Annelen Brunner, Stefan Engelberg & Katrin Hein

The distribution of constituent words in nominal compounds and its impact on semantic interpretation: an empirical study[1]

Abstract: The paper explores factors that influence the distribution of constituent words of compounds over the head and modifier position. The empirical basis for the study is a large database of German compounds, annotated with respect to the morphological structure of the compound and the semantic category of the constituents. The study shows that the polysemy of the constituent word, its constituent family size, and its semantic category account for tendencies of the constituent word to occur in either modifier or head position. Furthermore, the paper explores the degree to which the semantic category combination of head and modifier word, e.g., x=SUBSTANCE and y=ARTIFACT, indicates the semantic relation between the constituents, e.g., y_CONSISTS_OF_x.

Keywords: N-N compound, compound interpretation, compound family, corpus linguistics, GermaNet, polysemy, word formation in German

1. Introduction

A traditional view of endocentric N-N-compounding would probably adhere to most of the following assumptions:[2]

I) A compound is a complex word with a binary structure and a morphosyntactic head; the immediate constituents of a compound are words (in the sense of word stems).[3]

II) A compound is semantically asymmetric in that – in German or English – the left constituent modifies the right constituent. (Thus, a compound is usually a hyponym of its right constituent.)

1 We are grateful to Felix Bildhauer and other colleagues from the project "Corpus grammar: grammatical variations in standard language and near-standard German" of the *Leibniz-Institut für Deutsche Sprache* for sharing their corpus data with us. Many thanks go to Alexander Koplenig for statistical consultations and to Lara Reichling and especially Julia Steinke for their support in manually annotating our data. The article also benefited greatly from the comments of the anonymous reviewers.

2 Cf. introductory texts such as Schlücker (2012) and Olsen (2015).

3 There are some exceptions such as phrasal compounds or confix formations; cf. Hein (2015: 40) for an overview.

 https://doi.org/10.3726/zwjw.2021.01.01

III) The relation between the two constituents is unmarked and underspecified and has thus to be determined from extra-morphological sources, e.g., from context.
IV) Compounding is fully productive.

As long as no further restrictions or distribution factors are assumed, it should follow that words are freely combined into compounds. If we pick three German words, for example *Garten* 'garden', *Problem* 'problem', and *Gemüse* 'vegetables', we can indeed form the six possible combinations and can easily imagine contexts in which they are interpreted as indicated: *Gartenproblem*, 'a problem with the garden'; *Problemgarten*, 'a problematic garden'; *Gartengemüse*, 'vegetables from the garden'; *Gemüsegarten*, 'a garden for vegetables'; *Problemgemüse*, 'problematic vegetables'; *Gemüseproblem*, 'a problem with vegetables'. We would also expect all the words to be more or less equally suited to occur in the position of the modifier and the position of the head. And if we check this, e.g., for *Garten*, in a large database of compounds (which is described in section 2), we can find evidence for this assumption: The database contains 1,077 compound types with *Garten* in modifier position and 1,028 compound types with *Garten* in head position.

However, as we know from literature on compounding, there are serious doubts as to whether assumptions I to IV are the whole truth about compounding; well-behaved words such as *Garten* in our introductory example might be an exception. From a usage-based perspective, it can be assumed (i) that words tend to prefer either the modifier or the head position in compounds, (ii) that there are fairly conventionalized interpretation patterns for compounds, and (iii) that productivity differs according to different domains of N-N-compounding.
In this paper we will consider the extent to which constituent words show a preference for either modifier or head position and which factors determine this preference. We will approach these questions from an empirical perspective, exploring the trends in a large database of German compounds.[4]

The paper proceeds as follows: In section 2, we will present the compound database that is the empirical resource for our investigation. Section 3 gives an overview of the research that pertains to our investigation. Sections 4 to 8 contain the descriptions of five corpus linguistic studies. For 8,363 constituent types in 707,910 compound types, we determine the general distribution patterns with respect to head and modifier position (Study A, section 4) and investigate the effects of the size of the constituent family (Study B, section

4 The research presented in this paper was conducted within the research project "Wortbildungsmuster / Patterns of Word Formation" at the *Leibniz-Institut für Deutsche Sprache* in Mannheim.

5) and the polysemy of constituents on this distribution (Study C, section 6). On the basis of a smaller dataset of 6,232 constituents in 184,823 compounds, we consider the role of the semantic category of the constituent word in head-modifier distribution (Study D, section 7). In this context, we also explore how strongly the semantic category combination is linked to a dominant interpretation pattern for the compound. The paper ends with a conclusion in section 9.

2. Database of German compounds

To compile our database, we used a subset of the German Reference Corpus (DeReKo, Release 2017-II),[5] the "KoGra Untersuchungskorpus",[6] which comprises roughly 7 billion tokens. Over 90% of the corpus is German newspaper texts, but it also contains some literary texts, and about 6% is spoken language material (cf. Bubenhofer, Konopka, and Schneider 2014). A custom word analyzer based on the Canoo Language Tools[7] was used to add detailed morphological annotation. On this basis, it was possible to automatically identify and extract a large collection of nominal compounds that serves as the basis for our studies on word formation.

In an earlier paper (Hein and Brunner 2020), we used an excerpt from these data (100,000 compounds) to study the role of the morphological complexity of the head constituent for the productivity of compound formation. For the studies in this paper, we started out with the whole database of 489,684,273 compound tokens and extracted all compounds comprising two simplex nouns, e.g., *Stadthalle* ('town hall'). Linking elements between the two constituents were allowed (e.g., *Arbeit-s-wut*, 'work mania'). We excluded derivative nouns as well as compounds with more than two constituents to reduce the number of variables in our study. 107,243,702 compound tokens, i.e., 21.9% of all identified nominal compound tokens from the KoGra corpus, matched our criteria.

From these data, we created two types of tables: a compound table, comprising all compound types, their frequencies, and immediate constituents, and a constituent table which contains a row for each immediate constituent type. For each constituent type, the constituent table lists its constituent family

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

6 *Korpus des Projekts Korpusgrammatik*. Leibniz-Institut für Deutsche Sprache: „Korpusgestützte Grammatik". Grammatisches Informationssystem grammis. DOI: 10.14618/korpusgrammatik. URL: https://grammis.ids-mannheim.de/korpusgrammatik/6615.

7 *canoonet. Deutsche Wörterbücher und Grammatik*. http://www.canoonet.eu.

size, i.e., the number of all compound types the constituent appears in. It also gives the number of all cases in which the compound appears in modifier position (its modifier constituent family) and all the cases where it appears in head position (its head constituent family). In a second preprocessing step, we assigned labels for semantic-thematic classes based on GermaNet (Hamp and Feldweg 1997; Henrich and Hinrichs 2010) to the constituents. GermaNet is a lexical-semantic net of the WordNet family. It relates German nouns, verbs, and adjectives by grouping lexical units that express the same concept into so-called synsets and by defining semantic relations between synsets. The way we used GermaNet to enrich our data was inspired by a study by Maguire, Wisniewski, and Storms (2010), who used WordNet categories in the same way to label their constituent data derived from English compounds[8]: We made use of the fact that the GermaNet vocabulary is organized in XML files that correspond to broad semantic-thematic fields such as ARTIFACT, PLANT, EMOTION, etc. These fields were created to organize work on GermaNet into cohesive packages, but are closely related to major nodes in the semantic network (cf. Hamp and Feldweg 1997: 10). We will go into more detail about the specifics of these categories in Studies D and E (sections 7 and 8). For nouns, 23 such XML files are available. We extracted the node words from these files and matched them against the constituent table.

Some constituents appeared in several of the GermaNet XML files, which is an indicator that the word has multiple senses. In contrast to Maguire, Wisniewski, and Storms (2010: 59–60), who identified the most common reading of such words using the WordNet hierarchy, we kept all labels. Multiple labels for a constituent were interpreted as an indicator of polysemy. This allowed for analyses contrasting the behavior of polysemous and monosemous constituents (Study C, section 6).[9]

---

8 The WordNet categories used by Maguire, Wisniewski, and Storms (2010) are nearly equivalent to the GermaNet categories. WordNet has three additional categories (ACT, STATE, and PROCESS) that are not available for German nouns. For a comparison between WordNet and GermaNet, cf. http://www.sfs.uni-tuebingen.de/GermaNet/germanet_structure.shtml. In the following, we will use the names of the corresponding WordNet categories as translations for the GermaNet categories.

9 To obtain a sense of whether this measure of polysemy is reliable, we carried out a study of 211 constituents where the number of senses assigned by GermaNet was compared to the number of senses assigned by two other lexicographical resources, Duden online (www.duden.de) and DWDS (www.dwds.de). We calculated the correlation coefficients between the three resources and found that they were in the same range (0.72–0.74) for all pairings. We conclude that though agreement between the resources is not perfect, GermaNet sense counts are not wildly different from or less reliable than the sense counts determined using other lexicographical resources.

Since the Canoo analysis was fully automatic, it contains errors with regard to compound segmentation and lemmatization of the constituents. We performed an evaluation based on a sample of 500 compound types and found that in 84.8% of the cases, the analysis was completely correct. In an additional 11.8%, the segmentation of the compound was correct, but there were errors in the lemmatization of one or both of the constituents (e.g., *Müller*, 'miller', was in some cases lemmatized to *Müll*, 'garbage'). The remaining 3.4% of the cases were serious errors, such as non-compound words. For a complex automatic analysis of such a large number of tokens, we deemed these error rates acceptable, given that our focus is on broad quantitative trends rather than detailed analysis. We nonetheless carried out some additional cleanup. Due to the large number of compounds in our study, it was impossible to correct the errors on compound type level, but a manual cleanup was performed on the list of labeled constituents: Constituents were removed if they were not recognizable as words at all or if they were not simplex nouns but either derivatives or compounds that had not been analyzed correctly. A remarkable characteristic of our constituent list is the fact that it contains a high number of proper names as well as foreign words.[10] We opted not to remove those words, as we believe they are legitimate elements of German compounds.

Once labeled and cleaned, the constituent list was used to filter the list of compound types such that only compounds remained that comprise a labeled first and second constituent, meaning that a list of 707,910 compound types (about 85% of the unlabeled and uncleaned list of N-N compound types) was retained. On this basis, the constituent table was re-created with a final count of 8,363 unique constituents. Studies A, B, and C in this paper are based on these lists. For Studies D and E, which examine the effects of the semantic category of the constituent word on its head-modifier distribution, we wanted to exclude polysemous constituents. We therefore only retained compound types that comprise a monosemous modifier and a monosemous head constituent ('bi-monosemous compounds') and created a new constituent table on this basis. Table 1 gives an overview of the statistics of the relevant datasets.

10 A random sample of 100 constituents from the list contained 14% proper names and 20–30% foreign words (range depending on whether foreign words that are well established in German were counted or not). These constituents tended to have smaller family sizes (average of 44–54) than regular nouns (average 267); about half of them had family sizes of 10 or fewer. In addition to that, we observed that the proper nouns (almost exclusively categorized as LOCATION by GermaNet) showed a strong tendency towards appearance as modifier, while the foreign words behaved in a similar way to normal nouns in this respect.

Tab. 1: Statistics for the datasets.

Studies	Compound types	Constituent types
A, B, C	707,910	8,363
D, E	184,823	6,232
(bi-monosemous compounds)	(26.11% of comp. types in A, B, C)	(74.52% of const. types in A, B, C)

To give an impression of what our data looks like, table 2 shows an excerpt from the compound table, table 3 an excerpt from the constituent table.

Tab. 2: Excerpt from the compound type table.

frequency	lemma	canoo analysis	modifier	head	m-category	h-category
1,269,620	Bürger-meister	(cmp:N&N bürgermeister_N (bürger_N) (meister_N))	bürger_N	meister_N	PERSON	PERSON
18,011	E-Mail-Adresse	(cmp:N&N:hy e-mail-adresse_N (e-mail_N) (adresse_N))	e-mail_N	adresse_N	COMMUNI-CATION	ARTIFACT, COMMU-NICATION, LOCATION
299	Schwanen-familie	(cmp:N&N schwanenfamilie_N (schwan_N)(en_xl) (familie_N))	schwan_N	familie_N	ANIMAL	GROUP, COGNI-TION
1	Asphalt-bett	(cmp:N&N asphaltbett_N (asphalt_N) (bett_N))	asphalt_N	bett_N	SUBSTANCE	ARTIFACT, PLACE

Tab. 3: Excerpt from the constituent table.

lemma	family size	mod. fam. size	head fam. size	m-compounds	h-compounds	%m	%h	category	senses
haus_N	3919	1468	2455	[Hausherr, Haustür, Hausarzt, ...]	[Gasthaus, Bürgerhaus, Schulhaus, ...]	37.46	62.64	ARTIFACT, GROUP, MOTIVE, PLACE	4
amsel_N	122	95	27	[Amselmutter, Amselnest, Amselstimme, ...]	[Wasseramsel, Stadtamsel, Spottamsel, ...]	77.87	22.13	ANIMAL	1
bö_N	5	0	5	[]	[Sturmbö, Regenbö, Hagelbö, ...]	0	100	PHENOM-ENON	1

3. Theoretical background

In literature on compounding from the last decade, we often encounter the assumption that "some nouns are more likely to occur as modifiers, whereas others show the preference of being modified" (Bauer, Beliaeva, and Tarasova 2019: 50), and that therefore the head constituent family of a constituent word and the modifier constituent family can differ considerably in size (Libben 2010: 319, and similarly Baayen 2010: 3; Fleischer and Barz 2012: 135; Tarasova 2013: 159; Roth 2015: 170). Even where a tendency towards an uneven distribution of constituent words is not stated explicitly, certain assumptions about the semantic structure of the field of compounds imply such a tendency: Many approaches to word formation assume – couched in the terminology of very different theoretical frameworks – that the interpretation of compounds is based on a certain number of interpretation patterns, such as Mätzner (1860: 469–477), Paul (1920: 9–10), Henzen (1947: 54), Hatcher (1960: 363–366), Lees (1960: 124–194), Brekle (1976: 141–187), Adams (1973: 68–83), Kürschner (1974), Levi (1978: 75–106), Warren (1978: 229–259), Fanselow (1981), Fleischer (1982), Ortner et al. (1991), Meyer (1993), Fandrych and Thurmair (1994: 39–40), Gagné and Shoben (1997: 72), Motsch (1999), Jackendoff (2009: 15), Hein (2015: 224–238), and Ortner and Ortner (2015: 1042–1046). These patterns describe semantic relations between the two immediate constituents of a compound $[X\ Y]_Y$, such as y_IS_MADE_OF_x (*Holztisch*, 'wood'–'table'), y_IS_LOCATED_IN_x (*Gebirgsbach*, 'mountain'–'creek'), or y_IS_PART_OF_x (*Türgriff*, 'door'–'handle'). The predicates that constitute the core of these patterns formulate selectional restrictions over their arguments. Thus, the predicate IS_MADE_OF presupposes a noun denoting a substance or material in modifier position (MP) and a word for an artifact in head position (HP). Looking at this pattern only, we would expect words for substances only to occur as the left, words for artifacts only as the right constituent of compounds. Theories based on interpretation patterns might thus explicitly or implicitly assume that constituent words show a tendency towards either modifier or head position within compounds. However, since there are many interpretation patterns, a word such as *Tisch* ('table'), denoting an artifact, occurs on the right side of some patterns (y_IS_MADE_OF_x, *Holztisch*, 'wood'–'table') and on the left side of others (y_IS_PART_OF_x, *Tischbein*, 'table'–'leg'). If we had a complete list of interpretation patterns with information about their selectional restrictions and their productivity, we could deduce from that information which words show a tendency towards left or right position and to what extent. However, opinions about the number of interpretation patterns differ widely between 4 (Hatcher 1960: 356) and 34 main patterns (with 142 sub-patterns) in Ortner et al. (1991). In conclusion,

the existence of interpretation patterns might at least urge words belonging to some semantic categories to occur more often in one of the immediate constituent positions in compounds. In Studies A to C, we explore factors (constituent family size, polysemy) that are indirectly related to the association of constituent words with few or many of these interpretation patterns and their particular positional restrictions. In Studies D and especially E, we delve deeper into the relation between combinations of semantic word categories and interpretation patterns.

A major debate in the literature on compounding is concerned with whether the modifier or the head of the compound is more crucial for accessing the semantic relation expressed by the compound. It is generally agreed that semantic and conceptual information about the immediate constituents of a compound plays a crucial role in the interpretation process (cf. Spalding et al. 2010: 283; Maguire, Wisniewski, and Storms 2010: 64–65). Thus, our studies – though they are not connected with experimental tasks – also build on the large amount of psycholinguistic literature on compound interpretation (cf. Spalding et al. 2010: 284 for an overview).

One well-known approach in the head-modifier debate is the so-called "CARIN theory" (Gagné and Shoben 1997) or "RICE theory" (Spalding et al. 2010) as a further development of CARIN. Presuming that compound constituents are associated with specific semantic relations and that such relations are also stored with constituents within the mental lexicon, it is assumed that semantic relations stored with the head and semantic relations stored with the modifier compete with each other when a novel compound is being processed (cf. Olsen 2012a: 137–138). CARIN or RICE theory claims that this competition is resolved as follows: "[…] the modifier suggests relations that compete with each other for selection and the head noun then plays an important role in evaluating whether a suggested relational interpretation is a plausible meaning for the combination" (Spalding et al. 2010: 284–285). This can be illustrated using the example of *chocolate bee*: As the 'made of' relation is often connected with the modifier *chocolate*, it can be assumed that this semantic relation is readily accessible for the compound, too. If the interpretation suggested by the modifier is indeed a plausible interpretation for the compound as a whole, then it is evaluated by the properties of the head as the next part of the process. In some cases, the interpretation which is suggested by the modifier can be rejected due to semantic properties of the head, e.g., *mountain planet*. In this compound, the head rejects the locative relation which is highly frequent for the modifier *mountain*, and which is instantiated in many other compounds, such as *mountain cabin* (Spalding et al. 2010: 284–286).

In contrast to this "'suggest-evaluate' framework" (Spalding et al. 2010: 286), schema-based theories of compound interpretation (e.g., Wisniewsky

1997; Murphy 1988) give more weight to the head than to the modifier. In their view, the head opens a schema with a slot into which the modifier is inserted. In a second step, world knowledge comes into play by recognizing "the need for a second stage of processing beyond slot filling, i.e., 'concept elaboration'. Once a slot is filled, world knowledge is used to refine the resulting combination" (Olsen 2012b: 2132; Maguire, Wisniewski, and Storms 2010: 50–51). This two-stage process can be illustrated using the example of *plastic chair*: During the slot-filling mechanism, *plastic* is inserted in the 'made of' slot which is opened by the concept *chair*. With the help of world knowledge – i.e., the use of plastic chairs as garden furniture – the concept of 'plastic chair' undergoes further elaboration. In addition to the modifier-orientated and head-orientated approaches sketched above, a third perspective on compound meaning adopts a pattern-based approach (e.g., Maguire, Wisniewski, and Storms 2010; Tarasova 2013), which postulates that compound interpretation relies on – or is at least facilitated by – semantic patterns. In this context – as in CARIN theory – aspects of frequency are taken into consideration. It is assumed that "people rely on statistical knowledge about how nouns tend to be used in combination in order to facilitate the interpretation of novel compounds" (Maguire, Wisniewski, and Storms 2010: 50). In contrast to schema-based theories, pattern-based approaches reject the idea that a full conceptual schema must be activated when a noun occurs in the head role. Instead, they start from the premise that exploiting regular patterns in compounding allows for the selective activation of conceptual knowledge that is connected to the constituents. For example, the complex word *mountain bird* can be correctly interpreted as "a bird located in the mountains" (locative interpretation) without having a detailed idea of a mountain bird (Maguire, Wisniewski, and Storms 2010: 65–66).

Taking up assumptions from Maguire, Wisniewski, and Storms (2010), we explored in Studies D and E whether the semantic categorization of constituent words helps to explain MP-HP preferences and to infer interpretation patterns.

4. Study A: Overall tendency of words towards occurrence in head or in modifier position

Without any further restrictions, the basic structural position as expressed in I to IV (section 1) would lead us to expect constituent types to be evenly distributed over modifier and head position. That is, the more compounds are formed with a particular constituent word, the closer the set of compounds should approach a 50-50 distribution with respect to the position of this constituent word. The opposite assumption, namely that constituent types

are unevenly distributed over head and modifier position, would presuppose restrictions or distribution factors that go beyond the basic assumptions in I to IV. Possible semantic reasons for this assumption were presented in section 3. Constituent family size and constituent polysemy as factors that are (probably) indirectly connected to the existence of interpretation patterns will be examined in Study B (section 5) and Study C (section 6). While an even distribution over head and modifier position is associated with a clear quantitative pattern, an operationalization of the opposite assumption, i.e., that constituents show a tendency to occur either in modifier or in head position, is harder to establish. How is this "tendency" to be understood? Do words show almost affix-like restrictions towards first or second position in compounds or do they usually only exhibit slight tendencies towards an uneven distribution? Do words cluster around particular distribution patterns or do they differ widely in their distributional behavior within compounds?

In order to gain an initial idea of how constituent words behave with respect to preferences for modifier and head position, we sorted our 8,363 constituent words into 101 distribution classes, with the first class containing words that always occur in second position (modifier position MP = 0%, head position HP = 100%; rounded values), the second class showing a distribution of MP 1%, HP 99%, etc., and the last class containing words with a distribution of MP 100%, HP 0%.

We can now verify three conceivable assumptions about the distribution of the constituent words over the 101 distribution classes: (i) If the basic, unrestricted structural assumption that words are equally likely to occur in modifier and in head position is correct, we would expect a Gaussian distribution with the expected value at 50:50. (ii) The opposite assumption that words move towards extremely uneven distributions would yield a bimodal curve with peaks at 0:100 and 100:0. (iii) Words could instead be expected to show a tendency towards either modifier or head position while still retaining a certain positional flexibility. It might be an unusual idea to bring Zipf's Law into play here since it usually serves to account for the token distribution of large sets of types, such as all the words of a vocabulary. However, Zipfian distributions reflect the balance between conventionalization and expressivity in many domains of the lexicon. Uneven, right-skewed distributions appear not only in the vocabulary as a whole but also in domains with small sets of types, i.e., the different meanings of a word or the different argument structures that can be realized with a verb (cf. Piantadosi 2014 for Zipfian distributions in sets with very few types). For the sake of argument, let us assume a Zipfian distribution over just two types: a constituent word in modifier vs. head position. Since Zipf's Law predicts that the most frequent option will occur twice as often as the second most frequent option, this would amount to the prediction that in two thirds

of the cases a constituent word assumes one of the two possible positions and in one third of the cases it assumes the other, i.e., a distribution with peaks either around the 67:33 or 33:67 classes.

If we select all constituent words from our database with a constituent family size of at least 2, we obtain 7,942 constituent words. Their distribution with respect to modifier and head position is visualized in figure 1.

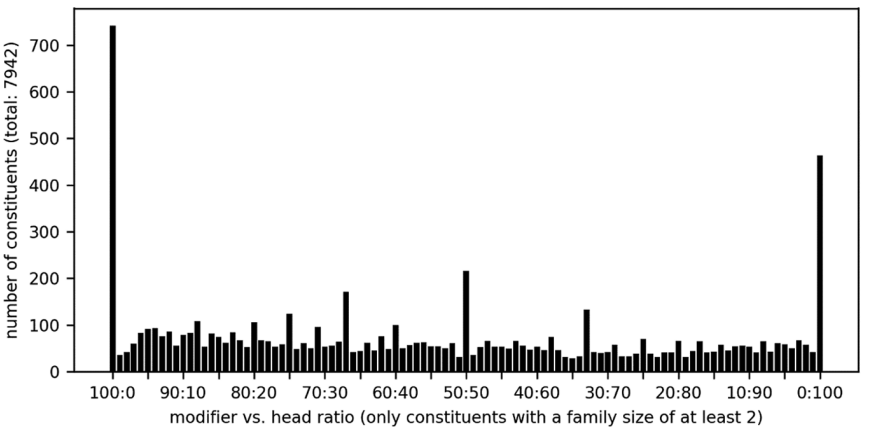

Fig. 1: Membership of constituent words with a constituent family size ≥ 2 in MP-HP distribution classes.

At first glance, the distribution seems to follow assumption (ii), namely that words tend to accumulate in the extreme outer classes. However, this is of course an artifact, since words with a constituent family size of 2 can only show 3 different MP:HP distributions: 100:0, 50:50 and 0:100. Thus, even if distributed by chance, two thirds of the constituent words with a constituent family size of 2 will fall into the two distribution classes on the extreme left and right. In order to avoid this effect, we will only look at those simplex constituents that have a constituent family of at least 100 compound types. Figure 2 shows the distribution of the 2,827 words that match these criteria.

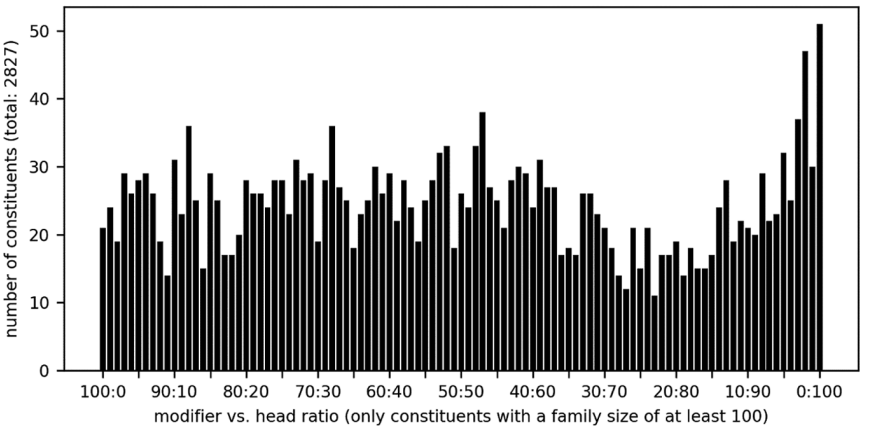

Fig. 2: Membership of constituent words with a constituent family size ≥ 100 in MP-HP distribution classes.

Discussion: Disregarding slight differences, figure 2 shows that even though most words deviate from a 50:50 distribution, there seems to be no preferred distribution pattern, i.e., none of the three assumptions finds particular support in the data. There is no generally preferred distribution pattern for constituent words in compounds. Since it would be surprising if words were really distributed by chance over compound positions, we will now examine factors that might account for the distribution.

5. Study B: The influence of family size on the distribution of constituent words

A factor that might account for the distributional behavior of words as constituents of compounds was investigated by Tarasova (2013, 2019). She claims not only that constituent words show an uneven distribution with respect to the modifier and the head position but also that this effect is more pronounced with constituent words with a large family size. In order to check this within our data, we restricted our assignment of words to distribution classes to the 284 words with a family of at least 1,000 compounds (figure 3).

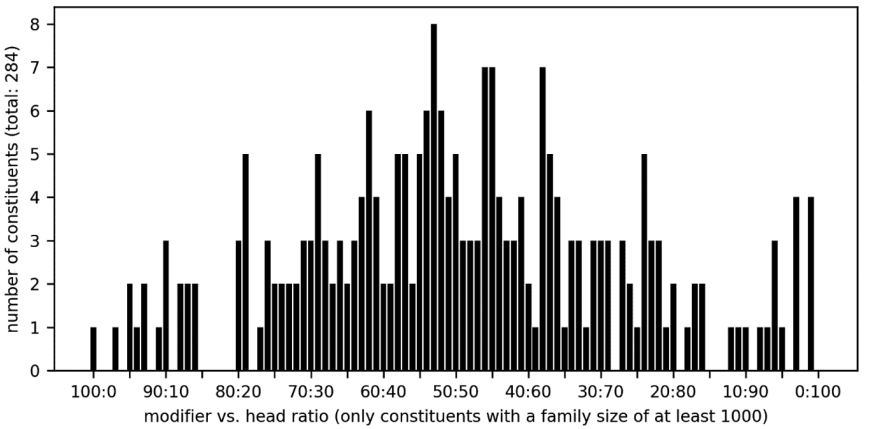

Fig. 3: Membership of constituent words with a constituent family size ≥ 1,000 in MP-HP distribution classes.

As figure 3 shows, there is no evidence for Tarasova's claim in our data.[11] On the contrary, words with very large families tend to distribute more evenly with respect to modifier/head positions than words with smaller families. In order to check this observation more thoroughly, we performed a second computation starting from the null hypothesis that constituent words are distributed evenly over modifier and head position independently of a word's constituent family size. We divided the 8,363 constituent words into 10 classes according to the size of their constituent families (cf. table 4). The 10 classes are of approximately similar size and are characterized by the following properties: "N class" = number of constituent word types in this class; "CFS" = range of constituent family size for words in this class; "Cmp" = number of compound types with words of this class as constituents; "M-Cmp" = number of compound types where the words of this class occur

11 We will not discuss the reasons for the differences between Tarasova's (2013, 2019) results and ours in detail. They might be due either to the languages investigated – Tarasova's study is about compounding in English – or to the fact that Tarasova's results are based on a much smaller number of constituent words (100 constituent words with 7,332 compounds; Tarasova 2013: 98–101), the selection of which was balanced with respect to the proportion of semantic compound patterns found in a larger part of the corpus. Tarasova (2019: 56) also emphasizes that she considers only 'non-lexicalized' compounds. However, it seems that she is not referring to hapax legomena here but to semantically transparent compounds. The proportion of non-transparent compounds in our data is fairly small, such that this difference in the design of our studies does probably not account for the difference in our results.

in modifier position; "H-Cmp" = number of compound types where the words of this class occur in head position; "Prop M" = proportion of compound types in this class in which the constituent word occurs in modifier position. If Prop M takes the value 50, we have an even distribution within a group. For example, class IV is defined as the set of constituent words that occur in 14 to 24 different compounds. It contains 881 different constituent words, which account for 16,504 compound types. In 59.6% of the cases (9,836 compound types), the constituent word occurs in modifier position.

Tab. 4: Distribution of constituent words over 10 classes reflecting constituent family size.

Class	N class	CFS	Cmp	M-Cmp	H-Cmp	Prop M
I	760	1–2	1,099	677	429	61.60
II	793	3–6	3,416	2,062	1,371	60.36
III	873	7–13	8,516	5,233	3,320	61.45
IV	881	14–24	16,504	9,836	6,730	59.60
V	846	25–41	27,526	15,585	12,016	56.62
VI	852	42–70	46,341	25,374	21,063	54.75
VII	841	71–121	78,974	41,543	37,550	52.60
VIII	854	122–225	143,527	73,364	70,354	51.12
IX	832	226–468	274,273	136,988	137,624	49.95
X	831	469–4008	813,788	397,248	417,453	48.81

Figure 4 displays boxplots for the 10 groups, showing the distribution of the Prop M values of the individual constituents within the group.

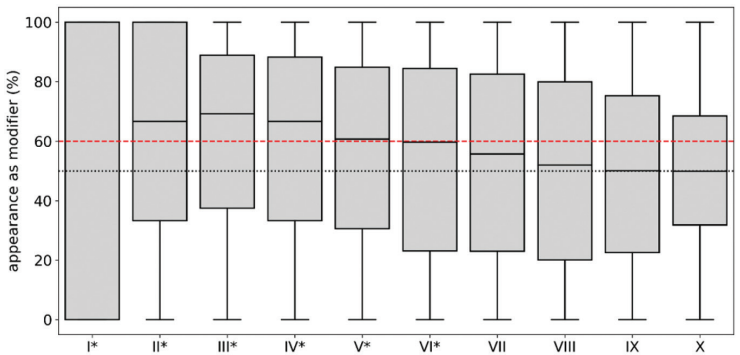

Fig. 4: Classes of constituent words according to their family size, and deviation of the classes from an even MP-HP distribution; the higher the values on the y-axis, the stronger the tendency to occur in modifier position. The star marks the classes that deviate statistically significantly from 50% (dotted line). The red dashed line marks the median of the Prop M values for all constituents.

Discussion: Two observations can be made: (i) The larger the family size, the more evenly distributed are the constituent words over modifier and head position.[12] (ii) Except for constituents with very large constituent families, constituent words tend to occur more often in modifier position. In fact, we find that in our dataset there is an overall tendency towards modifier position (visualized by the red dashed line). This means that there is a greater number of modifier types than head types in our data.[13] However, the boxplots also show that the scattering of the Prop M values of the individual constituents is extremely high.[14]

6. Study C: The influence of polysemy on the distribution of constituent words

Another factor we considered was polysemy: Does the fact that a word has multiple senses have an effect on its propensity to appear in modifier or head position? We sorted the constituents into two groups: monosemous (only one sense) and polysemous (more than one sense). Table 5 shows the statistics for these groups.

Tab. 5: Distribution of monosemous vs. polysemous constituent words.

Class	N class	Cmp	M-Cmp	H-Cmp	Prop M
Monosemous	6,601	726,811	382,560	345,320	52.64
Polysemous	1,762	687,153	325,350	362,590	47.35

12 We performed one-sample t-tests to determine whether the deviation from 50% is statistically significant (p=0.05, adjusted for 10 tests). The tendencies of the four classes VII to X towards uneven distribution are not statistically significant.

13 It might seem confusing that the distribution of modifier and head can be unequal over the whole dataset – after all, the overall number of modifiers and heads in compounds that were used to compile the constituent list has to be identical: *n* compound types consist of *n* modifiers and *n* heads. However, the constituent list contains *types* – i.e., *unique* words appearing as part of a compound and the number of modifier types and head types can of course differ. In the set of compound types {*blackbird*, *blackmail*, *blackberry*}, we count three compound types and therefore three heads and three modifiers; however, the modifiers belong to only one type and the heads to three different types.
With respect to figure 4, it must be considered that the words in class X account for more than half of the 1,413,964 compound constituents in our database, such that the slight tendency towards head position within this class makes up for the tendency towards modifier position observed in all the other classes.

14 Tarasova (2019: 62) also observes strongly scattering patterns with respect to preferences for head or modifier position. She assumes that a number of semantic, morphological, and usage factors might influence the distribution.

Monosemous constituents show a tendency towards appearance as a modifier, while polysemous constituents tend to occur in head position. One-sample t-tests confirmed that the divergence from 50% is statistically significant in both cases (p=0.05, adjusted for 2 tests). At the same time, we observed that the monosemous group is much more diverse, with a larger number of constituent types and smaller family sizes. The average family size for a monosemous constituent is 110 compound types, as opposed to 390 for a polysemous constituent.

Discussion: In Study B, we observed a tendency of constituent words with small family sizes towards modifier position. As Study C shows, monosemous words also tend towards modifier position, while polysemous words show a preference for head position. This probably partly reflects the tendency of monosemous words to have smaller family sizes than polysemous words. However, the tendency towards head position in the set of polysemous words is stronger even than in the set of words with the largest family sizes (cf. group X, table 4). Thus, polysemy seems to be a stronger predictor for head position than family size.

When looking for explanations for our empirical findings, it is first of all not surprising that polysemous words have much larger constituent families, assuming that each distinct meaning of a word has its own potential in word formation. Regarding their tendency towards head position, polysemous words need to be disambiguated and the formation of a hyponym can provide this disambiguation. Modifying *Herz* ('heart') by a compound modifier uncovers the different meanings of *Herz*: *Hundeherz* ('dog'–'heart'), *Schokoladenherz* ('chocolate'–'heart'), *Palmherz* ('palm'–'heart'), *Kunstherz* ('art(ificial)'–'heart'). For the same reason, polysemous words in modifier position are less suited to disambiguate a head constituent. Polysemous words can also be expected to occur in more different interpretation patterns, where some of the patterns assign them to modifier position and others to head position.

Conversely, the fact that the class of monosemous words with its larger lexical variability shows a tendency towards modifier position might be due to the basic function of compounding as a device to produce hyponyms. For our data in particular, the modifier bias of monosemous words might also partly be due to the high number of proper names contained in our constituent table (15% in a sample of 100, cf. section 2). Proper names are mostly monosemous and do not lend themselves to hyponymic subcategorization. Therefore, they are unlikely to appear as heads.

7. Study D: The influence of semantic category membership on the distribution of constituent words

In Study D, we investigate the potential influence of semantic-thematic category membership on the distribution of constituent words between head and modifier. The study will also show the extent to which the tendency of monosemous constituent words to occur in modifier position (cf. Study C) depends on their semantic class.

GermaNet categories were used to determine the category of each constituent. As explained in section 2, we based this study on a smaller dataset (184,823 compounds, 6,232 constituents), which contains only compounds that comprise a monosemous modifier and a monosemous head ('bi-monosemous' compounds). Thus, we had an unambiguous label for each constituent. Similar to our approach in Study B (section 5), we created boxplots to show the distribution of Prop M for each of the 23 semantic/thematic GermaNet classes (cf. figure 5) and performed one-sample t-tests to determine whether the divergence from 50% was statistically significant (p=0.05 adjusted for 23 tests) within each group.

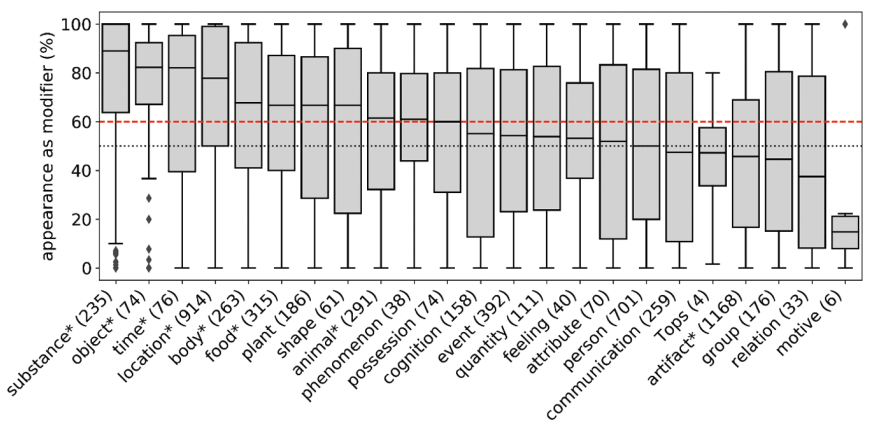

Fig. 5: MP-HP tendency of constituents belonging to particular GermaNet categories, ordered according to their median; groups with a statistically significant divergence from 50% are marked with stars; the dashed red line shows the median of the Prop M values of all constituents in the dataset; the numbers after the category names show the number of constituent types belonging to each group.[15]

15 Two groups are extremely small – TOPS (4) and MOTIVE (6). TOPS is also not a true semantic-thematic group, but rather an organizational top-level group for GermaNet, so it is not suitable for our purposes anyway. We did not include these two groups in our further analyses.

Discussion: Several observations can be made: (i) For most categories, the whiskers in the box plots reach out to the top and the bottom of the MP-HP distribution, indicating that constituent words occur across the board of distribution options. As we have argued with respect to figure 1, this is mainly a logical consequence of the fact that many constituent words have very small constituent families. (ii) Most categories show a tendency towards modifier position. This reflects the fact that, overall, we find a greater variety of constituent word types in modifier than in head position; this trend is slightly more pronounced than in the dataset for Studies A-C and is indicated by the red dashed line (cf. figure 5).[16] (iii) Seven categories (SUBSTANCE, (NATURAL) OBJECT, TIME, LOCATION, BODY, FOOD, and ANIMAL) exhibit a statistically significant tendency towards modifier position. ARTIFACT is the only GermaNet group that shows a statistically significant propensity towards head position.

There are fairly straightforward explanations for the disproportionate tendency of some categories towards modifier position. We will briefly look at the four leftmost categories in figure 5: SUBSTANCE, (NATURAL) OBJECT, TIME, and LOCATION. Words for substances occur particularly often in one frequent interpretation pattern, namely x^{ARTIFACT} IS MADE OF $y^{\text{SUBSTANCE}}$, where they occupy the modifier position (e.g., *Messingdach*, 'brass'–'roof'). This pattern might also be one of the factors that explain why ARTIFACT nouns tend towards head position. Due to the specific categorization conventions in GermaNet, all words for natural substances are subsumed under the category of (NATURAL) OBJECT, e.g., *Marmor* ('marble'), *Schiefer* ('slate'), *Torf* ('peat'); therefore, many of the words in this category also show a strong affinity for the MADE-OF pattern. The category of LOCATION predominantly contains place names, which can be used in modifier position to subcategorize non-toponymic words (e.g., *Neckardeich*, 'Neckar'–'dike') but, as mono-referential words, they can rarely be subcategorized themselves by other modifiers. Within the category of TIME, we find words for defined time spans such as *Mai* ('May'), *Ostern* ('Easter'), *Herbst* ('fall'), which, like place names, can modify other words, but do not lend themselves easily to modification. In summary, the MP-HP distribution is sensitive to semantic-thematic categories to a certain degree. However, within most categories, the words show a fairly dispersed distribution with respect to MP versus HP position.

16 Surprisingly, in their similar study for English, Maguire, Wisniewski, and Storms (2010: 60) found the reverse trend for their data: "11,765 different nouns were used as modifiers in the BNC and 13,550 nouns were used as heads."

8. Study E: The connection between interpretation patterns and the distribution of constituent words

As we assume that interpretation patterns play a crucial role in explaining the tendency of constituents towards modifier or head position, we will now focus on the semantic category membership of both head and modifier noun in combination, in an attempt to identify established semantic patterns in compounding. In doing this, we follow Maguire, Wisniewski, and Storms' (2010: 63) assumption that "separating nouns into a small number of broad semantic categories is sufficient for revealing consistent patterns in modifier and head use". The hypotheses in Maguire, Wisniewski, and Storms (2010: 58) are as follows: (i) The combinations of modifier category plus head category are not randomly distributed over compounds. Combinations occur more or less often than would be expected from a random distribution, taking the group size of the single categories into account. (ii) Combinations that occur beyond expectation reflect productive interpretation patterns. "For example, we expected combinations of the form [substance – artifact] to be predominantly associated with the <made of> relation, since artifacts typically have a material constitution that can be denoted by a substance concept. Accordingly, we predicted that a relatively large proportion of combinations should fall into the [substance – artifact] category" (Maguire, Wisniewski, and Storms 2010: 58).

In order to verify Maguire, Wisniewski, and Storms' (2010) first hypothesis, we reproduced their study on the basis of our own data. The 23 semantic-thematic GermaNet categories yielded 529 possible combinations, of which 499 are attested in our set of 184,823 bi-monosemous compound types. The ten most frequently attested combinations are shown in table 6.

Tab. 6: The ten most frequently attested category combinations (on the basis of 184,823 bi-monosemous compound types).

combination	frequency	examples
ARTIFACT | ARTIFACT	10,594	*Skilift* ('ski'–'lift'), *Briefbombe* ('letter'–'bomb'), *Kaminzimmer* ('fireplace'–'room'), *Käfigtür* ('cage'–'door')
LOCATION | ARTIFACT	5,788	*Hotelzimmer* ('hotel'–'room'), *Gartenmöbel* ('garden'–'furniture'), *Donauschiff* ('Danube'–'boat'), *Grablaterne* ('grave'–'lantern')
LOCATION | LOCATION	5,675	*Rheintal* ('Rhine'–'valley'), *Strandhotel* ('beach'–'hotel'), *Inselparadies* ('island'–'paradise'), *Heimatplanet* ('home'–'planet')
PERSON | ARTIFACT	5,158	*Bürgersaal* ('citizen'–'hall'), *Arztkittel* ('doctor'–'coat'), *Brautschuh* ('bride'–'shoe'), *Papstbrief* ('pope'–'letter')

ARTIFACT | PERSON	5,065	*Zimmermädchen* ('room'–'maid'), *Computerexperte* ('computer'–'expert'), *Buspassagier* ('bus'–'passenger'), *Tresordieb* ('safe'–'thief')
LOCATION | PERSON	4,270	*Gartenfreund* ('garden'–'friend'), *Ozeanriese* ('ocean'–'giant'), *Hotelchef* ('hotel'–'boss'), *Europaexperte* ('Europe'–'expert')
SUBSTANCE | ARTIFACT	4,142	*Lederjacke* ('leather'–'jacket'), *Dampfboot* ('steam'–'boat'), *Betontreppe* ('concrete'–'staircase'), *Staubbrille* ('dust'–'glasses')
EVENT | ARTIFACT	3,891	*Festzelt* ('festival'–'tent'), *Mordwaffe* ('murder'–'weapon'), *Schlachtschiff* ('battle'–'ship'), *Therapiezimmer* ('therapy'–'room')
PERSON | PERSON	3,762	*Ministerkollege* ('minister'–'colleague'), *Kaisersohn* ('emperor'–'son'), *Nazischerge* ('Nazi'–'henchman'), *Bügermädchen* ('citizen'–'girl')
ARTIFACT | LOCATION	3,757	*Yachthafen* ('yacht'–'harbour'), *Dommuseum* ('cathedral'–'museum'), *Dachgarten* ('roof'–garden'), *Mauerbereich* ('wall'–area')
...	...	

As we saw in figure 5, the number of constituents belonging to each group varies considerably, with ARTIFACT being by far the largest group (1,168 constituents). It is therefore not surprising to find the ARTIFACT | ARTIFACT combination at the top of the list. However, looking just at the most frequent combinations might lead us to overlook more remarkable combinations. Following Maguire, Wisniewski, and Storms (2010) once again, we corrected for the factor of group size and computed the ratio of compound types observed for a particular group combination vs. the number expected given the modifier and head counts for the two relevant groups (cf. Maguire, Wisniewski, and Storms 2010: 62). The higher this value, the more unexpected it is to find this particular combination pattern. Table 7 lists the highest-ranking combinations according to the ratio value.[17]

17 In table 7, all combinations containing the sparse categories TOPS and MOTIVE were omitted. These combinations tended to rank very high with the ratio measure, as their sparseness made their combination with nearly any other category unexpected, but they contained a very small number of compounds which were not representative and often contained errors.

Tab. 7: Category combinations ranked by their predictability ratio.

combination	ratio	freq.	examples
PLANT | PLANT	13.73485759	309	*Haselstaude* ('hazel'–'shrub'), *Yuccapalme* ('yucca'–'palm'), *Rapspflanze* ('rape'–'plant'), *Ligusterhecke* ('privet'–'hedge')
FOOD | FOOD	12.59698481	2,272	*Gulaschsuppe* ('goulash'–'soup'), *Obstsalat* ('fruit'–'salad'), *Hefeteig* ('yeast'–'dough'), *Butterkeks* ('butter'–'cookie')
(NAT.) OBJECT | (NAT.) OBJECT	11.99252780	114	*Felsklippe* ('rock'–'cliff'), *Sandtorf* ('sand'–'peat'), *Basaltschotter* ('basalt'–'gravel'), *Strohlehm* ('straw'–'clay')
(NAT.) PHEN. | (NAT.) PHEN.	9.553912008	26	*Monsunregen* ('monsoon'–'rain'), *Donnerknall* ('thunder'–'bang'), *Regenbö* ('rain'–'gust'), *Nebelregen* ('fog'–'rain')
EMOTION | EMOTION	7.719648007	57	*Angstlust* ('fear'–'desire'), *Spottlust* ('mockery'–'desire'), *Horrorangst* ('horror'–'fear'), *Wutkummer* ('rage'–'sorrow')
ANIMAL | BODY	6.897747044	715	*Tierkadaver* ('animal'–'carcass'), *Tigerkralle* ('tiger'–'claw'), *Elchgeweih* ('elk'–'antler'), *Dackelblut* ('dachshund'–'blood')
SUBSTANCE | SUBSTANCE	6.793769818	829	*Atommüll* ('atom'–'waste'), *Asbestfaser* ('asbestos'–'fiber'), *Uranerz* ('uranium'–'ore'), *Metallspan* ('metal'–'splint')
PROPERTY | PROPERTY	6.762244399	198	*Investmentfonds* ('investment'–'fund'), *Zinsmarge* ('interest'–'margin'), *Leasingbranche* ('leasing'–'sector'), *Mauttarif* ('toll'–'tariff')
QUANTITY | QUANTITY	6.734233797	131	*Kubikmeter* ('cubic'–'meter'), *Dollarmillion* ('dollar'–'million'), *Promillezahl* ('per mil'–'number'), *Meterscheit* ('meter'–'log')
TIME | (NAT.) PHENOMENON	6.516136413	94	*Sommermonsun* ('summer'–'monsoon'), *Herbstnebel* ('fall'–'fog'), *Julihitze* ('July'–'heat'), *Mairegen* ('May'–'rain')
...	...	...	...

Not surprisingly, and in accordance with our finding that semantic category membership is a factor for MP-HP distribution, Maguire, Wisniewski, and Storms' (2010) first hypothesis is confirmed by our data: The combinations of modifier category plus head category are not randomly distributed over compounds. According to Maguire, Wisniewski, and Storms' (2010) second hypothesis, we would expect the top combinations in table 7 to be closely connected to particular interpretation patterns. The most striking finding is that the combination of words of the same semantic category is very common. The five highest ranking combinations in table 7 show exactly this type of combination: PLANT | PLANT, FOOD | FOOD, (NATURAL) OBJECT | (NATURAL) OBJECT, (NATURAL) PHENOMENON | (NATURAL) PHENOMENON, and EMOTION | EMOTION. We had a closer look at the three top combinations – PLANT | PLANT, FOOD | FOOD, and (NATURAL) OBJECT | (NATURAL) OBJECT – and the top-ranked combination of non-identical categories – ANIMAL | BODY – and annotated a random sample of 100 compound types each for their combination patterns.

PLANT | PLANT: Due to the fairly coarse semantic classification of GermaNet, the category PLANT contains different semantic types of words, in particular: (i) words for species and other taxa (*Farn*, 'fern'), words for general types of plants and growth habits (*Blume*, 'flower'; *Busch*, 'bush'), and words for parts of plants (*Stengel*, 'stalk'). It is this category-internal diversity that accounts for the fact that in the extension of PLANT | PLANT, we find quite a number of different interpretation patterns: M_IS_TYPE_OF_H (48%) (*Schierlingsstaude*, 'hemlock'–'shrub'), H_IS_PART_OF_M (19%) (*Krokusknospe*, 'crocus'–'bud'), H_CONSISTS_OF_M (7%) (*Tanggestrüpp*, 'seaweed'–'thicket'), H_IS_LIKE_M (7%) (*Palmfarn*, 'palm'–'fern'), M_IS_PART_OF_H (6%) (*Knollenbegonie*, 'tuber'–'begonia'), H_IS_TYPE_OF_M (6%) (*Strauchmalve*, 'bush'–'mallow'), and some minor and unidentifiable relations (7%).

FOOD | FOOD: As with the category PLANT, the group of FOOD words contains different types of words, in particular: (i) words for natural (edible) objects (*Apfel*, 'apple'), (ii) words for natural (edible) substances (*Mehl*, 'flour'), (iii) words for dishes as objects (*Knödel*, 'dumpling'), and (iv) words for dishes as substances (*Suppe*, 'soup'). The interpetation relations found in the sample are: H_CONTAINS_M (69%) (*Lakritzbier*, 'licorice'–'beer'), H_IN_COMBINATION_WITH_M (11%) (*Spinatomelett*, 'spinach'–'omelette'), H_IS_FOR_M (5%), (*Fonduesoße*, 'fondue'–'relish'), H_IS_MADE_OF_M (3%) (*Brezelbrösel*, 'pretzel'–'crumbs'), and some minor patterns and compounds with unidentified relations (12%).

(NATURAL) OBJECT | (NATURAL) OBJECT: The category (NATURAL) OBJECT subsumes colloquial words for geological formations (*Felsen*, 'rock'; *Klippe*, 'cliff') and weather phenomena (*Hagel*, 'hail'; *Nebel*, 'fog') as well as words

that are known to most speakers but might cause uncertainties with respect to their meaning and reference (*Löss*, 'loess'; *Schiefer*, 'shale'; *Basalt*, 'basalt'; *Alabaster*, 'alabaster'). This uncertainty often makes it difficult to determine the relation between the two constituents, e.g., *Lehmlöss* ('clay'–'loess'): Does *Lehm* contain *Löss* or the other way round? Is one the origin of the other? Is it a mixture? The dominant interpretation pattern for (NATURAL) OBJECT | (NATURAL) OBJECT is H_CONSISTS_OF_M (42%) (*Schieferflöz*, 'shale'–'seam'). Other patterns are H_CONTAINS_M (18%) (*Topasfelsen*, 'topaz'–'rock/boulders'), H_IS_LIKE_M (7%) (*Mahagoniobsidian*, 'mahogany'–'obsidian'), M_CONTAINS_H (5%) (*Schlicksand*, 'silt'–'sand'), and a number of less frequent and unidentified relations (28%).

ANIMAL | BODY: The category ANIMAL mainly contains words for animal species; the category BODY is dominated by words for body parts, but also contains words like *Embryo* ('embryo') and *Urin* ('urine'). The combination is dominated by the interpretation pattern H_IS_PART_OF_M (56%) (*Pavianhand*, 'baboon'–'hand'), followed by H_IS_PRODUCED_BY_M (18%) (*Drachenkot*, 'dragon'–'feces'), M_IS_ORIGIN_OF_H (7%) (*Krokodilfötus*, 'crocodile'–'fetus'), and several other patterns and unidentified relations (19%).

Discussion: The attempt to derive interpretation patterns from GermaNet semantic categories was partly successful for category combinations that occur more often than expected. The most homogenous group was FOOD | FOOD (69% M_IS_TYPE_OF_H), followed by ANIMAL | BODY (58% H_IS_PART_OF_M), PLANT | PLANT (48% M_IS_TYPE_OF_H), and (NATURAL) OBJECT | (NATURAL) OBJECT (42% H_CONSISTS_OF_M). However, it is unclear whether the ratio rankings in table 7 are really a good indicator for groups with dominant interpretation patterns, as the combination SUBSTANCE | ARTIFACT that Maguire, Wisniewski, and Storms (2010) discuss as the model case for the relation between category combination and interpretation pattern achieves only rank 77 in our table and is not among the 10 top-ranking combinations in their own data either (Maguire, Wisniewski, and Storms 2010: 63).

A closer look at Maguire, Wisniewski, and Storms' (2010) idea to derive the interpretation relations from the combination of semantic categories reveals several flaws. Firstly, even with category combinations that occur with frequencies well beyond chance, the proportion of compounds whose meaning is correctly deduced is often below 50%. Secondly, in order to use the category combinations as a helpful general device for interpretation, they should provide correct interpretations in particular for combinations with high absolute frequencies. The most frequent combination, ARTIFACT | ARTIFACT, with 10,594 attested compounds, occurs more than twice as often as the top ten combinations in table 7 together (4,745 attestations). We annotated

a random sample of 100 compounds of the type ARTIFACT | ARTIFACT for interpretation patterns. The combination shows affinities to a large number of interpretation patterns. Among them, the most frequently attested pattern H_IS_PART_OF_M (*Moscheefenster*, 'mosque'–'window'; *Jackettkragen*, 'jacket'–'collar'; *Revolvertrommel*, 'revolver'–'drum') accounts for only 20% of the compounds. Thus, the annotated sample suggests that Maguire, Wisniewski, and Storms' (2010) strategy might only work for a small number of compounds that reflect specific semantic category combinations. Thirdly – and this is not a problem specific to Maguire, Wisniewski, and Storms' (2010) approach – it is unclear how many compounds can be interpreted context-free and based on established interpretation patterns at all. Context-free interpretation is sometimes easy with lexicalized compounds (*Basaltfelsen*, 'basalt'–'rock', (NATURAL) OBJECT | (NATURAL) OBJECT, H_CONSISTS_OF_M) as well as with non-lexicalized hapax legomena (*Korianderlikör*, 'coriander'–'liqueur', FOOD | FOOD, H_CONTAINS_M). However, the annotation practice showed that assigning context-free interpretations to compounds is often difficult for several reasons: (i) Sometimes, two or more interpretation patterns can be equally plausible (*Majolikaofen*, 'majolica'–'oven', ARTIFACT | ARTIFACT, either M_IS_PART_OF_H, 'oven decorated with majolica tiles', or H_PRODUCES_M, 'oven used to burn majolica tiles'). (ii) A lexicalized compound can be opaque (*Rattenschwanz*, 'rat'–'tail', ANIMAL | BODY, usually means 'string of unpleasant things that occur one after the other', e.g., *ein Rattenschwanz an Problemen*, 'a string of problems that occur one after the other'). (iii) A lexicalized compound can have a specialized meaning such that establishing the relation between head and modifier requires specialist knowledge (*Keupermergel*, 'Keuper'–'marl', NATURAL_OBJECT | NATURAL_OBJECT, 'marl as a lime-rich mudstone that contributes as sediment to the Keuper as one of the lithostratigraphic units in Central Europe'). (iv) A non-lexicalized compound can be dependent on its context (*Flohgrippe*, 'flea'–'influenza', ironic reference to 'a possible illness that killed all the fleas in a flea circus'). If we had not restricted our investigation to compounds with monosemous constituents, we would probably have encountered even more problems.

The nature of the GermaNet categories exerted a considerable influence on our results. This is to a large extent due to the unclear semantic status of the categories.[18] Some categories, such as ANIMAL, establish a hyperonymic relation to almost all of their members: *Tier* ('animal') is a hyperonym of *Luchs* ('lynx'), *Hase* ('hare'), and *Krokodil* ('crocodile'). Other categories,

18 Cf. for a discussion of similar problems with ontological systems Engelberg & Meyer (2015: 150–161).

e.g., COGNITION, constitute something more like a thematic field, to which member words such as *Stil* ('style'), *Jazz* ('jazz'), and *Theorie* ('theory') are assigned. Still other categories, such as FEELING, are partly hyperonymic (*Wut*, 'rage') and partly thematic (*Spott*, 'mockery') in nature, or – like BODY – partly meronymic (*Kopf*, 'head') and partly thematic (*Schweiß*, 'sweat').

The fact that most categories are more of a thematic than a taxonomic nature accounts partly for the observation that combinations of categories of the same semantic group rank very high in table 7, something that was also observed by Maguire, Wisniewski, and Storms (2010), who had five such combinations among their ten top-rated combinations: Concept formation by compounding often requires words from the same thematic field. However, in order to map categories onto the selectional restrictions for slots in interpretation relations, we would fare much better with hyperonymical relations. Nevertheless, even in a non-thematic, taxonomical ontology, a slightly more differentiated categorization might have improved the interpretation process considerably. For example, subcategorizing the FOOD domain into the two classes of words that dominate this category, namely words for natural edible objects (*Rosine*, 'raisin') and words for dishes and products of cooking and baking (*Stollen*, 'stollen'), would have allowed us to distinguish the frequent Y-CONTAINS-X interpretation (*Rosinenstollen*, 'raisin'–'stollen') from the Y-IS-(SUITABLE)-FOR-X interpretation (*Stollenrosine*, 'stollen'–'raisin').

In conclusion, Study E was mainly motivated by the considerations in section 3 that interpretation patterns impose selectional restrictions on their head and their modifier argument and thereby account for the distributional properties of constituent types. We combined this idea with the assumption that interpretation patterns can to a large degree be deduced from the combination of the semantic categories of head and modifier. With respect to this second assumption, Study E remains somewhat inconclusive.

9. Conclusion

In this paper we examined possible factors that govern the preference of constituent words for either the head or the modifier position in German N-N compounds empirically on the basis of a large dataset.

We found (i) that words with large constituent families tend towards a balanced distribution with respect to head and modifier position, (ii) that monosemous constituent words show a preference for the modifier position while polysemous words tend towards head position, and (iii) that words of some semantic classes (SUBSTANCE, (NATURAL) OBJECT, TIME, LOCATION, BODY, FOOD, ANIMAL) show a statistically significant propensity for modifier position while ARTIFACT shows a propensity for head position. We also took a closer look at how constituents from specific semantic classes combine and

we explored a hypothesis put forward by Maguire, Wisniewski, and Storms (2010) that knowing the semantic categories of the two constituents helps to determine the interpretation pattern the compound is based on and determines the assignment of constituent words to constituent positions. Though we could confirm that some types of semantic category combinations show preferences for specific semantic interpretation patterns, our general findings regarding this hypothesis remained inconclusive, mainly due to the nature of the semantic classifications obtained from GermaNet. A more strictly hyperonymic categorization might have helped to explore the true limits of this strategy.

The fact that our studies are based on a very large database forced us to accept a certain amount of error in the automatic and semi-automatic categorizations, and the annotation using GermaNet categories was a rather rough approximation of semantic classification. At the same time, we were able to observe trends in a database larger than any of the studies conducted before, and thus give a bird's eye view of German compounding. As in our earlier study, which used a smaller excerpt from the same compound database (Hein and Brunner 2020), we believe that studying quantitative trends on the basis of large datasets can lead to interesting new insights.

Of course, this paper could only address a few of the many possible factors that might influence head-modifier distribution. We did not consider the phonological properties of the constituents, and the influence of their morphology was purposely excluded by focusing exclusively on simplex-simplex compounds. Another interesting aspect that was not taken into account in this paper is compound token frequency; like Tarasova (2019) or Maguire, Wisniewski, and Storms (2010), we only looked at compound types. Token frequencies might play an important role in the question of which interpretation patterns become entrenched, which in turn probably influences modifier-head preferences for certain compounds.

References

Adams, Valerie. 1973. *An Introduction to Modern English Word-Formation*. London: Longman.

Baayen, Harald. 2010. The directed compound graph of English: An exploration of lexical connectivity and its processing consequences. In Susan Olsen (ed.), New Impulses in Wordformation. [Special issue]. *Linguistische Berichte*, Sonderheft 17. 383–402.

Bauer, Laurie, Natalia Beliaeva & Elizaveta Tarasova. 2019. Recalibrating Productivity: Factors Involved. *Zeitschrift für Wortbildung / Journal of Word Formation* 3(1). 44–80. https://www.ingentaconnect.com/contentone/plg/jwf/2019/00000003/00000001/art00003# (accessed 14 February 2020).

Brekle, Herbert Ernst. 1976. *Generative Satzsemantik im System der englischen Nominalkomposition*. München: Fink.

Bubenhofer, Noah, Marek Konopka & Roman Schneider (eds.). 2014. *Preliminarien einer Korpusgrammatik*. Tübingen: Narr.

Engelberg, Stefan, & Peter Meyer. 2015. Das Lehnwortportal Deutsch als kontaktlinguistisches Forschungsinstrument. In Emmerich Kelih, Jürgen Fuchsbauer & Stefan Michael Newerkla (eds.), *Lehnwörter im Slawischen. Empirische und crosslinguistische Perspektiven*, 149–170. Frankfurt am Main.: Lang.

Fandrych, Christian, & Maria Thurmair. 1994. Ein Interpretationsmodell für Nominalkomposita: linguistische und didaktische Überlegungen. *Deutsch als Fremdsprache* 31. 34–45.

Fanselow, Gisbert. 1981. *Zur Syntax und Semantik der Nominalkomposition. Ein Versuch praktischer Anwendung der Montague-Grammatik auf die Wortbildung im Deutschen*. Tübingen: Niemeyer.

Fleischer, Wolfgang. 1982. *Phraseologie der deutschen Gegenwartssprache*. Leipzig: Bibliograph. Inst.

Fleischer, Wolfgang & Irmhild Barz. 2012. *Wortbildung der deutschen Gegenwartssprache*. 4., völlig neu bearbeitete Auflage. Berlin: De Gruyter.

Gagné, Christina L., & Edward J. Shoben. 1997. Influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 23. 71–87.

Hamp, Birgit & Helmut Feldweg. 1997. GermaNet – a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications. Madrid 1997*. 9–15.

Hatcher, Anna Granville. 1960. An Introduction to the Analysis of English Noun Compounds. *Word* 16. 356–373.

Hein, Katrin. 2015. *Phrasenkomposita im Deutschen. Empirische Untersuchung und konstruktionsgrammatische Modellierung*. Tübingen: Narr Francke Attempto.

Hein, Katrin & Annelen Brunner. 2020. Why do some lexemes combine more frequently than others? – An empirical approach to productivity in German compound formation. In Jenny Audring, Nikos Koutsoukos & Christina Manouilidou (eds.), *Rules, Patterns, Schemas and Analogy. Online Proceedings of the 12th Mediterranean Morphology Meeting (MMM12), Ljubljana (Slovenia), June 27–30, 2019*, 28–41. Patras: University of Patras.

Henrich, Verena & Erhard Hinrichs. 2010. GernEdiT – The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 2228–2235. Valletta, Malta.

Henzen, Walter. 1947. *Deutsche Wortbildung*. Halle an der Saale: Niemeyer.

Jackendoff, Ray. 2009. Compounding in the parallel architecture and conceptual semantics. In Rochelle Lieber & Pavol Stekauer (eds.), *The Oxford handbook of compounding*, 105–128. Oxford: Oxford University Press.

Kürschner, Wilfried. 1974. *Zur syntaktischen Beschreibung deutscher Nominalkomposita. Auf der Grundlage generativer Transformationsgrammatiken*. Tübingen: Niemeyer.

Lees, Robert B. 1960. *The Grammar of English Nominalizations*. The Hague: Mouton.

Levi, Judith N. 1978. *The Syntax and Semantics of Complex Nominals*. New York, San Francisco & London: Academic Press.

Libben, Gary. 2010. Compound Words, Semantic Transparency, and Morphological Transcendence. In Susan Olsen (ed.), New Impulses in Word-Formation. [Special issue]. *Linguistische Berichte*, Sonderheft 17. 317–330.

Mätzner, Eduard.1860. *Englische Grammatik. Erster Theil: Die Lehre vom Worte*. Berlin: Weidmannsche Buchhandlung.

Maguire, Phil, Edward Wisniewski & Gert Storms. 2010. A corpus study of semantic patterns in compounding. *Corpus Linguistics & Linguistic Theory* 6. 49–73.

Meyer, Ralf. 1993. *Compound Comprehension in Isolation and in Context*. Tübingen: Niemeyer.

Motsch, Wolfgang. 1999. *Deutsche Wortbildung in Grundzügen*. Berlin & New York: De Gruyter.

Murphy, Greg L. 1988. Comprehending complex concepts. *Cognitive Science* 12. 529–562.

Olsen, Susan. 2012a. Der Einfluss des Mentalen Lexikons auf die Interpretation von Komposita. In Livio Gaeta & Barbara Schlücker (eds.), *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte*, 135–170. Berlin & Boston: De Gruyter.

Olsen, Susan, 2012b. Semantics of Compounds. In Claudia Maienborn, Klaus von Heusinger & Paul Portner (eds.), *Semantics. An International Handbook of Natural Language Meaning*, vol. 3, 2120–2150. Berlin, New York & Boston: De Gruyter.

Olsen, Susan. 2015. Composition. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation. An International Handbook of the Languages of Europe*. Vol. 1, 364–386. Berlin & Boston: De Gruyter.

Ortner, Lorelies, Elgin Müller-Bollhagen, Hanspeter Ortner, Hans Wellmann, Maria Pümpel-Mader & Hildegard Gärtner. 1991. *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache. Eine Bestandsaufnahme des Instituts für Deutsche Sprache, Forschungsstelle Innsbruck. Vierter Hauptteil: Substantivkomposita (Komposita und Kompositionsähnliche Strukturen 1)*. Berlin & New York: De Gruyter.

Ortner, Lorelies & Hanspeter Ortner. 2015. Schemata and semantic roles in word-formation. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-Formation. An International Handbook of the Languages of Europe*, Vol. 2, 1035–1056. Berlin, Boston: De Gruyter Mouton.

Paul, Hermann. 1920. *Deutsche Grammatik. Band V, Teil IV: Wortbildungslehre*. Halle a. S.: Niemeyer.

Piantadosi, Steven T. 2014. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review* 21. 1112–1130.

Roth, Tobias. 2015. Kompositum oder Kollokation? Konkurrenz an der Syntax-Morphologie-Schnittstelle. In Regula Schmidlin, Heike Behrens & Hans Bickel (eds.), *Sprachgebrauch und Sprachbewusstsein: Implikationen für die Sprachtheorie*, 155–176. Berlin, Boston: De Gruyter.

Schlücker, Barbara. 2012. Die deutsche Kompositionsfreudigkeit. Übersicht und Einführung. In Livio Gaeta & Barbara Schlücker (eds.), *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte*, 1–25. Berlin: De Gruyter.

Spalding, Thomas L., Christina L. Gagné, Allison Mullaly & Hongbo Ji. 2010. Relation-Based Interpretation of Noun-Noun Phrases: A New Theoretical Approach. In Susan Olsen (ed.), New Impulses in Wordformation. [Special issue]. *Linguistische Berichte Sonderheft* 17. 283–315.

Tarasova, Elizaveta. 2013. *Some new insights into the semantics of English N+N compounds*. Wellington: Victoria University dissertation.

Tarasova, Elizaveta. 2019. Productivity of form and productivity of meaning in N+N compounds. In Vesna Kalafus Antoniová, Sandra Jiménez-Pareja & Alba E. Ruz (eds.), Selected papers from the Word-Formation Theories III & Typology and Universals in Word-Formation IV Conference, Košice, Slovakia, 27–30 June, 2018. [Special issue]. *SKASE Journal of Theoretical Linguistics* 16. 49–69. http://www.skase.sk/Volumes/JTL39/pdf_doc/04.pdf (accessed 14 February 2020).

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

Wisniewski, Edward J. 1997. When concepts combine. *Psychonomic Bulletin and Review* 4. 167–183.

Annelen Brunner
Leibniz-Institut für Deutsche Sprache
R5 6-13
D-68161 Mannheim
brunner@ids-mannheim.de

Stefan Engelberg
Leibniz-Institut für Deutsche Sprache
R5 6-13
D-68161 Mannheim
engelberg@ids-mannheim.de

Katrin Hein
Leibniz-Institut für Deutsche Sprache
R5 6-13
D-68161 Mannheim
hein@ids-mannheim.de