Chapter 8

Collecting and investigating features of compositionality ratings

Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart, Germany

Developing computational models to predict degrees of compositionality for multiword expressions typically goes hand in hand with creating or using reliable lexical resources as gold standards for formative intrinsic evaluation. Not much work however has looked into whether and how much both the gold standards and the computational prediction models vary according to the properties of the compounds within the lexical resources. In the current study, we focus on English and German noun compounds and suggest a novel route to assess the interactions between compound and constituent properties with regard to the compounds' degrees of compositionality. Our contributions are two-fold: (1) a novel collection of compositionality ratings for 1,099 German noun compounds, where we asked the human judges to provide compound and constituent properties (such as paraphrases, meaning contributions, hypernymy relations, and concreteness) before judging the compositionality; and (2) a series of analyses on rating distributions and interactions with compound and constituent properties for our novel collection as well as existing gold standard resources in English and German. Following the analyses we discuss to what extent one should aim for an even distribution of ratings across the pre-specified scale, and to what extent one should take into account properties of the compound and constituent targets when creating a novel resource and when using a resource for evaluation. We suggest as a minimum requirement to balance targets across frequency ranges, and optimally to balance targets across their most salient properties in a post-collection filtering step. Above all, we recommend to assess computational models not only on the full dataset but also with regard to subsets of targets with coherent task-relevant properties.

Sabine Schulte im Walde. 2024. Collecting and investigating features of compositionality ratings. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 269–308. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998645

1 Motivation

Combinations of words are considered multiword expressions (MWEs) in the field of natural language processing (NLP), if they are semantically idiosyncratic to some degree, i.e., the meaning of the combination is not entirely (or even not at all) predictable from the meanings of the constituents (Sag et al. 2002, Baldwin & Kim 2010, Savary et al. 2018). Hence, computational modelling of MWEs has been a long-standing task and is important for both theoretical and applied research, in order to investigate multiword expressions from a large-scale, empirical perspective, and to integrate the compositionality models into NLP applications that require natural language understanding (NLU), such as domain-specific interpretation (Clouet & Daille 2014, Hätty & Schulte im Walde 2018, Hätty et al. 2019, Bettinger et al. 2020, Hätty et al. 2021, Eichel et al. 2023) and machine translation (Carpuat & Diab 2010, Cholakov & Kordoni 2014, Weller et al. 2014, Cap et al. 2015, Salehi et al. 2015b, Gamallo et al. 2019, Dankers et al. 2022).

In the current study, the focus of interest is on noun compounds, such as *climate change* and *crocodile tears* in English, and *Ahornblatt* 'maple leaf' and *Fliegenpilz* 'toadstool' in German. The representation, processing and modelling of noun compounds has previously received an immense attention across disciplines and languages, e.g., regarding the theoretical definition of compoundhood, typologies of compounds, structural properties of compounds, and compound and constituent meanings (Levi 1978, Plag 2003, Bauer 2017, Schulte im Walde & Smolka 2020, i.a.); regarding the question whether compounds are stored in the mental lexicon and processed as units, via their constituents, or via a dual route (Taft & Forster 1975, Butterworth 1983, i.a.); regarding conceptual combinations of modifiers and heads (Murphy 1990, Wisniewski 1996, Costello & Keane 2000, Benczes 2014, i.a.); regarding the role of compound relations (Gagné 2002, Nastase 2003, Girju et al. 2005, Spalding et al. 2010, i.a.); regarding association and feature norms of noun compounds and their constituents (Roller & Schulte im Walde 2014, Schulte im Walde & Borgwaldt 2015, i.a.); etc.

Standard computational approaches define and compare corpus-based representations of compounds and their constituents, in order to compute the degrees of semantic relatedness as a basis for predicting the degrees of compositionality of the compounds; for example, the representation of a compound such as *climate change* is supposedly more similar to the representations (or a combination of the representations) of the constituents *climate* and *change* than the representation of a more semantically idiosyncratic compound such as *crocodile tears* would be. Such distributional models are rather successful and obtain correlations of $\rho \approx 0.7$ when evaluated against gold standard resources.

Developing computational models of compositionality typically goes hand in hand with creating reliable lexical resources as gold standards for formative intrinsic evaluation. Accordingly, we find datasets of noun compounds with ratings on compositionality across languages, such as English (Reddy et al. 2011b, Cordeiro et al. 2019), German (Schulte im Walde et al. 2013, 2016b), and French and Portuguese (Cordeiro et al. 2019). Not much work however has looked into whether and how much both the gold standards and the prediction models vary according to properties of the targets within the lexical resources. For example, what are the empirical, corpus-based properties of the noun compound targets, such as frequencies and constituent productivities? What are their lexical-semantic properties, such as degrees of ambiguity and concreteness? And how do these properties interact with the compounds' degrees of compositionality? The distributions of target properties and compositionality ratings differ across compound datasets, and potential skewness hinders us from a generalised assessment of prediction models. I.e., does a system's correlation of $\rho \approx 0.7$ hold across targets and target properties, or is this merely an average result and therefore opaque regarding any gold standard subsets? As to our knowledge, up to date only a few computational studies on noun compounds have described the variance of prediction results across compound and constituent properties (Schulte im Walde et al. 2016a, Köper & Schulte im Walde 2017, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023), thus pointing out the need for a more systematic investigation.

The current study suggests a novel route to assess the interactions of compound and constituent properties with regard to the compounds' degrees of compositionality, which we consider as indispensible ground knowledge when interpreting the results of computational models. We provide two contributions to move forward both theoretical and computational investigations of compositionality for noun compounds:

(1) We created a *novel collection of compositionality ratings for 1,099 German noun compounds* where – differently to previous related work – we asked the human judges to provide (a) paraphrases of the compounds' meanings, (b) constituent features contributing to the compounds meanings, (c) judgements on the hypernymy relations between the compounds and their head constituents, and (d) judgements on the concreteness of the compounds and constituents, before they provided their judgements on the compounds' degree of compositionality with regard to the respective constituents. The elaborate information enables us to relate compositionality judgements to a range of compound and constituent properties.

(2) We present a *series of analyses* on (a) distributions of compositionality ratings, and (b) relations between compositionality ratings and compound and constituent properties (such as frequency, productivity, ambiguity, hypernymy and concreteness). Next to relying on our own novel collection as basis for our study, we also make use of the predominantly used lexical resources of noun compound compositionality for English (Reddy et al. 2011b, Cordeiro et al. 2019) and German (Schulte im Walde et al. 2013, 2016b), and exploit web corpora for the same two languages (Baroni et al. 2009, Schäfer & Bildhauer 2012).

Based on our insights from (1) and (2), we then discuss distributions of compositionality ratings across resources, and to what extent (and how) one should take into account properties of targets when creating a novel resource, and when using a resource in the evaluation of computational models.

In the remainder of this article, Section §2 presents an overview of existing lexical resources with compositionality ratings for noun compounds, as well as standard computational prediction models across languages. In Section §3, our article introduces the creation of the novel gold standard of German compositionality ratings, before we dive into analyses and discussions of rating distributions and rating properties in Section §4.

2 Previous work on compositionality datasets and models

As a starting point for discussing the interactions and potential strategies for optimisations of gold-standard compositionality ratings, we provide an overview of the predominantly used English and German datasets (Section §2.1) and approaches towards predicting degrees of compositionality (Section §2.2).

2.1 Datasets of compositionality ratings

Reddy et al. (2011b) created the probably first dataset with compositionality judgements for noun compounds that were explicitly collected as gold standard ratings to evaluate computational models of compositionality. Henceforth, we will refer to this dataset as Reddy-NN. For the Reddy-NN dataset, Reddy et al. (2011b) selected 90 English noun compounds with two simplex noun constituents. The compound target construction was done such that Reddy et al. distinguished between four classes of modifier and head combinations regarding the constituents'

contributions to the compound meanings,[1] based on heuristics using relations and definitions in WordNet (Fellbaum 1998): a compound was considered compositional with regard to a constituent if it either represented a hyponym of that constituent (e.g., a *swimming pool* is a *pool*), or if the constituent occurred in its definition (e.g., *swimming* occurs in the definition of a *swimming pool*). Then Reddy et al. asked 30 annotators via Amazon Mechanical Turk (AMT) to provide judgements on compositionality ratings for the compound as a whole (which they refer to as "phrase compositionality"), and for the strengths of meaning contributions of the constituents, all on a scale $[0, 5]$ from 0 (clearly non-compositional) to 5 (clearly compositional). The upper part in Table 1 provides a selection of examples from the Reddy-NN target compounds, together with the mean compositionality ratings across the raters and the respective standard deviations. The basic dataset was subsequently extended in various respects: Bell & Schäfer (2013) added semantic relations; Schulte im Walde et al. (2016a) added frequencies and scores for productivity and ambiguity; and Cordeiro et al. (2019), henceforth Cordeiro-N, extended the dataset to 280 English noun compounds, however varying the modifier word class, and then following the same rating procedure as Reddy et al. (2011b). In our own work, we created two datasets of German noun compounds:

(1) In Schulte im Walde et al. (2013), we presented a set of 244 German noun–noun compounds with two simplex nominal constituents, based on a larger set of 450 *concrete* noun compounds from von der Heide & Borgwaldt (2009), who had collected compound–constituent compositionality ratings from 30 annotators in a paper-and-pen annotation. We collected and added to the resource between 27–34 compositionality ratings via AMT for the compound as a whole. All ratings were collected on a scale $[1, 7]$ from 0 (clearly non-compositional) to 7 (clearly compositional). Henceforth, we will refer to this dataset as Concrete-NN. The lower part of Table 1 provides a selection of examples, together with mean compositionality ratings and standard deviations. The basic dataset was subsequently extended by Schulte im Walde et al. (2016a), who added frequencies and scores for productivity and ambiguity; and Schulte im Walde & Borgwaldt (2015), who compiled and analysed association norms for the concrete compounds and their constituents.

[1]As to our knowledge, Libben and his colleagues (Libben et al. 1997, 2003) were the first in psycholinguistics research who systematically categorised noun-noun compounds with nominal modifiers and heads into four groups representing all possible combinations of modifier and head transparency (T) vs. opaqueness (O) within a compound. Examples for these categories were *car-wash* (TT), *strawberry* (OT), *jailbird* (TO), and *hogwash* (OO).

Table 1: Example compounds from REDDY-NN and CONCRETE-NN, with mean compositionality ratings and standard deviations. The compound column refers to compound phrase/whole ratings; the modifier and head columns refer to compound–modifier and compound–head ratings, respectively. Note that the collections use different scales: $[0, 5]$ in REDDY-NN and $[1, 7]$ in CONCRETE-NN.

	Mean ratings and std. dev.		
Compounds	compound	modifier	head
cheat sheet	2.89 ± 1.11	2.30 ± 1.59	4.00 ± 0.83
climate change	4.97 ± 0.18	4.90 ± 0.30	4.83 ± 0.38
couch potato	1.41 ± 1.03	3.27 ± 1.48	0.34 ± 0.66
crocodile tears	1.25 ± 1.09	0.19 ± 0.47	3.79 ± 1.05
diamond wedding	1.70 ± 1.05	0.78 ± 1.29	3.41 ± 1.34
guilt trip	2.19 ± 1.16	4.71 ± 0.59	0.86 ± 0.94
melting pot	0.54 ± 0.63	1.00 ± 1.15	0.48 ± 0.63
night owl	1.93 ± 1.27	4.47 ± 0.88	0.50 ± 0.82
polo shirt	3.37 ± 1.38	1.73 ± 1.41	5.00 ± 0.00
search engine	3.32 ± 1.16	4.62 ± 0.96	2.25 ± 1.70
Ahornblatt 'maple leaf'	6.03 ± 1.49	5.64 ± 1.63	5.71 ± 1.70
Feuerzeug (lit. 'fire stuff') 'lighter'	4.58 ± 1.75	5.87 ± 1.01	1.90 ± 1.03
Fleischwolf (lit. 'meat wolf') 'meat grinder'	1.70 ± 1.05	6.00 ± 1.44	1.90 ± 1.42
Fliegenpilz (lit. 'fly mushroom') 'fly agaric'	2.00 ± 1.20	1.93 ± 1.28	6.55 ± 0.63
Flohmarkt 'flea market'	2.31 ± 1.65	1.50 ± 1.22	6.03 ± 1.50
Löwenzahn (lit. 'lion tooth') 'dandelion'	1.66 ± 1.54	2.10 ± 1.84	2.23 ± 1.92
Maulwurf (lit. 'mouth throw') 'mole'	1.58 ± 1.43	2.21 ± 1.68	2.76 ± 2.10
Postbote (lit. 'mail messenger') 'post man'	6.33 ± 0.96	5.87 ± 1.55	5.10 ± 1.99
Seezunge (lit. 'sea tongue') 'sole'	1.85 ± 1.28	3.57 ± 2.42	3.27 ± 2.32
Windlicht (lit. 'wind light') 'lantern'	3.52 ± 2.08	3.07 ± 2.12	4.27 ± 2.36

(2) In Schulte im Walde et al. (2016b), we presented a dataset of German noun–noun compounds with two simplex nominal constituents. As to our knowledge, this dataset was the first that took properties of the compounds and the constituents into account during the selection of the targets: we induced a balanced set of 180 compounds with low/mid/high modifier productivity and low/mid/high head ambiguity (which we determined as the two most important balancing criteria) from a candidate compound set containing ≈ 150,000 noun-noun compounds occurring in a large web corpus (Schäfer & Bildhauer 2012). We also created an extended set of 868 compounds by systematically adding all compounds from the original candidate set with either the same modifier or the same head as any of the

compounds in the balanced set. For example, given the compound *Geduldspiel* 'puzzle' in the balanced set of compounds we added all compounds from the original candidate set with the modifier *Geduld* 'patience', and all compounds with the head *Spiel* 'game'. We then collected between 8–13 compound–constituent compositionality ratings via AMT, on a scale $[0, 6]$ from 0 (clearly non-compositional) to 6 (clearly compositional). Henceforth, we will refer to the two balanced/unbalanced versions of the dataset containing 180/868 noun–noun compounds as GHOST-NN/S and GHOST-NN/XL, in the same way as in Schulte im Walde et al. (2016a).

Table 2 provides a selection of examples, together with empirical and lexical compound and constituent properties, and mean compositionality ratings. The examples include compounds with the modifiers *Stadt* 'city' and *Sonne* 'sun' as well as compounds with the heads *Spiel* 'game' and *Kette* 'chain'. The corresponding properties are corpus frequencies for the compounds, modifiers and heads, as well as productivity and ambiguity scores for the constituents, relying on morphological family size (de Jong et al. 2002) and the number of senses defined in GermaNet (Hamp & Feldweg 1997, Kunze 2000), respectively. Semantic relations between modifiers and heads (e.g., in *Machtspiel* 'power game', the game is ABOUT power; in *Kartenspiel* 'card game', the cards represent the INSTRUMENT in the game) were annotated by the four authors of the paper, adopting the scheme by Ó Séaghdha (2007) using four relations defined by Levi (1978): BE, HAVE, IN, ABOUT; two relations referring to event participants (ACTOR, INST(rument)), and LEX indicating lexicalised compounds.

Overall, the described datasets REDDY-NN and CORDEIRO-N for English as well as CONCRETE-NN and GHOST-NN for German were created on different grounds for target compound selection, i.e., WordNet relations (REDDY-NN and CORDEIRO-N), concreteness (CONCRETE-NN), and partial balancing across empirical and lexical properties (GHOST-NN). The actual collection of human ratings was done similarly across datasets, while varying between paper-and-pen and crowdsourcing as well as the rating scales.

Figure 1 however presents a rather diverse picture regarding the distributions of compositionality ratings across the respective collection ranges. The boxplots show the four quartiles of the rating distributions, with the median lines in the boxes of the interquartile ranges, and the dots referring to outliers. Green boxes refer to compound ratings, blue/red boxes to compound–modifier and compound–head ratings, respectively. For the compound–constituent ratings in

Table 2: Examples of compounds from GHOST-NN/XL with empirical and lexical properties, and mean compositionality ratings.

Compounds	Relation	Frequencies			Productivities		Ambiguities		Ratings	
		compound	modifier	head	modifier	head	modifier	head	modifier	head
Stadthotel 'city hotel'	IN	3,405	4,053,206	1,199,856	543	59	1	1	3.35	5.35
Stadtrand (lit. 'city border') 'suburb'	HAVE	25,099	4,053,206	523,473	543	98	1	2	4.94	4.25
Stadtwerk (lit. 'city plant') 'public services'	ACTOR	107,754	4,053,206	1,354,148	543	366	1	6	3.81	3.69
Sonnenenergie 'solar energy'	INST	25,398	832,636	1,191,333	155	30	3	2	4.58	5.44
Sonnenkönig 'Sun King'	LEX	2,680	832,636	494,221	155	109	3	3	1.94	5.50
Sonnenmasse 'sun mass'	HAVE	3,433	832,636	468,284	155	108	3	3	4.56	4.75
Sonnenscheibe 'solar disc'	BE	3,155	832,636	364,567	155	96	3	4	4.56	3.75
Sonnenseite 'sunny side'	IN	7,279	832,636	5,508,445	155	256	3	6	4.00	4.31
Sonnenstrahl 'sun beam'	HAVE	44,612	832,636	32,182	155	27	3	3	5.13	4.69
Sonnenuhr (lit. 'sun clock') 'sundial'	INST	8,407	832,636	4,507,590	155	63	3	2	3.75	5.31
Kirchspiel (lit. 'church game') 'parish'	LEX	6,583	1,761,187	4,122,168	319	403	3	6	4.44	3.13
Machtspiel 'power game'	ABOUT	4,408	806,162	4,122,168	169	403	2	6	4.63	3.44
Testspiel (lit. 'test game') 'tryout'	BE	37,800	660,169	4,122,168	100	403	3	6	4.25	5.19
Trauerspiel (lit. 'mourning game') 'fiasco'	ABOUT	10,763	134,379	4,122,168	77	403	3	6	3.06	2.81
Windspiel (lit. 'wind game') 'wind chimes'	INST	2,284	551,317	4,122,168	88	403	3	6	4.31	2.94
Würfelspiel 'dice game'	INST	4,408	80,371	4,122,168	14	403	2	6	4.94	5.56
Bergkette 'mountain chain'	BE	8,799	564,178	207,479	205	139	2	4	5.13	2.56
Halskette (lit. 'neck chain') 'necklace'	IN	8,707	271,703	207,479	39	139	3	4	3.94	5.44
Handelskette 'trade chain'	INST	6,509	428,611	207,479	240	139	1	4	4.75	3.38
Hotelkette 'hotel chain'	BE	6,410	1,199,856	207,479	134	139	1	4	5.00	3.13
Menschenkette 'human chain'	BE	6,383	8,884,087	207,479	191	139	1	4	4.94	3.75
Produktionskette 'production chain'	HAVE	2,738	579,419	207,479	244	139	2	4	4.69	3.19
Schneekette 'snow chain'	INST	5,167	324,839	207,479	95	139	1	4	4.19	4.21
Zeichenkette (lit. 'character chain') 'string'	BE	8,836	749,903	207,479	62	139	3	4	4.34	2.95

the GHOST-NN variants, 75% of the mean ratings are in the range [4, 6], and the medians are between 4 and 5. CONCRETE-NN is less skewed, but still 75% of all ratings are in the range [3.5, 7]. Only REDDY-NN and the extension CORDEIRO-N (plots for the latter are in the appendix because they follow similar trends as REDDY-NN) cover a wide range of compositionality ratings. In the next section we will ask whether and how the skewness of the compounds' degrees of compositionality influences the reliability of predictions by computational models.

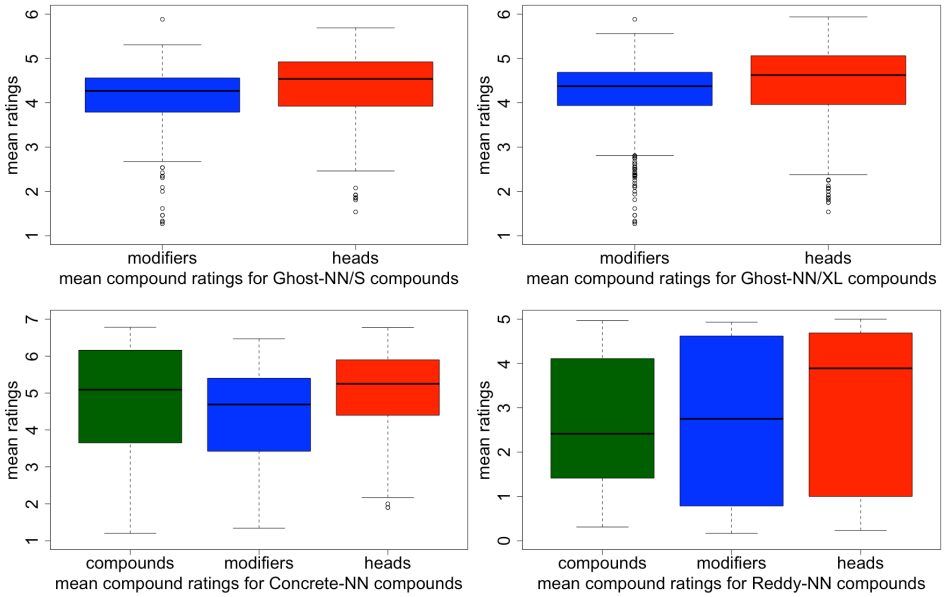

Figure 1: Compositionality rating distributions across rating datasets.

2.2 Compositionality prediction models

As introduced above, standard computational approaches define and compare corpus-based representations of compounds and their constituents, in order to determine the degree of semantic relatedness as a basis for predicting the degree of compositionality of the compounds. Existing models generally rely on the distributional hypothesis that the context of a linguistic unit contains indicators for the unit's usage and meaning (Harris 1954, Firth 1957), and thus exploit and represent corpus-based cooccurrences induced from large-scale corpora of the respective language, in combination with mathematical measures of similarity when comparing the representations. The most traditional approaches rely on

distributional count vector spaces, either using window-based or syntax-based cooccurrences (Reddy et al. 2011b,a, Schulte im Walde et al. 2013, 2016a), while later approaches use embeddings as representations (Salehi et al. 2015a, Cordeiro et al. 2019, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023). The work by Salehi combined corpus-based textual information with dictionary information (Salehi et al. 2014a) and integrated translation knowledge (Salehi & Cook 2013, Salehi et al. 2014b), and the work in our group extended textual to multimodal approaches (Roller & Schulte im Walde 2013, Köper & Schulte im Walde 2017). While most approaches were directly applied to type-level representations, Bott & Schulte im Walde (2017) applied soft clustering to access the sense level, and Miletic & Schulte im Walde (2023) compared token- and type-level BERT representation layers. The actual predictions of degrees of compositionality then compare the respective representations by computing the cosine distance (or other vector-based distance measures) between vector representations of compounds and vector representations of constituents, or apply composite functions to the vectors of the constituents (such as vector multiplication) before computing the similarity with the compound vector (Mitchell & Lapata 2010, Reddy et al. 2011b, Hermann 2014, Dima et al. 2019, Alipoor & Schulte im Walde 2020, i.a.).

While the exact details of the various approaches are not relevant to the current study, we would like to point out that the majority of approaches predicted the degrees of compositionality across all compound and constituent targets of the respective datasets, i.e., disregarding target subsets and potential influences of such subsets on the prediction. As such, existing compositionality prediction models have overall proven very successful, obtaining Spearman's rank-order correlation coefficients (Siegel & Castellan 1988) of $\rho \approx 0.7$ when evaluated against the gold standard datasets. In the following we present three studies demonstrating that the results differ, however, when compound and constituent properties are taken into account in the evaluation of the models.

Schulte im Walde et al. (2016a) implemented a standard window-based vector space model relying on cooccurrence in a sentence-internal window of ± 20 words, and predicted degrees of compositionality based on the cosine distance measure. For evaluation they used Reddy-NN, Concrete-NN and Ghost-NN as well as an English noun compound dataset with semantic relations by Ó Séaghdha (2007). In a preparatory effort, they extended the datasets such that information on compound and constituent frequency, constituent productivity, compound and constituent ambiguity, and semantic relations was available for all English and German resources. Coccurrences, frequencies and productivities were induced from the respective COW corpora (Schäfer & Bildhauer 2012, Schäfer

2015); ambiguities from WordNet/GermaNet (Fellbaum 1998, Hamp & Feldweg 1997, Kunze 2000), and semantic relations and compositionality ratings were annotated, if not available. Crucially, Schulte im Walde et al. (2016a) then ran their prediction models on all targets within the respective datasets, but also on subsets of targets with extreme properties, such as the least and the most frequent compounds, the least and the most productive constituents, by relation type, etc. Their results showed that – among other insights – the same models make overall better predictions for (i) more frequent compounds, and for (ii) compounds with less frequent, less productive and less ambiguous heads, while (iii) the modifier properties did not have a consistent effect.

In a similar vein, Alipoor & Schulte im Walde (2020) implemented a standard window-based vector space model and word2vec embeddings for English, relying on a sentence-internal window of ±10 words in the English COW corpus. They focused on the effect of various kinds of dimensionality reductions on compositionality prediction, and they also zoomed into compound subsets regarding compound and constituent properties. Similarly to the study by Schulte im Walde et al. (2016a), they found that in most vector-space variants the predictions (i) were better for mid-/high-frequency compounds in comparison to low-frequency compounds, and (ii) did not behave in a consistent way for modifier properties; but in contrast to the previous work, their predictions were (iii) better for compounds with mid-/high-productivity than low-productivity heads. In addition, they looked into the effect of target compositionality, and found that predictions were (iv) generally better for mid-/high-compositional than low-compositional compound–constituent combinations. Miletic & Schulte im Walde (2023) also zoomed into the influence of frequencies, productivities and ambiguities in our study regarding BERT representation layers. Focusing on head properties of the Cordeiro-N compounds, we found better compositionality predictions for low-frequency, low-productivity, and low-ambiguity heads across compound and compound–constituent rating predictions.

Finally, Köper & Schulte im Walde (2017) compared multimodal models combining textual and visual vector spaces when predicting degrees of compositionality for German noun compounds and particle verbs. They zoomed into the effects of constituent properties: frequency, ambiguity, concreteness, imageability and compositionality. As in previous work, they did not find consistent effects of modifier properties, but as Schulte im Walde et al. (2016a) they found overall better predictions for (i) compounds with low-frequency and low-ambiguity in comparison to high-frequency/-ambiguity heads; and for (ii) compounds with concrete and imaginable in comparison to abstract and low-imageability heads.

The described studies and their insights clearly demonstrate that – across variants of textual (and also multimodal) vector-space models – compound and con-

stituent properties strongly influence the prediction quality. We are thus asking two questions that we address in the current study. First of all, is there a way to understand better how humans perceive interactions between compound properties and compositionality ratings? We address this question by providing a novel collection (strategy) in Section §3. And secondly, how exactly do compound and constituent properties interact with compositionality ratings, in our novel collection and also in existing datasets? We will address this question by analysing the distributions and correlations of compositionality ratings and compound and constituent property distributions in Section §4.

3 Novel collection: Feature-based compositionality

In this section we present our novel compositionality ratings for German compounds. As target compounds, we rely on the union of targets from the above-described previous German datasets, Concrete-NN and Ghost-NN, resulting in a total of 1,099 German noun-noun compounds (i.e., 244 compound targets from Concrete-NN and 868 compound targets from Ghost-NN, minus 13 overlapping compound targets). Given that we aimed for a better understanding of what's on an annotator's mind when providing a judgement on a compound's degree of compositionality, we compiled a series of tasks for the annotators to fulfill in addition to providing the actual judgements. In the following we list these tasks, accompanied by the respective motivations. The full annotation guidelines are available in the appendix. The annotators were five graduate students of computational linguistics at the University of Stuttgart.

1. *Compound meaning*: We wanted the annotators to consciously pay attention to the overall meaning of the compound and therefore asked them to paraphrase the compound meaning within a phrase or a sentence. Similar tasks have previously been defined by, e.g., Wisniewski (1996) and Marsh (2015).

2. *Constituent meaning contribution*: Similarly, we wanted the annotators to consciously pay attention to the constituents' meaning components and their contributions to the meaning of the compound. We therefore asked them to explicitly provide one or more features of constituent meaning that contribute to the compound meaning, such as *failure* regarding the contribution of the head *Fehler* 'mistake' to the meaning of the compound *Kunstfehler*.

3. *Super-/sub-ordination (hypernymy/hyponymy)*: We wanted the annotators to be aware of potential hypernymy relationships between compounds and

head constituents, because we hypothesised that a large portion of the compound targets represent sub-ordinate categories (Gagné et al. 2019, 2020). We focused on the compound–head relationship and asked the annotators to judge if the compound is a hyponym *(is a kind)* of the compound head, on a scale $[0, 5]$.

4. *Abstractness/concreteness*: We wanted the annotators to be aware of the concreteness of the compounds and the constituents, because we hypothesised that the degree of concreteness might have an influence on the compositionality of the compound. We therefore asked them to judge about the concreteness (in contrast to abstractness) of compounds and constituents on a scale $[0, 5]$.

5. *Degree of compositionality*: Finally, we wanted the annotators to provide their judgements about the degrees of compositionality of the compounds with regard to their constituents on a scale $[0, 5]$ *after* fulfilling the above-listed tasks about compound and constituent properties.

All annotations are publicly available from http://www.ims.uni-stuttgart.de/data/feature-comp-nn, which also includes the spreadsheet for annotation that we gave to the annotators. In the following we provide insights into the various kinds of annotations we collected.

Regarding task 1 (compound meaning), Table 3 shows examples of paraphrases of compound meanings that were provided by the annotators. We can see that the paraphrases are strongly overlapping in some cases, e.g., for the compound *Autozug* 'car train' we find four almost identical phrases *Zug, der Autos transportiert* 'train that transports cars'. Yet, the paraphrases offer different aspects of meanings, such as *schwingen* 'to swing', *Instrument* 'instrument' and *Dekoration* 'decoration' for *Windspiel* (lit. 'wind game') 'wind chimes'. Overall, we judge the paraphrases as useful materials to approach the compound meanings, similarly to dictionary definitions and WordNet glosses.

Regarding task 2 (constituent meaning contribution), Table 4 shows examples of modifier and head features which the annotators considered as contributing to the compound meanings. When comparing these features with the compound paraphrases in Table 3, we can see that the overlap in the materials differs for constituents with more vs. less contributions to compound meaning, e.g., three annotators refer to *Panzer* 'carapace' as the meaning contribution of *Schild* 'shield' to *Schildkröte* (lit. 'shield toad') 'turtle', and *Instrument* 'instrument' for *Spiel* 'game' in *Windspiel* (lit. 'wind game') 'wind chimes'.

Table 3: Examples of compound paraphrases in FEATURE-NN.

Autozug 'car train'	(*ein*)*Zug, der Autos transportiert* (4 annotators) '(a) train that transports cars' *ein Zug für den Fernverkehr, der neben Personen auch Fahrzeuge befördert* 'a train for long-distance traffic that also carries vehicles, next to persons'
Eifersucht 'jealousy', (lit. 'eagerness addiction')	*Besitzanspruch auf eine Person* 'claim of ownership to a person' *eine Form des Neides im Kontext romantischer Beziehungen* 'a form of jealousy in the context of romantic relations' *Angst die Liebe oder Zuneigung eines Anderen mit jemanden teilen zu müssen* 'fear of having to share someone's love or affection' *anderer Ausdruck für Neid* 'different expression for jealousy' *Angst jemanden zu verlieren* 'fear to lose someone'
Schildkröte 'turtle', (lit. shield toad)	*Reptil mit Panzer* 'reptile with carapace' *eine Reptilienart mit einem charakteristischen Panzer auf dem Rücken* 'a type of reptile with characteristic carapace on back' *Reptilien mit Panzer* 'reptile with carapace' *ein Reptil mit einem harten Panzer um den Torso* 'a reptile with a hard carapace around the torso' *ein Reptil mit einem Panzer* 'a reptile with a carapace'
Windspiel 'wind chimes', (lit. 'wind game')	*Objekt, das im Wind schwingt* 'object that swings in the wind' *eine Art Instrument, das außerhalb von Gebäuden aufgehängt und vom Wind gespielt wird* 'a kind of instrument that hangs outside buildings and is played by the wind' *Dekoration die im Wind sich bewegt* 'decoration that moves in the wind' *Konstrukt, das sich im Wind bewegt und Geräusche macht* 'construct that moves in the wind and makes sounds' *eine hängende Dekoration, die im Wind Töne erzeugt* 'a hanging decoration that makes sounds in the wind'

Table 4: Examples of constituent features contributing to compound meaning in FEATURE-NN.

Bahnhof 'train station'	*Bahn* 'train'	*verkehrstechnisch, ziehend* 'transport connecting, pulling' *Bahnverkehr, Zugverkehr* 'rail/train traffic' *Transportmittel* 'means of transport' *Transportmittel, Zug* means of transport, train' *Zug* 'train'
Schildkröte 'turtle', (lit. 'shield toad')	*Schild* 'shield'	*schildförmig, schützend* 'shield-shaped', 'protective' *gepanzert, geschützt* 'armoured', 'protected' *mechanischer Schutz* 'mechanical protection' *Panzer, Schutz, robust* 'carapace', 'protection', 'robust' *gepanzert* 'shielded'
Windspiel 'wind chimes', (lit. 'wind game')	*Wind* 'wind'	*windig* 'windy' *Wind* 'wind' *Bewegung in der Luft* 'movement in the air' *Luft, Böe, wehen* 'air', 'gust', 'blow' *beweglich* 'movable'
Luftzug 'draught', (lit. 'air train')	*Zug* 'train'	*ziehend* 'pulling' *bewegt* 'moved' *Transportmittel* 'means of transport' *Bewegung* 'movement' *Richtung* 'direction'
Schildkröte 'turtle', (lit. 'shield toad')	*Kröte* 'toad'	*kriechend* 'creeping' *Reptil* 'reptile' *Amphibien die am Wasser leben* 'amphibians that live in the water' *Tier* 'animal', *Frosch* 'frog' *Reptil* 'reptile'
Windspiel 'wind chimes', (lit. 'wind game')	*Spiel* 'game'	*spielend* 'playing' *Instrument* 'instrument', *Klang* 'sound' *Vergnügen* 'pleasure' *unterhaltend* 'entertaining' *Musik* 'music'

Regarding task 3 (hypernymy relation between compounds and their head constituents), Table 5 shows examples of mean hypernymy ratings for a subset of the target compounds with heads *Spiel* 'game', *Werk* 'work'; 'factory' and *Zug* 'train'; 'draught'. The dataset FEATURE-NN contains a total of 39/76/28 compound types (i.e., 39/76/28 different modifiers) with heads *Spiel*, *Werk* and *Zug*, respectively. We can see that these heads strongly differ regarding their hypernymy relation strengths to the respective compounds. Figure 2 shows the distributions of the ratings across all compound heads (red box), and also for only the compounds with the three example heads (orange boxes). The boxplots show that (i) overall we have a target set of compounds that is highly skewed towards super-/subordination, but also that (ii) the hypernymy strength distribution varies according to specific compound heads.

Regarding task 4 (abstractness/concreteness), Figure 3 shows the distributions of the ratings across all compounds, all modifiers and all heads (green, blue and red boxes, respectively, as in Section §2.1), and Figure 4 shows the distribution across all compounds in comparison to the distributions across compounds with the same example heads as above, *Spiel*, *Werk* and *Zug*. In Figure 3 we can see that we have similar overall concreteness distributions for the compounds, the modifiers and the heads. When zooming into compounds with specific heads in Figure 4, we observe a more diverse picture: while the compounds, modifiers and heads of *Spiel* and *Zug* compounds are again skewed towards concreteness, the compounds and constituents of *Werk* compounds exhibit more diversity in their concreteness ratings.

Figure 5 and Table 6 look into the compositionality ratings in our novel dataset, making use of two perspectives. Figure 5 shows boxplots of compound–modifier and compound–head compositionality ratings. For both constituent types we can see skewed distributions towards strongly compositional compounds, similarly to the distributions in GHOST-NN, cf. Figure 1. Table 6 compares the novel ratings against the original ratings in the datasets CONCRETE-NN and GHOST-NN, relying on Spearman's rank-order correlation coefficient ρ. The correlations are between 0.663 and 0.792 and therefore all point towards strong agreement between the novel mean ratings and the original mean ratings. On the one hand, this allows us to judge our novel collection as reliable, even though a smaller number of annotators was involved; on the other hand, the strong correlations tell us that the additional rating tasks we asked the annotators to perform did not have a strong influence on their compositionality judgements.

Table 5: Examples of mean hypernymy ratings in FEATURE-NN for a subset of the target compounds with heads *Spiel* 'game', *Werk* 'work'; 'factory' and *Zug* 'train'; 'draught'.

Angriffsspiel 'offensive play'	3.2	*Mundwerk* 'gab'	0.2
Ballspiel 'ball game'	4.8	*Netzwerk* 'network'	0.4
Computerspiel 'computer game'	4.8	*Stahlwerk* 'steel plant'	5.0
Farbenspiel 'play of colours'	3.0	*Stockwerk* 'floor'	0.0
Gedankenspiel 'intellectual game'	1.6	*Tagewerk* 'day's work'	4.0
Glockenspiel 'chimes'	2.2	*Teufelswerk* 'devil's work'	2.8
Glücksspiel 'gambling'	4.4	*Triebwerk* 'power unit'	3.2
Kartenspiel 'card game'	5.0	*Uhrwerk* 'clockwork'	2.4
Kinderspiel 'children's game'; 'easy'	3.6	*Wunderwerk* 'miracle'	3.8
Kirchspiel 'parish'	0.8	*Zementwerk* 'cement plant'	4.8
Liebesspiel 'amorous play'	2.6	*Atemzug* 'breath'	0.8
Machtspiel 'power game'	3.2	*Autozug* 'car train'	5.0
Orgelspiel 'organ playing'	4.0	*Beutezug* 'foray'	3.2
Ritterspiel 'knights game'	4.0	*Charakterzug* 'character trait'	2.6
Schattenspiel 'shadow play'	4.2	*Dampfzug* 'steam train'	5.0
Trauerspiel 'fiasco'	2.9	*Fackelzug* 'torchlight procession'	2.4
Wasserspiel 'water game'	3.4	*Feldzug* 'campaign'	1.4
Windspiel 'wind chimes'	2.6	*Gebirgszug* 'mountain range'	1.6
Wortspiel 'pun'	3.2	*Gesichtszug* 'facial feature'	1.0
Würfelspiel 'game of dice'	5.0	*Kriegszug* 'military expedition'	2.6
Bergwerk 'mine'	4.6	*Luftzug* 'draught'	2.8
Blattwerk 'foliage'	1.6	*Nachtzug* 'night train'	5.0
Erstlingswerk 'first work'	3.4	*Protestzug* 'protest march'	3.6
Feuerwerk 'fireworks'	2.4	*Schachzug* 'chess move'; 'gambit'	2.5
Hexenwerk 'sorcery'; 'difficult'	2.8	*Schriftzug* 'lettering'	0.6
Klavierwerk 'piano work'	3.0	*Seilzug* 'cable pull'	3.6
Kraftwerk 'power station'	4.4	*Siegeszug* 'triumphal march'	1.4
Mauerwerk 'masonry'	2.0	*Trauerzug* 'funeral procession'	3.6
Meisterwerk 'masterpiece'	3.2	*Triumphzug* 'triumphal march'	3.4
Menschenwerk 'man-made'	4.0	*Vogelzug* 'bird migration'	1.8

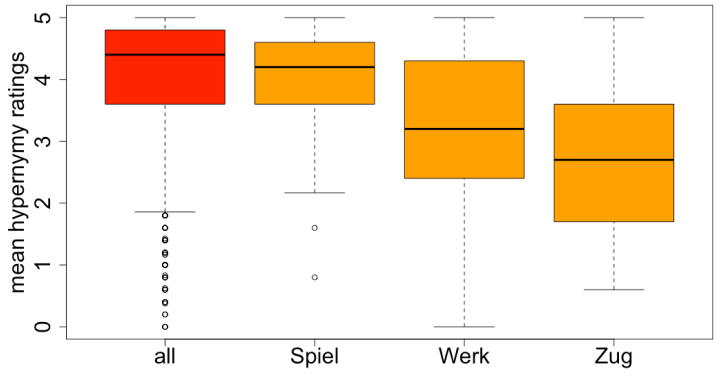

Figure 2: Strengths of hypernymy relation ratings in FEATURE-NN regarding all compound–head combinations in comparison to compounds with heads *Spiel, Werk* and *Zug*.

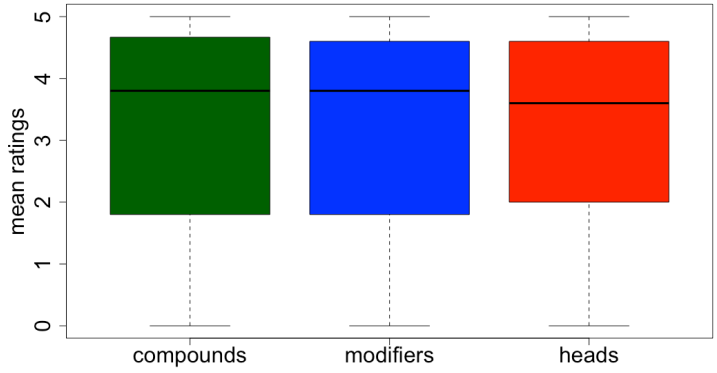

Figure 3: Concreteness ratings in FEATURE-NN.

Table 6: Correlations (ρ) between original and feature-based compositionality ratings for CONCRETE-NN and Ghost-NN compounds.

	constituent	ρ
CONCRETE-NN	modifier	0.792
	head	0.728
Ghost-NN/S	modifier	0.770
	head	0.687
Ghost-NN/XL	modifier	0.663
	head	0.687

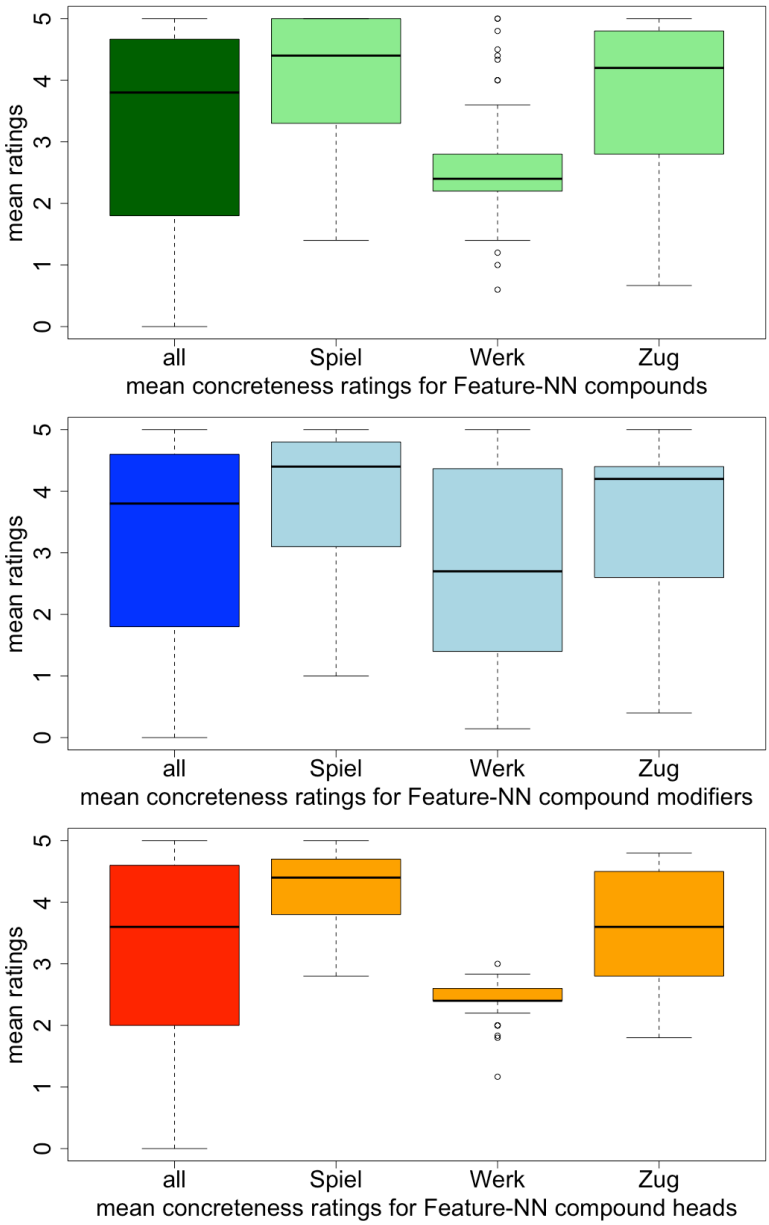

Figure 4: Concreteness ratings in FEATURE-NN, comparing ratings across all compounds (top), all compound–modifier combinations (middle), and all compound–head combinations (bottom) against those for compounds with heads *Spiel*, *Werk* and *Zug*, respectively.

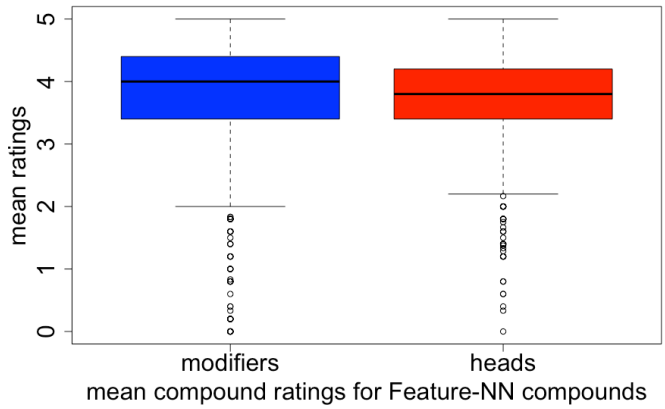

Figure 5: Compositionality ratings in Feature-NN.

4 Analyses

In this section, we raise and discuss two issues that we consider important for the creation of datasets with compositionality ratings, and potentially also for the creation of datasets with ratings on further semantic variables. (1) On the one hand, we are asking whether the distribution of ratings across a pre-specified scale of ratings should be even, as opposed to being skewed towards parts of the rating scale. (2) On the other hand, we are asking to what extent one should take into account properties of targets when creating a novel resource, and also when using a resource for evaluating computational models. In the following, we will look into rating distributions across datasets regarding issue (1), and into interactions between target properties and rating distributions regarding issue (2). As datasets, we will make use of the existing German and English resources Concrete-NN, Ghost-NN, Reddy-NN and Cordeiro-N[2] introduced in Section §2.1, as well as our novel resource Feature-NN introduced in the previous Section §3. As properties, we will make use of frequency, productivity and ambiguity values provided by Schulte im Walde et al. (2016a) and Miletic & Schulte im Walde (2023), hypernymy and concreteness ratings for the German targets collected in Feature-NN, and concreteness ratings for the English compound and constituent targets collected by Muraki et al. (2022) and Brysbaert et al. (2014), respectively.

Figure 1 on page 277 presented the distributions of compositionality ratings across the targets in the existing German and English rating datasets; Figure 5 on the previous page presented the distributions for our novel dataset Feature-NN.

[2]Plots for the Reddy-NN extension Cordeiro-N can be found in the appendix.

The two GHOST-NN variants and also our novel dataset FEATURE-NN are skewed towards strongly compositional targets, while the targets in CONCRETE-NN and even more so in REDDY-NN exhibit more even distributions. Figures 6 and 7 provide an additional view on the ratings in the latter two datasets, where the mean ratings on the x-axes are plotted in relation to the respective standard deviations (y-axes). The plots in Figures 6 and 7 confirm that there are more strongly compositional than strongly non-compositional or mid-scale targets in CONCRETE-NN, while REDDY-NN predominantly includes strongly compositional and also strongly non-compositional targets, in contrast to the mid-range which is covered rather sparsely. Overall, we induce from the distribution plots that (a) the concreteness-focused selection of targets for CONCRETE-NN, (b) the property-based balancing selection of targets for GHOST-NN, and (c) the target selection combining WordNet-based hypernymy and gloss overlap resulted in target sets with rather different distributions across compositionality ratings.

Table 7 looks into relations between compositionality ratings for compounds and compound–constituent combinations, by presenting correlations between the compositionality rating distributions for compounds and constituents within datasets. While we do not see meaningful correlations between the compound–modifier or the compound–head ratings in the GHOST-NN variants or FEATURE-NN, we find a weak negative correlation for CONCRETE-NN ($\rho = -0.372$) and weak positive correlations for REDDY-NN ($\rho = 0.265$) and CORDEIRO-N ($\rho = 0.353$). Even more so, we find strong correlations between compound and compound–modifier ratings (CONCRETE-NN: $\rho = 0.600$; REDDY-NN: $\rho = 0.804$; CORDEIRO-N: $\rho = 0.798$), and also between compound and compound–head ratings (REDDY-NN: $\rho = 0.720$ and CORDEIRO-N: $\rho = 0.759$). I.e., in CONCRETE-NN and REDDY-NN strongly compositional compounds include strongly meaning-contributing modifiers (and heads, in the datasets REDDY-NN and CORDEIRO-N), and strongly non-compositional compounds include strongly non-contributing modifiers (and heads). We will discuss these insights further after we have looked into compound properties across datasets, i.e., issue (2).

Tables 8 and 9 look into interactions between compositionality ratings and properties of compounds and constituents, again relying on Spearman's ρ correlations. More specifically, Table 8 shows correlations between compound ratings and compound frequency (freq), hypernymy (hyp), concreteness (conc), and also between compound–modifier ratings (modifier) and compound–head ratings (head) and the respective modifier/head properties, as well as productivity (prod) referring to the family size, and ambiguity (amb) referring to the number of senses. For the REDDY-NN and the CORDEIRO-N datasets, we do not have hypernymy ratings, but we assume that hypernymy is strongly involved in compound–

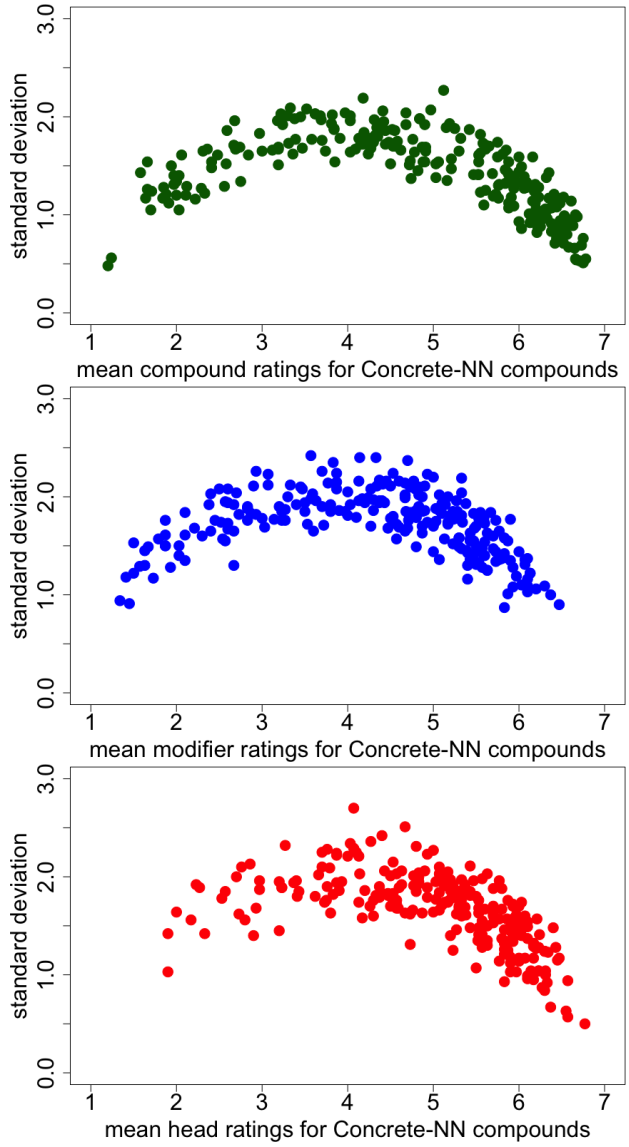

Figure 6: Mean compositionality ratings and standard deviations in CONCRETE-NN.

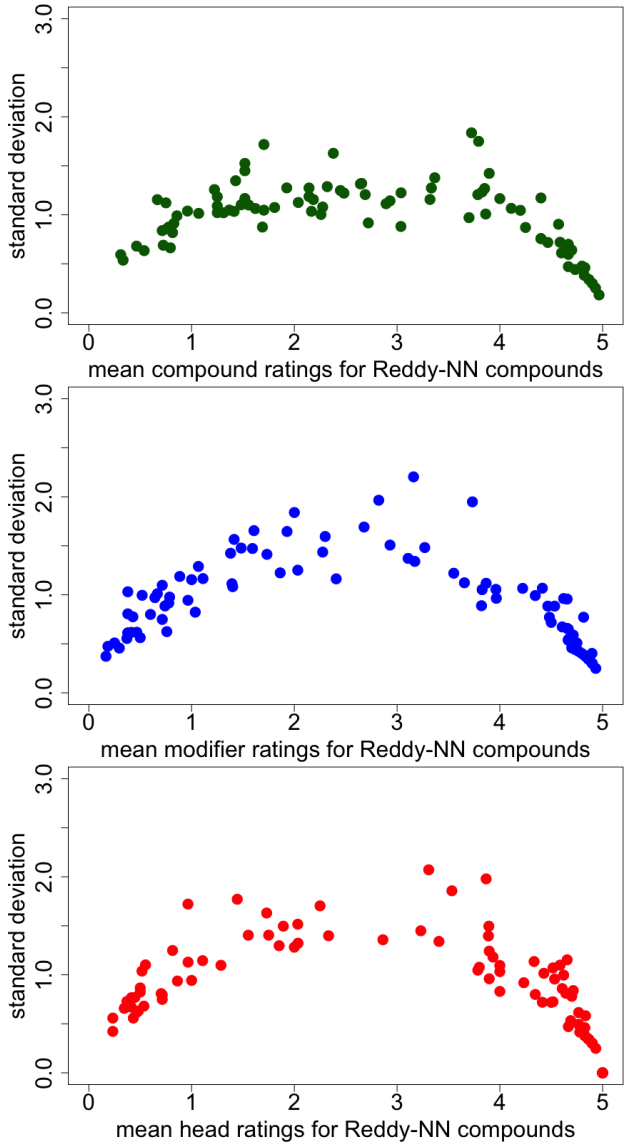

Figure 7: Mean compositionality ratings and standard deviations in Reddy-NN.

Table 7: Within-dataset correlations (ρ) between the compositionality ratings for compounds, modifiers and heads.

		ρ	
		modifier	head
German datasets			
CONCRETE-NN	compound	0.600	0.138
	modifier		−0.372
Ghost-NN/S	modifier		−0.087
Ghost-NN/XL	modifier		−0.123
Feature-NN	modifier		0.085
English datasets			
REDDY-NN	compound	0.804	0.720
	modifier		0.265
CORDEIRO-N	compound	0.798	0.759
	modifier		0.353

constituent relationships because of how targets were selected (cf. Section §2.1). We distinguish between original ratings (ORIG) and novel ratings (FEAT) in the German datasets, and we highlight cells with moderate-to-strong correlations $\rho > 0.4$.

The following observations are particularly striking: in the German dataset variants, we find a strong correlation between compound–head ratings and the degree of hypernymy ($0.624 \leq \rho \leq 0.797$), i.e., the stronger the degree of hypernymy, the more a head has been judged as contributing its meaning to the compound meaning, which we consider an indirect confirmation of the reliability of the ratings, because this is hypernymy per definitionem. In the FEATURE-NN ratings for the CONCRETE-NN compound–head combinations we further see a moderate correlation between the ratings and the heads' degrees of concreteness ($\rho = 0.414$). For compounds, the same type of correlation is even stronger in the REDDY-NN and the CORDEIRO-N datasets ($\rho = 0.592$ and $\rho = 0.469$, respectively), and negative for the concreteness of compound–modifier ratings in REDDY-NN ($\rho = -0.492$). Most striking in the table are the moderate correlations for REDDY-NN between all compound and compound–constituent ratings and their empirical properties frequency and productivity ($0.454 \leq \rho \leq 0.579$), while there are no moderate correlations between compositionality ratings and frequency and productivity in the German datasets.

Table 8: Correlations (ρ) between compound and constituent compositionality ratings and compound and constituent properties.

			Properties				
			freq	prod	amb	hyp	conc
Concrete-NN	orig	compound	−0.075	–	–	0.424	0.113
Concrete-NN	orig	modifier	0.080	0.164	−0.157	–	0.079
Concrete-NN	orig	head	−0.147	−0.178	−0.279	0.689	0.228
Concrete-NN	feat	modifier	0.020	0.114	−0.177	0.080	0.182
Concrete-NN	feat	head	−0.070	−0.061	−0.230	0.762	0.414
Ghost-NN/S	orig	modifier	0.032	0.024	−0.235	–	0.002
Ghost-NN/S	orig	head	−0.220	−0.271	−0.305	0.797	0.344
Ghost-NN/S	feat	modifier	0.020	0.071	−0.192	–	0.142
Ghost-NN/S	feat	head	−0.164	−0.197	−0.119	0.624	0.281
Ghost-NN/XL	orig	modifier	−0.088	−0.023	−0.231	–	0.119
Ghost-NN/XL	orig	head	−0.202	−0.204	−0.356	0.692	0.171
Ghost-NN/XL	feat	modifier	−0.130	−0.087	−0.164	–	0.212
Ghost-NN/XL	feat	head	−0.246	−0.250	−0.294	0.645	0.224
Reddy-NN		compound	0.579	–	–	–	0.592
Reddy-NN		modifier	0.547	0.471	0.172	–	−0.492
Reddy-NN		head	0.454	0.484	0.224	–	−0.207
Cordeiro-N		compound	0.385	–	–	–	0.469
Cordeiro-N		modifier	0.340	0.269	−0.100	–	−0.381
Cordeiro-N		head	0.307	0.331	0.110	–	−0.283

Table 9: Correlations (ρ) between compound compositionality ratings and compound and constituent properties.

	frequency			productivity		ambiguity	
	comp	mod	head	mod	head	mod	head
Concrete-NN	−0.075	0.049	0.099	0.101	0.199	−0.182	−0.060
Reddy-NN	0.579	0.535	0.393	0.517	0.464	0.219	0.133
Cordeiro-N	0.385	0.188	0.257	0.132	0.314	−0.140	0.072

In Table 9, we focus on compound ratings, this time looking into correlations between compound ratings and compound and constituent properties. We can see that the compound phrase/whole ratings in the Reddy-NN dataset are also moderately correlated with modifier and head frequencies and productivities.

We now turn towards a discussion of the analyses with regard to the two issues we raised: (1) to what extent one should aim for an even distribution of ratings across the pre-specified scale of ratings, and (2) to what extent one should take into account properties of targets when creating a novel resource and when using a resource for evaluation. We saw in our analyses that the datasets we explored are skewed towards certain ranges of compositionality in different ways, some contain more compositional than non-compositional compounds, and some contain many more ratings at either extreme of the compositionality scale than in the mid-range. Furthermore, in some datasets (but not in others) we find strong correlations between compound and compound–constituent ratings as well as moderate correlations between compositionality ratings and corpus-based frequencies and productivity scores. Which of these inter-dependencies are desired, and which are artefacts created by the specific strategies of how to select compound targets for the dataset? Optimally, one should aim for ratings on a scale that are evenly distributed across targets, both overall and also with regard to salient target properties, in order to ensure full coverage of the phenomenon. This goal is very difficult to achieve, however, because we can only check on rating distributions once we have collected the ratings. We therefore suggest to pay attention to a subset of target properties that are considered most salient and influential regarding the desired rating types. This was done for Ghost-NN by Schulte im Walde et al. (2016a), for example, whose resulting ratings are however highly skewed towards compositionality, so in retrospect our specific choice of salient properties may be considered suboptimal.

We see two alternative routes to follow, individually or in combination: (a) Balance your targets across frequency ranges as the minimally required target property, because we know that target frequency has generally a strong influence on language processing and comprehension (Ellis 2002). (b) If time and money allow, go for a large set of targets in the selection phase, such that the collected ratings may be analysed and the targets then be post-balanced across the most salient target properties in a post-processing filtering step. Realistically, many datasets that are available or will be available in the future still incorporate artefacts with regard to one or the other target property, so we need a workaround when evaluating our computational models on the basis of such datasets. Our baseline for this workaround is to assess models not only on the full dataset, but also with regard to subsets of targets with coherent task-relevant properties, similarly to our

studies described in Section §2.2 (Schulte im Walde et al. 2016a, Köper & Schulte im Walde 2017, Alipoor & Schulte im Walde 2020, Miletic & Schulte im Walde 2023). In this way we obtain a fine-tuned set of model results, rather than "just" an overall result score.

5 Conclusion

The current study started off with the observation that evaluations of computational models predicting degrees of compositionality for noun compounds typically evaluate their models across all targets, disregarding the fact that prediction models might vary according to properties of the targets within the gold standard resources. We suggested a novel route to assess the interactions between compound and constituent properties with regard to degrees of compositionality: (1) We created a novel collection Feature-NN with compositionality ratings for 1,099 German compounds, where we asked the human judges to provide compound and constituent properties (such as paraphrases, meaning contributions, hypernymy relations, and concreteness) before judging the compositionality; and (2) We performed a series of analyses on rating distributions and interactions with compound and constituent properties for our novel collection as well as previous gold standard resources for German (Concrete-NN and Ghost-NN) and English (Reddy-NN and Cordeiro-N). Our novel collection of ratings provides useful materials to investigate the meanings of the 1,099 compound targets and their constituents and is available from http://www.ims.uni-stuttgart.de/data/feature-comp-nn under a CC BY-NC-SA license. The obtained compositionality ratings are strongly correlated with previous ratings on the same targets, from which we induce (a) that we judge our novel ratings as reliable, and at the same time (b) that the additional ratings on compound and constituent properties that we asked the human judges to provide did not have a strong influence on their judgements.

Making use of our novel annotations as well as information on frequencies, productivities, ambiguities and degrees of concreteness regarding the target compounds and their constituents, we gained insight into distributions over compositionality ratings as well as interactions between these distributions and a range of target properties, most importantly: (a) The previous and also our novel collection of compositionality ratings all show skewed distributions, however in various ways: Ghost-NN and Feature-NN are skewed towards strongly compositional targets, while Reddy-NN includes strongly compositional and also strongly non-compositional targets while the mid-range is covered more sparsely.

(b) Regarding relations between compound and constituent ratings, Concrete-NN and Reddy-NN show moderate-to-strong correlations between compound and compound–modifier ratings (Concrete-NN: $\rho = 0.600$; Reddy-NN: $\rho = 0.804$) and between compound and compound–head ratings (Reddy-NN: $\rho = 0.720$). (c) Looking into the interactions between compound and constituent properties and their compositionality ratings, we found moderate-to-strong correlations with concreteness (Concrete-NN: $\rho = 0.414$; and Reddy-NN: $\rho = 0.592$ for compounds and $\rho = 0.492$ for heads), and we also found moderate correlations with frequency and productivity (Reddy-NN: $0.393 \geq \rho \geq 0.579$).

Following the analyses we discussed to what extent one should aim for an even distribution of ratings across the pre-specified scale, and to what extent one should take into account properties of targets when creating a novel resource and when using a resource for evaluation. We suggest as a minimum requirement to balance targets across frequency ranges, and optimally to balance targets across their most salient properties in a post-collection filtering step. Above all, we recommend assessing computational models not only on the full dataset but also with regard to subsets of targets with coherent task-relevant properties. We believe that especially the latter recommendation does not only apply to compositionality ratings (resources and models) but more generally to creating and using evaluation datasets across tasks.

Abbreviations

MWE	multiword expression
NLP	natural language processing
NLU	natural language understanding
BE	semantic compound relation: be
HAVE	semantic compound relation: have
IN	semantic compound relation: in
ABOUT	semantic compound relation: about
ACTOR	semantic compound relation: actor
INST	semantic compound relation: instrument
LEX	no semantic compound relation; lexicalised compound

Acknowledgements

We thank the five annotators for their contributions to the creation of the dataset Feature-NN, and we thank Chris Jenkins and Filip Miletic as well as the two

anonymous reviewers and the editors of this volume for their feedback on previous versions of this chapter. Our research received funding from the German Research Foundation (DFG) through projects *Sense Discrimination and Regular Meaning Shifts of German Particle Verbs* in the Collaborative Research Centre SFB 732, SCHU 2580/5 *Computational Models of the Emergence and Diachronic Change of Multi-Word Expression Meanings*, and SCHU 2580/2 *Distributional Approaches to Semantic Relatedness*.

Appendix A Annotation guidelines for FEATURE-NN ratings

A.1 Original German version: Guidelines für die Annotation von Eigenschaften komplexer Nomen und ihrer Konstituenten

In der Datei `anno-comp-ratings-feat.ods` findest Du eine Liste von komplexen Nomen und ihren zwei nominalen Konstituenten in den Spalten A, B und C (und für eine bessere Übersichtlichkeit wiederholt in den Spalten M, N und O). In den dazwischen liegenden Spalten bitten wir Dich um Deine spontanen Intuitionen bezüglich folgender Eigenschaften:

Spalte D: **Bedeutung des komplexen Nomens**

> Aufgabe: Erkläre die Bedeutung des komplexen Nomens in einer Phrase/einem Satz. Du darfst (musst aber nicht) die Konstituenten des Nomens in Deiner Erklärung verwenden.
>
> Beispiel: Die Bedeutung des komplexen Nomens *Eselsohr* ist *verknickte Ecke einer Buchseite*.

Spalten E und F: **Eigenschaften der Konstituenten**

> Welche Eigenschaften der ersten bzw. zweiten Konstituente finden sich in dem komplexen Nomen wieder? Falls Dir mehrere Eigenschaften einfallen, trenne diese bitte durch Komma. Falls Dir keine Eigenschaft einfällt, trage bitte "0" ein.
>
> Beispiel: Bei dem komplexen Nomen *Kunstfehler* trägt z.B. die erste Konstituente die Eigenschaften *sehr gut, Qualität* bei, die zweite Konstituente z.B. die Eigenschaft *Misserfolg*.
>
> Versuche, jede Eigenschaft auf ein oder wenige Worte zu beschränken. Die Wortarten sind beliebig.

Spalte G: **Über-/Unterordnung**

Ist das komplexe Nomen "eine Art" der zweiten Konstituente? Nutze eine Skala von 0 (nein, gar nicht) bis 5 (ja, absolut).

Beispiel: "Ein Ahornbaum **ist** eine Art von Baum", aber
"Ein Eselsohr **ist keine** Art von Ohr".

Spalten H–J: **Abstraktheit/Konkretheit**

Wie abstrakt bzw. konkret sind das komplexe Nomen sowie die erste bzw. zweite Konstituente? Nutze wiederum eine Skala von 0 (ganz abstrakt) bis 5 (ganz konkret).

Hinweis: Konkrete Wörter können durch die menschlichen Sinne (hören, riechen, schmecken, sehen, tasten) erfasst werden (z.B. *Tisch, Lärm*), abstrakte Wörter nicht (z.B. *Idee, Traum*).

Spalten K–L: **Kompositionalität**

Wie sehr lässt sich die Gesamtbedeutung des komplexen Nomens aus der Bedeutung der ersten bzw. zweiten Konstituente ableiten? Nutze wiederum eine Skala von 0 (gar nicht) bis 5 (sehr stark).

A.2 Tentative English translation: Guidelines for annotating properties of complex nouns and their constituents

The file `anno-comp-ratings-feat.ods` provides a list of complex nouns and their two nominal constituents in columns A, B and C (and repeated in columns M, N and O). In the intermediate columns we ask for your spontaneous intuitions regarding the following properties:

Column D: **Meaning of the complex noun**

Task: Explain the meaning of the complex noun within one phrase/sentence. You may (but you do not have to) use the constituents of the noun in your explanation.

Example: The meaning of the complex noun *Eselsohr* (lit. 'donkey ear') 'earmark' is a *folded corner of a page in a book.*

Columns E and F: **Properties of the constituents**

Which properties of the first/second constituent do you recognise in the complex noun? If you are aware of several properties, please separate them with commas. If you are not aware of any property, please enter "0".

Example: Regarding the complex noun *Kunstfehler* (lit. 'art mistake') 'malpractice' the first constituent contributes the properties *excellent* and *quality*, and the second constituent contributes the property *failure*.

Try to use only one or a few words for each property. You may use words of any word class.

Column G: **Super-/subordination**

Is the complex noun "a kind of" the second constituent? Please use a scale between 0 (no, not at all) and 5 (yes, absolutely).

Example: "An *Ahornblatt* 'maple tree' **is** a kind of tree", but
"An Eselsohr (lit. 'donkey ear') 'earmark' **is not** a kind of ear".

Columns H–J: **Abstractness/concreteness**

How abstract/concrete are the complex noun and the first and second constituent? Again, please use a scale between 0 (totally abstract) and 5 (totally concrete).

Hint: Concrete words can be perceived by human senses: hearing, smelling, tasting, seeing, touching (e.g., *table, noise*), abstract words cannot (e.g., *idea, dream*).

Columns K–L: **Compositionality**

To what degree can you induce the meaning of the complex nouns from the meanings of the first/second constituents? Again, please use a scale between 0 (not at all) and 5 (totally).

Appendix B Cordeiro dataset ratings

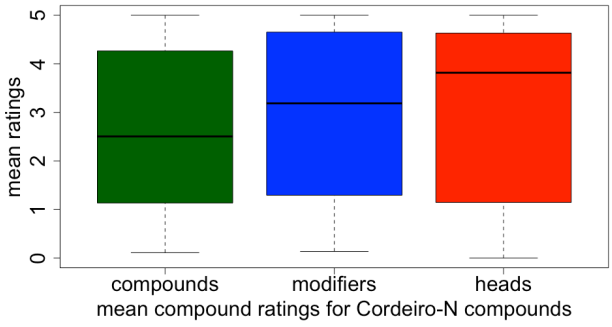

Figure 8: Compositionality rating distributions in Cordeiro-N.

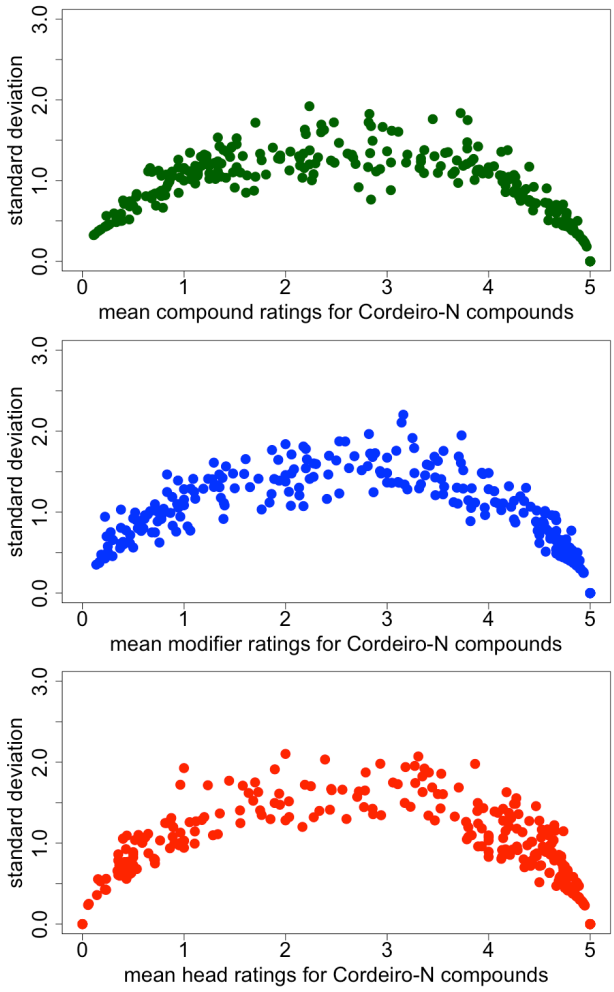

Figure 9: Mean compositionality ratings and standard deviations for compounds in CORDEIRO-N.

Appendix C Concreteness of Targets in Reddy-NN and Cordeiro-N

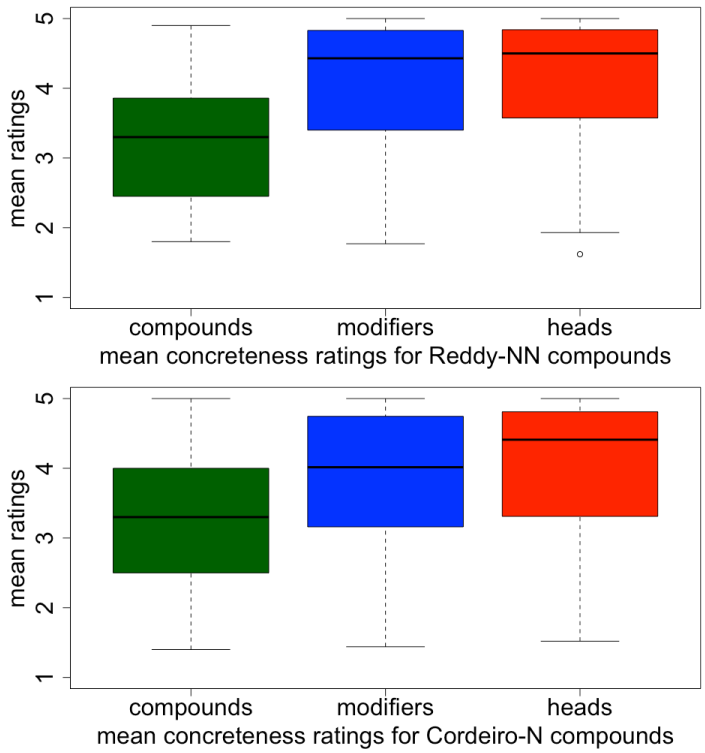

Figure 10: Concreteness ratings in Reddy-NN and Cordeiro-N.

References

Alipoor, Pegah & Sabine Schulte im Walde. 2020. Variants of vector space reductions for predicting the compositionality of English noun compounds. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'2020)*, 4379–4387. Marseille, France: ACL. https://aclanthology.org/2020.lrec-1.539.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 267–292. Boca Raton, FL: CRC Press.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3). 209–226.

Bauer, Laurie. 2017. *Compounds and compounding*. Cambridge University Press.

Bell, Melanie J. & Martin Schäfer. 2013. Semantic transparency: Challenges for distributional semantics. In Aurelie Herbelot, Roberto Zamparelli & Gemma Boleda (eds.), *Proceedings of the IWCS 2013 workshop on formal distributional semantics*, 1–10. Potsdam, Germany.

Benczes, Réka. 2014. What can we learn about the mental lexicon from non-prototypical cases of compounding? *Argumentum* 10. 205–220.

Bettinger, Julia, Anna Hätty, Michael Dorna & Sabine Schulte im Walde. 2020. A domain-specific dataset of difficulty ratings for German noun compounds in the domains DIY, cooking and automotive. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC'2020)*, 4359–4367. Marseille, France: European Language Resources Association (ELRA).

Bott, Stefan & Sabine Schulte im Walde. 2017. Factoring ambiguity out of the prediction of compositionality for German multi-word expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Proceedings of the 13th workshop on multiword expressions (MWE 2017)*, 66–72. Valencia, Spain. DOI: 10.18653/v1/W17-1708.

Brysbaert, Marc, Amy Beth Warriner & Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 64. 904–911. DOI: 10.3758/s13428-013-0403-5.

Butterworth, Brian. 1983. Lexical representation. In *Language production*, vol. 2: Development, writing and other language processes, 257–294. London: Academic Press.

Cap, Fabienne, Manju Nirmal, Marion Weller & Sabine Schulte im Walde. 2015. How to account for idiomatic German support verb constructions in statistical machine translation. In *Proceedings of the 11th workshop on multiword expressions*, 19–28. Denver, CO.

Carpuat, Marine & Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Proceedings of the 11th annual conference of the North American chapter of the Association for Computational Linguistics*. Los Angeles, CA.

Cholakov, Kostadin & Valia Kordoni. 2014. Better statistical machine translation through linguistic treatment of phrasal verbs. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, 196–201. Doha, Qatar.

Clouet, Elizaveta Loginova & Béatrice Daille. 2014. Splitting of compound terms in non-prototypical compounding languages. In Ben Verhoeven, Walter Daelemans, Menno van Zaanen & Gerhard van Huyssteen (eds.), *Proceedings of the 1st workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, 11–19. Dublin, Ireland: ACL. DOI: 10.3115/v1/W14-5702.

Cordeiro, Silvio, Aline Villavicencio, Marco Idiart & Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics* 45(1). 1–57.

Costello, Fintan J. & Mark T. Keane. 2000. Efficient creativity: Constraint-guided conceptual combination. *Cognitive Science* 24(2). 299–349.

Dankers, Verna, Elia Bruni & Dieuwke Hupkes. 2022. The paradox of the compositionality of natural language: A neural machine translation case study. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th annual meeting of the Association for Computational Linguistics*, 4154–4175. Dublin, Ireland. DOI: 10.18653/v1/2022.acl-long.286.

de Jong, Nicole H., Laurie B. Feldman, Robert Schreuder, Michael Pastizzo & R. Harald Baayen. 2002. The processing and representation of Dutch and English compounds: Peripheral morphological and central orthographic effects. *Brain and Language* 81. 555–567.

Dima, Corina, Daniel de Kok, Neele Witte & Erhard Hinrichs. 2019. No word is an island: A transformation weighting model for semantic composition. *Transactions of Computational Linguistics* 7. 437–451.

Eichel, Annerose, Helena Schlipf & Sabine Schulte im Walde. 2023. *Made of Steel?* Learning plausible materials for components in the vehicle repair domain. In *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics*, 1420–1435. Dubrovnik, Croatia.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2). 143–188.

Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database* (Language, Speech, and Communication). Cambridge, MA: MIT Press.

Firth, John R. 1957. *Papers in linguistics 1934–1951*. London, UK: Longmans.

Gagné, Christina L. 2002. Lexical and relational influences on the processing of novel compounds. *Brain and Language* 81. 723–735.

Gagné, Christina L., Thomas L. Spalding & Daniel Schmidtke. 2019. LADEC: The large database of English compounds. *Behavior Research Methods* 51. 2152–2179.

Gagné, Christina L., Thomas L. Spalding, Patricia Spicer, Dixie Wong & Beatriz Rubio. 2020. Is *buttercup* a kind of cup? Hyponymy and semantic transparency in compound words. *Journal of Memory and Language* 113. DOI: 10.1016/j.jml.2020.104110.

Gamallo, Pablo, Susana Sotelo, Jose Ramom Pichel & Mikel Artetxe. 2019. Contextualized translations of phrasal verbs with distributional compositional semantics and monolingual corpora. *Computational Linguistics* 45(3). 395–421.

Girju, Roxana, Dan Moldovan, Marta Tatu & Daniel Antohe. 2005. On the semantics of noun compounds. *Journal of Computer Speech and Language* 19(4). 479–496.

Hamp, Birgit & Helmut Feldweg. 1997. GermaNet: A lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*. https://aclanthology.org/W97-0802.

Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146–162.

Hätty, Anna, Julia Bettinger, Michael Dorna, Jonas Kuhn & Sabine Schulte im Walde. 2021. Compound or term features? Analyzing salience in predicting the difficulty of German noun compounds across domains. In *Proceedings of the 10th joint conference on lexical and computational semantics*, 252–262. Bangkok, Thailand.

Hätty, Anna, Ulrich Heid, Anna Moskvina, Julia Bettinger, Michael Dorna & Sabine Schulte im Walde. 2019. AkkuBohrHammer vs. AkkuBohrhammer: Experiments towards the evaluation of compound splitting tools for general language and specific domains. In *Proceedings of the 15th conference on natural language processing (KONVENS 2019)*, 59–67. Erlangen, Germany: German Society for Computational Linguistics & Language Technology.

Hätty, Anna & Sabine Schulte im Walde. 2018. Fine-grained termhood prediction for German compound terms using neural networks. In *Proceedings of the COLING joint workshop on linguistic annotation, multiword expressions and constructions*, 62–73. Santa Fe, NM.

Hermann, Karl Moritz. 2014. *Distributed representations for compositional semantics*. University of Oxford. (Doctoral dissertation).

Köper, Maximilian & Sabine Schulte im Walde. 2017. Complex verbs are different: Exploring the visual modality in multi-modal models to predict compositionality. In *Proceedings of the 13th workshop on multiword expressions*, 200–206. Valencia, Spain.

Kunze, Claudia. 2000. Extension and use of GermaNet, a lexical-semantic database. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.), *Proceedings of the 2nd international Conference on Language Resources and Evaluation (LREC'00)*, 999–1002. Athens, Greece: European Language Resources Association (ELRA). https://aclanthology.org/L00-1274/.

Levi, Judith N. 1978. *The syntax and semantics of complex nominals*. London: Academic Press.

Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 1997. Semantic transparency and compound fracture. *CLASNET Working Papers* (9). 1–13.

Libben, Gary, Martha Gibson, Yeo Bom Yoon & Dominiek Sandra. 2003. Compound fracture: The role of semantic transparency and morphological headedness. *Brain and Language* 84. 50–64.

Marsh, Charles. 2015. *Cigarette helmets and horse wars: Towards a better understanding of noun compound interpretability*. Department of Computer Science, Princeton University. (Bachelor thesis).

Miletic, Filip & Sabine Schulte im Walde. 2023. A systematic search for compound semantics in pretrained BERT architectures. In *Proceedings of the 17th conference of the European chapter of the Association for Computational Linguistics*, 1499–1512. Dubrovnik, Croatia.

Mitchell, Jeff & Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science* 34. 1388–1429.

Muraki, Emiko J., Summer Abdalla, Marc Brysbaert & Penny M. Pexman. 2022. Concreteness ratings for 62 thousand English multiword expressions. DOI: 10.31234/osf.io/m397u.

Murphy, Gregory L. 1990. Noun phrase interpretation and conceptual combination. *Journal of Memory and Language* 29. 259–288.

Nastase, Viviana A. 2003. *Semantic relations across syntactic levels*. School of Information Technology & Engineering, University of Ottawa. (Doctoral dissertation).

Ó Séaghdha, Diarmuid. 2007. Designing and evaluating a semantic annotation scheme for compound nouns. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics (CL2007)*. Birmingham, UK: University of Birmingham, UK. https://ucrel.lancs.ac.uk/publications/cl2007/.

Plag, Ingo. 2003. *Word-formation in English*. Cambridge University Press.

Reddy, Siva, Ioannis P. Klapaftis, Diana McCarthy & Suresh Manandhar. 2011a. Dynamic and static prototype vectors for semantic composition. In *Proceedings of the 5th international joint conference on Natural Language Processing*, 705–713. Chiang Mai, Thailand.

Reddy, Siva, Diana McCarthy & Suresh Manandhar. 2011b. An empirical study on compositionality in compound nouns. In *Proceedings of the 5th international joint conference on natural language processing*, 210–218. Chiang Mai, Thailand.

Roller, Stephen & Sabine Schulte im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the conference on empirical methods in natural language processing*, 1146–1157. Seattle, WA, USA.

Roller, Stephen & Sabine Schulte im Walde. 2014. Feature norms of German noun compounds. In *Proceedings of the 10th workshop on multiword expressions*, 104–108. Gothenburg, Sweden.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Alexander F. Gelbukh (ed.), *Proceedings of the third international conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, 1–15. Springer.

Salehi, Bahar & Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In *Proceedings of the 2nd joint conference on lexical and computational semantics*, 266–275. Atlanta, GA.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014a. Detecting non-compositional MWE components using Wiktionary. In *Proceedings of the conference on empirical methods in Natural Language Processing*, 1792–1797. Doha, Qatar.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2014b. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In Shuly Wintner, Sharon Goldwater & Stefan Riezler (eds.), *Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics*, 472–481. Gothenburg, Sweden: ACL. DOI: 10.3115/v1/E14-1050.

Salehi, Bahar, Paul Cook & Timothy Baldwin. 2015a. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics/human language technologies*, 977–983. Denver, Colorado, USA.

Salehi, Bahar, Nitika Mathur, Paul Cook & Timothy Baldwin. 2015b. The impact of multiword expression compositionality on machine translation evaluation. In *Proceedings of the 11th workshop on multiword expressions*, 54–59. Denver, CO.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejcek, Fabienne Cap, Slavomir Ceplo, Silvio Ricardo Cordeiro, Gulsen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartin, Lonneke van der

Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87–147. Berlin: Language Science Press. DOI: 10.5281/zenodo.14715.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.), *Proceedings of the 3rd workshop on Challenges in the Management of Large Corpora (CMLC-3)*, 28–34. Mannheim, Germany.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul. http://www.lrec-conf.org/proceedings/lrec2012/pdf/834_Paper.pdf.

Schulte im Walde, Sabine & Susanne Borgwaldt. 2015. Association norms for German noun compounds and their constituents. *Behavior Research Methods* 47(4). 1199–1221.

Schulte im Walde, Sabine, Anna Hätty & Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of English and German noun-noun compounds: A vector-space perspective. In *Proceedings of the 5th joint conference on lexical and computational semantics*, 148–158. Berlin, Germany.

Schulte im Walde, Sabine, Anna Hätty, Stefan Bott & Nana Khvtisavrishvili. 2016b. $G_h$ost-NN: A representative gold standard of German noun-noun compounds. In *Proceedings of the 10th international conference on Language Resources and Evaluation*, 2285–2292. Portoroz, Slovenia.

Schulte im Walde, Sabine, Stefan Müller & Stephen Roller. 2013. Exploring vector space models to predict the compositionality of German noun-noun Compounds. In Mona Diab, Tim Baldwin & Marco Baroni (eds.), *Proceedings of the 2nd joint conference on Lexical and Computational Semantics*, 255–265. Atlanta, GA. https://aclanthology.org/S13-1038.

Schulte im Walde, Sabine & Eva Smolka (eds.). 2020. *The role of constituents in multi-word expressions: An interdisciplinary, cross-lingual perspective* (Phraseology and Multiword Expressions 4). Berlin: Language Science Press. DOI: 10.5281/zenodo.3598577.

Siegel, Sidney & N. John Castellan. 1988. *Nonparametric statistics for the behavioral sciences*. Boston, MA: McGraw-Hill.

Spalding, Thomas L., Christina L. Gagné, A. C. Mullaly & Hongbo Ji. 2010. Relation-based interpretation of noun-noun phrases: A new theoretical approach. In Susan Olsen (ed.), *New impulses in word-formation* (Linguistische Berichte Sonderhefte 17), 283–316.

Taft, Marcus & Kenneth I. Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior* 14. 638–648.

von der Heide, Claudia & Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen: Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, 51–74.

Weller, Marion, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde & Alexander Fraser. 2014. Distinguishing degrees of compositionality in compound splitting for statistical machine translation. In *Proceedings of the 1st workshop on computational approaches to compound analysis*, 81–90. Dublin, Ireland.

Wisniewski, Edward J. 1996. Construal and similarity in conceptual combination. *Journal of Memory and Language* 35. 434–453.