *Untersuchung der Produktivität von Komposita. Eine Korpusanalyse der Simplizia Furcht, Angst, Wut, Zorn, Hass und Scham anhand des Deutschen Referenzkorpus*. Universität Mannheim: unpublished manuscript.

Tarasova, E. (2013) *Some new insights into the semantics of English N+N compounds*. PH.D. thesis. Victoria University of Wellington: unpublished manuscript.