Automatic Classification of Attributes in German Adjective-Noun Phrases

Neele Falk[†,‡*], Yana Strakatova[†*], Eva Huber[†,¶], Erhard Hinrichs[†]
[†]SfS, University of Tübingen / Tübingen, Germany
[‡]IMS, University of Stuttgart / Stuttgart, Germany
[†]firstname.lastname@uni-tuebingen.de
[‡]neele.falk@ims.uni-stuttgart.de
[¶]eva.huber@uzh.ch

Abstract

Adjectives such as *heavy* (as in *heavy rain*) and *windy* (as in *windy day*) provide possible values for the attributes `intensity` and `climate`, respectively. The attributes themselves are not overtly realized and are in this sense implicit. While these attributes can be easily inferred by humans, their automatic classification poses a challenging task for computational models. We present the following contributions: (1) We gain new insights into the attribute selection task for German. More specifically, we develop computational models for this task that are able to generalize to unseen data. Moreover, we show that classification accuracy depends, inter alia, on the degree of polysemy of the lexemes involved, on the generalization potential of the training data and on the degree of semantic transparency of the adjective-noun pairs in question. (2) We provide the first resource for computational and linguistic experiments with German adjective-noun pairs that can be used for attribute selection and related tasks. In order to safeguard against unwelcome memorization effects, we present an automatic data augmentation method based on a lexical resource that can increase the size of the training data to a large extent.

1 Introduction

There is ample evidence that humans decompose the meaning of objects and events into a set of prototypical semantic relations and their values. These relations, referred to in different frameworks as *attributes* (Barsalou, 1992), *frame elements* (Fillmore, 1982), *thematic relations* (Gruber, 1965), or *thematic roles* (Jackendoff, 1972), serve as an effective means to cluster classes of objects and events by degrees of semantic similarity. For example, thematic roles such as `buyer` and `seller` help distinguish among different participants in a financial transaction, and adjectives, such as *young* and *old*, group individuals into different equivalence classes for the relation `age`. Likewise, adjectives such as *heavy* (as in *heavy rain*) and *windy* (as in *windy day*) provide possible values for the attributes `intensity` and `climate`, respectively. The attributes themselves are not overtly realized and are in this sense implicit. While these attributes can be easily inferred by humans, their automatic classification poses a challenging task for computational models, as shown in the recent study by Shwartz and Dagan (2019) for English data. Compared to automatic role assignment for verbal arguments, attribute selection for adjective-noun pairs has received relatively little attention in computational semantics.

Attribute selection is highly relevant in different NLP tasks, such as information retrieval, topic modelling, and sentiment analysis. Consider a sentiment analysis task. If there is positive/negative sentiment expressed about something or someone, it is useful to know what triggers that sentiment. This requires from a system the ability to generalize over specific adjectives to more abstract attributes:

(1) *I {like/don't like} her siblings. They are*

a. *{bright/stupid} people.*
Attribute: `intelligence`

b. *{friendly/rude} people.*
Attribute: `behaviour`

For polysemous adjectives, the attribute selection task can be viewed as a coarse-grained word sense disambiguation. For instance, the adjective *bright* in example (1a) may acquire different meanings when it combines with different nouns, e.g. *bright room*, where the attribute is not `intelligence`, but `perception`.

In this paper, we frame the attribute selection task as a multiclass classification problem. We conduct experiments on the German dataset GerCo (Strakatova et al., 2020) of adjective-noun phrases. To the best of our knowledge, this is the first at-

[*] denotes equal contribution

Proceedings of the 14th International Conference on Computational Semantics, pages 239–249
June 17–18, 2021. ©2021 Association for Computational Linguistics

tribute analysis for German. Our main contributions are the following: (1) We gain new insights into the attribute selection task for German. More specifically, we develop computational models for this task that are able to generalize to unseen data. Moreover, we show that classification accuracy depends, inter alia, on the degree of polysemy of the lexemes involved, on the generalization potential of the training data and on the degree of semantic transparency of the adjective-noun pairs in question. (2) We provide the first resource for computational and linguistic experiments with German adjective-noun pairs that can be used for attribute selection and related tasks. In order to safeguard against unwelcome memorization effects, we present an automatic data augmentation method based on a lexical resource that can increase the size of the training data to a large extent.

This paper is structured as follows. We discuss related work in section 2. Section 3 describes the dataset in more detail. In section 4, we present the experiments and their results. Finally, we draw conclusions and give directions for future work in section 5.

2 Related work

Earlier studies of attribute selection focus primarily on English data. Hartung (2015) and Hartung et al. (2017) investigate the attributes in AN phrases and create a dataset for English adjective-noun phrases and their corresponding attributes based on the English WordNet. Hartung et al. (2017) try to model the task of selecting underlying attributes such as `age` for a phrase such as *old car* with representation learning: they experiment with different composition models to construct a single vector for the adjective-noun combination from the embeddings of the adjective and the noun. This composed vector is then used as a proxy for the underlying attribute, e.g. `age` and ranked with possible alternative values for other candidate attributes. Shwartz and Dagan (2019) evaluate different types of word embeddings on a number of lexical semantics tasks, including attribute selection and probe their ability to model lexical composition. For that purpose they reformulate the task of attribute selection into a binary classification: given an adjective-noun pair and an attribute, the classifiers predict whether the target attribute is selected for the pair in question. Their findings on the English dataset reveal that this task remains a challenge for all embedding types, though contextualized embeddings clearly outperform static embeddings.

Our work differs this from previous work in several aspects: we create the first dataset for the annotation of attributes in adjective-noun pairs for German. The taxonomy of 16 attributes is not as fine-grained as in Hartung (2015), who distinguishes between 254 attribute labels. Our more compact label set is thus more coarse-grained and more suitable for automatic modeling. We test the automatic models in a multiclass-classification setup with the adjective and noun embedding as input.

Unlike previous work on attribute selection, we take into account whether the semantics of an adjective-noun pair is transparent or not. Since the GerCo dataset contains both collocations and free phrases, we can partition the data accordingly and can compare the results obtained by a given classifier for the two classes. In earlier work (Strakatova et al., 2020), we report on binary classifiers for collocational and free adjective-noun pairs, which did not include prediction of the target attributes. In the present paper, the relevant attributes are taken into account. Therefore, our research contributes to a growing number of studies of semantic transparency, which up to now have focused on multiword expressions and nominal compounds (Reddy et al., 2011; Bell and Schäfer, 2013; Jana et al., 2019; Shwartz and Dagan, 2019) in particular, and extends this body of literature to the empirical domain of adjective-noun pairs. Our ability to distinguish between free phrases and collocations, allows us to test the finding of Espinosa Anke et al. (2019), who show that semantic relations in collocations are more difficult to predict in comparison to other types of relations such as hyponymy, meronymy, etc.

In sum, previous studies confirm that (i) revealing lexical relations in compounds and AN phrases is a challenge in NLP and (ii) relations found in collocations are more difficult to predict than other types of lexical relations. We combine these two findings in our study and model the lexical-semantic relations, which we call *attributes*, for both collocations and free phrases.

3 Data

In our experiments, we use the German dataset of adjective-noun phrases GerCo (Strakatova et al., 2020) which we annotate with additional semantic information.[1] This dataset is suitable for our study due to several reasons: (1) it contains highly polysemous adjectives; (2) half of the dataset is represented by collocations; (3) it is based on a lexical resource – the German wordnet GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) which can assist us in augmenting the data and obtaining attribute information about it.

The original GerCo dataset contains 3,652 AN phrases manually annotated as "collocations" and "free phrases". The distinction between the two types is based on the transparency of the adjective in the phrase that is operationalized as literality (Reddy et al., 2011). For instance, in the phrase *grober Sand* 'coarse sand', the adjective has its literal sense of "rough in texture" – it is annotated as *free phrase*. In the phrase *grober Fehler* 'gross mistake', the meaning of the adjective is shifted: it does not describe texture in combination with the noun *Fehler* 'mistake', but refers to its intensity.

The adjectives in GerCo have been chosen on the basis of the semantic classes that they are assigned to in GermaNet. The advantage of GermaNet as a lexical resource is that, in contrast to the English WordNet, it models adjectives in a hierarchical manner similarly to nouns and verbs. From each of the 16 semantic classes for German adjectives, three adjectives have been selected. Each adjective is paired with the most frequent co-occurring nouns, thus all adjective-noun pairs in the dataset have a strong association.[2] In the present study, we excluded two relational adjectives from the data: *barock* 'baroque' and *steinig* 'stony'. Out of the remaining 46 adjectives, 44 have at least two senses (Strakatova et al., 2020). The top nodes of the GermaNet hierarchy of adjectives represent the 16 semantic classes and the direct hyponyms of the top nodes represent more fine-grained classes of adjectives.[3] Figure 1 shows a part of the taxonomy for one sense of adjectives *tief* 'deep' and *salzig* 'salty'. The top nodes are used as attribute labels to annotate the data (see section 3.1).

We make use of this hierarchical structure for adjectives in GermaNet in two ways: extracting attribute information (subsection 3.1) and automatic augmentation of the dataset (subsection 3.2).

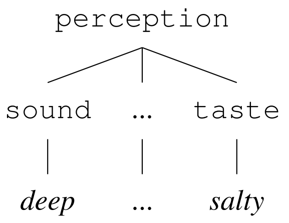

Figure 1: A part of the taxonomy of adjectives in GermaNet for *tief* 'deep' and *salzig* 'salty'. The top node is used as attribute label to annotate the GerCo dataset

3.1 Gold standard

For the present study, we add two layers of semantic annotation to the GerCo dataset: (1) by manual annotation: word sense IDs in GermaNet for all the adjectives and nouns in the dataset; (2) by automatic annotation: attributes for all the phrases.

Manual annotation. Manual annotation has been performed by two advanced students of computational linguistics with a solid background in lexical semantics and lexicography. Each adjective and noun from the GerCo dataset has been disambiguated and annotated with the corresponding sense IDs in GermaNet. We need these annotations for two reasons: to obtain attribute information about the phrases and to augment the data automatically.

Automatic annotation. To add the attribute annotations, we made use of the hierarchical structure of adjectives in GermaNet. Based on the manually annotated sense IDs of the adjectives, we assign an attribute label to each phrase automatically. For instance, *tief* 'deep/low' in *tiefe Stimme* 'deep voice' has been annotated with the sense "having a low pitch". The top node in the hierarchy for this sense is `perception` (see figure 1) – the phrase is assigned this label as an attribute. In *tiefe Liebe* 'deep love', the adjective is annotated with a different sense – "very strong, intense", the attribute label for this sense is `intensity`. Table 1 provides an overview of all the 16 labels with examples from the dataset (codenamed GerCo+).

Collocations. Half of the GerCo+ dataset is represented by collocations. Their distribution, however, is not balanced for each attribute. It concurs with the previous observations in literature that certain meanings tend to be expressed collocationally and certain meanings are usually found in free phrases. For instance, `intensity` is usually expressed in collocations whereas `color` in free

[1] The dataset, the splits and the code for running the models on the data are available at `https://github.com/Blubberli/IWCS-attributes.git`

[2] Based on the logDice score (Rychly, 2008); 75% of the data has a logDice $>$ 4.14.

[3] Based on the semantic classification of German adjectives proposed by Hundsnurscher and Splett (1982).

attribute	**example**
behaviour	*frecher Bursche* ‘rude guy’
body	*blindes Kind* ‘blind child’
climate	*windiger Tag* ‘windy day’
evaluation	*herrliches Wetter* ‘wonderful weather’
feeling	*bitteres Lachen* ‘bitter laugh’
intensity	*leichter Regen* ‘light rain’
location	*tiefer See* ‘deep lake’
manner	*wilder Tanz* ‘wild dance’
intelligence	*schlauer Junge* ‘smart boy’
motion	*starres Gesicht* ‘rigid face’
quantity	*karger Lohn* ‘meager salary’
perception	*schwarzer Rock* ‘black skirt’
relation	*sicherer Tod* ‘certain death’
society	*reiche Verwandten* ‘rich relatives’
substance	*grober Sand* ‘coarse sand’
time	*alter Freund* ‘old friend’

Table 1: Attributes in the GerCo+ dataset.

phrases (van der Wouden, 1997). Figure 2 shows the frequency distribution of collocations and free phrases in GerCo+. Four labels (intensity, relation, manner, feeling) are represented to a large extent by collocations, for perception, substance, on the other hand, the number of free phrases is very high. We expect collocations to be more challenging for the models.

Additional adjectives. The number of distinct adjectives in the original GerCo dataset is small. For some attributes (e.g. evaluation), very few adjectives are available. To be able to test each attribute for at least three distinct adjectives, we added 8 adjectives. We manually combined them with suitable nouns from the original dataset and annotated the phrases with the corresponding attributes. The adjectives in the final dataset can select between one and six different attributes (see figure 3). Most of the adjectives can select more than one attribute: this ambiguity is expected to pose another challenge for the automatic modelling.

3.2 Automatic augmentation

Lexical memorization is the tendency of a classifier to memorize the relations between words it has seen in training and corresponding labels (Levy et al., 2015). The generalisation ability of classifiers and the phenomenon of *lexical memorization* in classifying lexical inference relations and relations in noun compounds have been investigated by Levy et al. (2015); Dima (2016); Shwartz and Waterson (2018). Since the GerCo+ dataset is rather small, the danger of the classifier falling into the trap of lexical memorization effects needs to be safeguarded against. We therefore propose an automatic data augmentation to be able to create different training and test splits: either with modifier overlap, with head overlap or no overlap. We also expect a larger dataset to have positive effects on the precision of the machine-learning models. In order to increase the amount of training data, we perform automatic data augmentation relying on lexical and conceptual relations in GermaNet.

In GermaNet, senses of words are grouped into sets of synonyms (synsets). Synsets are connected to each other via conceptual relations, the main type of such relations is hyponymy/hypernymy as in *pie→pastry→baked goods*. Apart from that, some lexical units are interlinked via lexical relations, such as synonymy and antonymy. Attributes are expected to carry over to adjectives and nouns linked in GermaNet via lexical and conceptual relations. Knowing the sense IDs of all the words in the dataset, we therefore only have to extract the semantically related adjectives and nouns to generate new phrases. The new phrases are annotated automatically with the attribute from the original phrase. For instance, the original dataset contains the phrase *tiefer Ton* ‘low-pitched sound’ (collocation) with the attribute perception. Both words are provided with the corresponding sense IDs from GermaNet. The antonym of *tief* in this sense is *hoch* ‘high-pitched’ and a co-hyponym of *Ton* is *Pfeifen* ‘whistle’. This results in a new phrase *hohes Pfeifen* ‘high-pitched whistle’ with the attribute perception.

Further phrases can be extracted via the adjectival top nodes in GermaNet: by combining non-ambiguous adjectives under those nodes with nouns that can select the corresponding attribute. Selecting only non-ambiguous adjectives, i.e. only adjectives that select a single possible attribute ensures that the resulting phrases is annotated with the correct attribute. For example, a new phrase for the attribute perception can be constructed by combining the adjective *salzig* ‘salty’ which can only express this attribute with other nouns that can have perception, e.g. *Suppe* ‘soup’. We create two augmented datasets:

1. **small** Augment only the adjectives by adding synonyms, antonyms, direct hypernyms, all hyponyms and co-hyponyms
2. **large** Augment the adjectives and nouns by

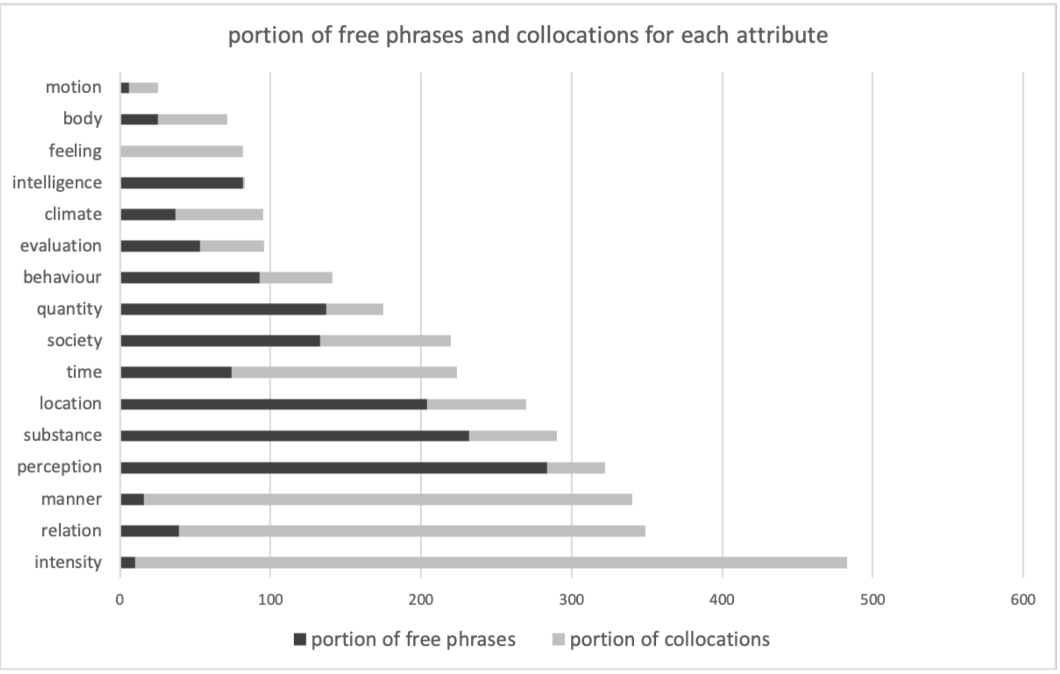

Figure 2: Distribution of free phrases and collocations in the GerCo+ dataset for each attribute.

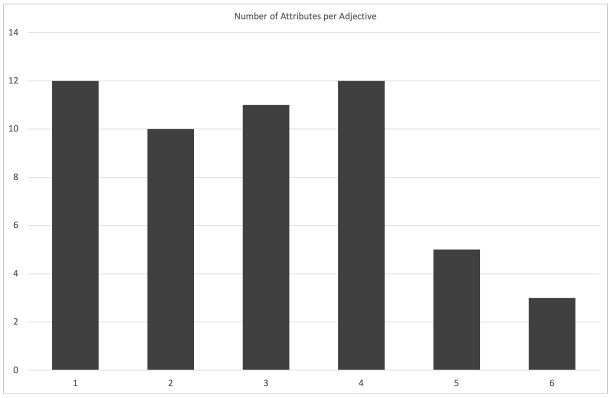

Figure 3: Distribution of the number of different attributes per adjective.

adding synonyms, antonyms, direct hypernyms, all hyponyms and co-hyponyms. Augment the attributes by combining all non-ambiguous hyponyms with suitable nouns.

In order to eliminate nonsensical phrases, the automatically created AN phrases are filtered by their bigram frequencies (>3) in a large corpus consisting of several German treebanks.[4]

Automatically augmented data is expected to be noisy to some extent. To estimate the amount of noise, we randomly extract 100 examples from each augmented dataset and manually assess the examples and the corresponding attributes. This study of random samples shows that around 20% of the automatically gained data is labeled incorrectly. Table 2 gives an overview of the data.

data	size	adj	nn	correct
gold standard	3,093	46	2,030	-
small	21,498	1,980	2,538	80%
large	232,389	4,630	36,659	79%

Table 2: Data overview: the amount of phrases, unique adjectives, unique nouns and the amount of correct phrases in the random sample extracted from each augmented dataset and evaluated manually.

3.3 Dataset splits

We create two test set ups: mixed and balanced. In the mixed setting, we test all the attributes and all the adjectives from the gold standard dataset. In the balanced setting, we use a subset of seven attributes with a balanced distribution of collocations and free phrases to compare the performance on the two types of phrases. The balanced attributes are `climate, quantity, time, society, location, behaviour, evaluation`.

The models are trained on the two automatically augmented datasets: small and large.

We create three splits of validation/test data from the gold standard GerCo+ dataset. Each test set contains roughly 700 phrases. To investigate the role of lexical memorization in the attribute selection task,

[4] TüBa-D/DP (de Kok and Pütz, 2019) and the corpus DECOW16AX (Schäfer, 2015; Schäfer and Bildhauer, 2012)

we create different lexical settings in the training data: (1) **No overlap** The validation/test and training have distinct vocabulary. (2) **Modifier overlap** The validation/test and training share modifiers (adjectives). (3) **Head overlap** The validation/test and training share heads (nouns).

4 Automatic classification

In the following experiment, we investigate to what extent attribute-selection can be computationally modeled. For that purpose, we use the data described in section 3.3 and train a simple neural network to predict one of the 16 possible attributes given the adjective and noun as input.

4.1 Modelling

We train a feed-forward non-linear classifier with one hidden layer. For each adjective-noun phrase, we extract the embedding for each constituent and apply a linear transformation to the concatenated input embeddings, followed by a `ReLU` non-linearity.

We experiment with two different embedding types:

- **fastText** (Bojanowski et al., 2017) non-contextualized German word embeddings with subwords trained on Common Crawl (Grave et al., 2018).
- **BERT** (Devlin et al., 2019) contextualized embeddings produced by a bidirectional transformer trained on Wikipedia, the EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl.[5] We treat the adjective-noun phrase as the context sentence, thus the embedding of the adjective is only contextualized given the noun (and the other way around respectively).

The size of the hidden layer corresponds to the embedding dimension of one constituent (300 for fastText, 768 for BERT), the output layer has size 16 which corresponds to the number of different attributes.

We optimize the cross-entropy loss with Adam and use class weights, with higher weights for the less frequent attributes because the distribution of the attributes is imbalanced. As BERT comes with 12 layers, we learn a scalar-weighted combination of them. We always apply a dropout of 0.8. As the best model, we pick the one that achieves the best macro F1 score on the validation set after not improving for 5 epochs.

We use two baselines: We train each model with either using only the adjective or only the noun embedding as input. For the contextualized embeddings, we use the respective embedding after contextualization.

Note that our goal was not to find the best model for the task but to investigate how well a simple model can generalize for the task if it has been trained on a sufficient amount of data.

4.2 Results and Evaluation

(i) Generalization One of the research questions we want to answer with the experiment is in which way the automatic models can learn abstractions only on the basis of semantically related adjective-noun pairs. If the model has seen phrases like *black limousine* and *yellow truck* in training, is it able to learn the abstract attribute `perception` and predict correctly for test phrases, such as *red car*? In the best case, although the model has neither seen *red* nor *car* in the training set, it can arrive at the correct solution via lexical similarities: it has learned that colors express `perception` when combined with e.g. artifacts.

As mentioned in Section 3.2, , it has been shown for other tasks in lexical semantics that the abstraction ability of automatic models in supervised learning is diminished if constituents of the phrase in the test set have already occurred in training. It may then be easier for the model to memorize the most frequent or only class label for specific words to solve the task. We investigate to what extent that phenomenon applies to attribute selection. Especially for adjectives that occur with only one attribute, this effect would be expected. This phenomenon could have a particularly negative effect for ambiguous adjectives: In the worst case, lexical memorization overwrites the less frequent sense as only the most dominant attribute is predicted.

Table 3 shows the results for both embedding types for the different training data and the adjective and noun baseline. We report the average macro F1 score for all attributes, so each attribute is scored equally, regardless of the number of test instances.

First, it becomes clear that both models are capable of abstracting to some degree with fastText outperforming BERT by 6%. It is particularly interesting that there is hardly any difference between

[5] https://github.com/dbmdz/berts

the small and the large data set, although the large data set contains ten times more training instances. This demonstrates that it is not the size of the training data alone that matters for the generalization ability of the models. A sufficient lexical variety is much more important. This variety seems to be covered in the smaller training data set, such that an increase in size does not have a large effect on the general result. It is also evident that a partial overlap of adjectives and nouns leads to a significant improvement especially for BERT. This effect is similar on the smaller data set for modifier and head overlap, on the larger one a modifier overlap brings more advantages. The number of unique nouns is much higher in this data set, so it is less likely that lexical memorization can occur with the head overlap.

The results for the adjectives and noun baseline illustrate that while it is necessary to have both constituents as input for the models with fastText embeddings, the contextualization of the BERT embeddings is sufficient to convey almost the same information via one of the two contextualized vectors. In both cases the adjective baseline is stronger, indicating that the adjective plays a more important role for the task than the noun.

(ii) Attributes Figure 4 and Figure 5 show the performance for each attribute on the large dataset, for no overlap, modifier overlap and head overlap. The attributes `time, climate, perception` and `evaluation` can be learned particularly well without overlap. A possible explanation is that adjectives and nouns selecting these attributes have a high semantic similarity. For example, adjectives selecting `time` are more similar to each other than adjectives selecting `intensity`. For such attributes, the generalization is more difficult. For instance, `manner` and `intensity` are not easy to predict despite a high amount of training data (14,084 and 8,714 training instances). Attributes that benefit most from lexical overlap are `body, feeling, behavior`, and `motion`.

(iii) Polysemy With respect to lexical memorization, the findings here are mixed. While across-the-board improvements for each attribute with modifier or head overlap indicate that this phenomenon takes place, the partial overlap does not automatically lead to predicting the attribute for the polysemous adjectives that has the highest frequency in the training data. Table 4 depicts how many of all the possible attributes for the ambiguous adjectives in the test set are covered. We sum the number of correctly recognized attributes for each adjective. Out of the total of 144, roughly two thirds are recognized by the models for each setup, the number is even higher for the modifier overlap. For instance, in the case of the adjective *zart* 'tender', `substance, intensity` and `manner` were recognized without overlap, while `body` was additionally recognized with the modifier overlap. Table 5 shows the average accuracy for adjectives with different degrees of ambiguity regarding their possible attributes. A lower degree of ambiguity leads to better results. For a higher degree of ambiguity the modifier overlap brings significant improvements so the models can learn to better distinguish the different senses for the adjectives based on the training data. It is also worth noting that there is a considerable jump in accuracy when we compare adjectives that co-occur with four or more attributes with those that select at most three attributes.

training data	**fastText**			**BERT**		
			small			
	both	*adj*	*noun*	*both*	*adj*	*noun*
no overlap	0.50	*0.42*	*0.29*	0.44	*0.44*	*0,33*
modifier overlap	0.66	*0.45*	*0.38*	0.61	*0.61*	*0.49*
head overlap	0.67	*0.45*	*0.46*	0.61	*0.59*	*0.56*
			large			
no overlap	0.53	*0.45*	*0.24*	0.45	*0.41*	*0.38*
modifier overlap	0.68	*0.49*	*0.26*	0.71	*0.68*	*0.62*
head overlap	0.60	*0.47*	*0.31*	0.57	*0.53*	*0.52*

Table 3: Average Macro F1 Score over all attributes for each training set. The results are presented for training on the adjective and noun (both), and for the two baselines: trained only on adjectives (adj) and only on nouns (noun)

training set	**no overlap**	**modifier overlap**	**head overlap**
fastText	97	105	99
BERT	95	105	99

Table 4: Number of correctly predicted senses of polysemous adjectives for each embedding type and each training setup trained on the large dataset; the total number of different senses in the test data: 144.

(iv) Transparency To investigate the difference in the performance between collocations and free phrases, we use a smaller balanced test set (described in Section 3.3). Table 6 presents the results as the average of the Macro F1 scores of all 7 attributes in the test set.

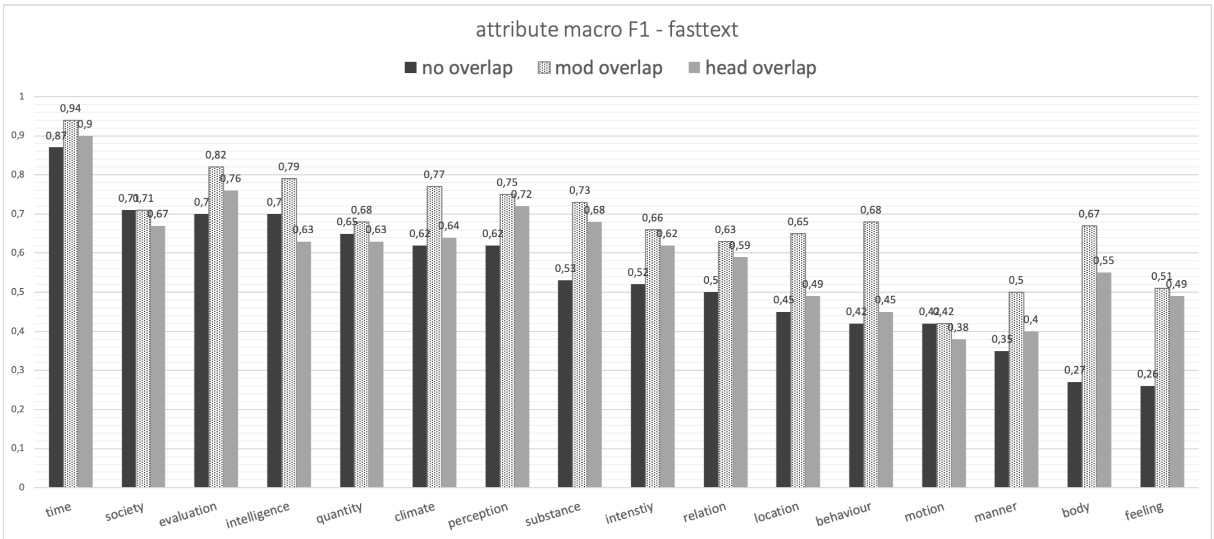

Figure 4: General Macro F1 for each attribute for fastText – each training set

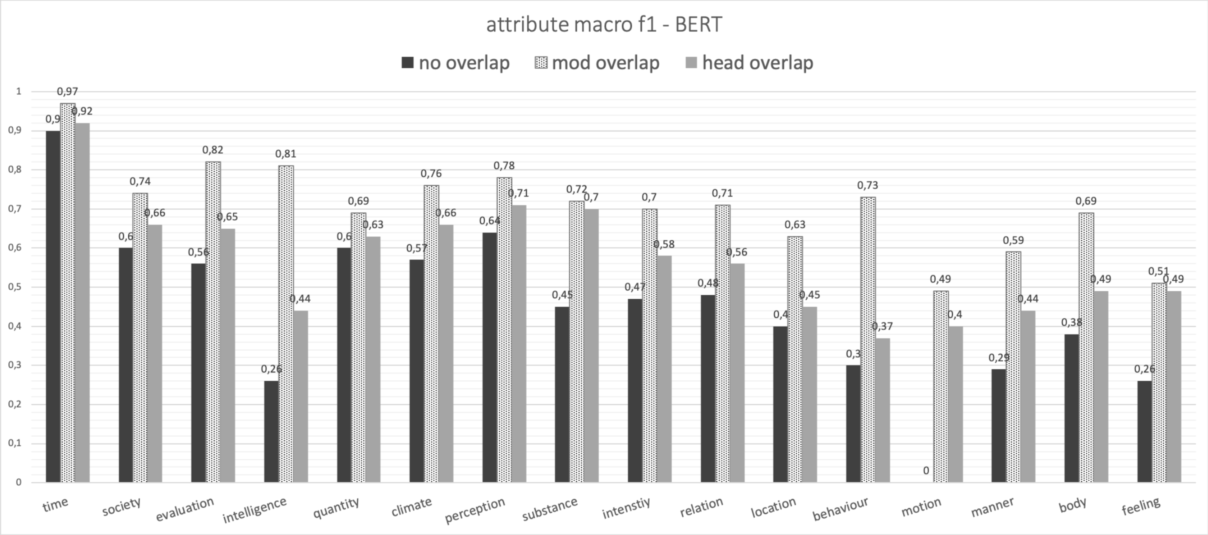

Figure 5: General Macro F1 for each attribute for BERT – each training set

no. attr	**fastText**			**BERT**		
	no	**mod.**	**head**	**no**	**mod.**	**head**
6	0.40	0.47	0.45	0.34	0.60	0.37
5	0.29	0.48	0.41	0.26	0.57	0.35
4	0.33	0.56	0.48	0.32	0.60	0.51
3	0.70	0.75	0.70	0.58	0.78	0.69
2	0.61	0.72	0.71	0.48	0.78	0.60
1	0.80	0.93	0.84	0.77	0.95	0.80

Table 5: Average accuracy for all adjectives with a specific number of possible attributes (no. attr) for the setup with no overlap (no), modifier overlap (mod) and head overlap (head).

Overall, there is a consistent difference between collocations and free phrases across all training data: free phrases are more accurately predicted in all cases. Contextualized embeddings were expected to yield better results for collocations because they are dynamically conditioned on the local context. Therefore, adjective and noun are represented by different vectors for different phrases. However, the model with BERT embeddings is worse if no lexical overlap is present. One reason for this may be that the contextualization of BERT does not give an advantage for a word-based task. It is more difficult to find regularities because the similarities between words could become blurred due to contextualization.

Although the performance for collocations is worse than for free phrases in general, for some attributes, the models are successful. This finding confirms the hypothesis that there are regularities also for collocations in spite of the general assumption of their idiosyncrasy. For instance, the attribute `climate` has a high F1 score for collocations in all experimental settings (between 0.67 and 0.87). It indicates that meaning shifts of the adjectives selecting this attribute are regular. Another example of such a regular meaning shift is provided by the polysemous adjective *süß* ‘sweet’. In its literal meaning, it refers to the attribute `perception` as in *süße Torte/Tee* ‘sweet cake/tea’. However, *süß* can also refer to the attribute `evaluation` when it is combined for instance with nouns from the semantic field ‘person’, as in *süßes Kind* ’sweet child’.

By contrast, other collocations are highly lexicalized. These cases are hard to classify and remain a challenge. For instance, the models fail to predict the attribute `evaluation` for examples such as *helle Zukunft* ‘bright future’.

training data	**fastText**		**BERT**	
	free phrase	collocation	free phrase	collocation
		small		
no overlap	0.66	0.53	0.59	0.44
modifier overlap	0.74	0.57	0.67	0.59
head overlap	0.80	0.73	0.73	0.67
		large		
no overlap	0.73	0.61	0.62	0.58
modifier overlap	0.84	0.73	0.87	0.72
head overlap	0.75	0.61	0.67	0.63

Table 6: Average Macro F1 score for the balanced set in terms of collocations and free phrases for each training set.

5 Conclusion and future work

In this paper we present a study on attribute selection in German adjective-noun phrases. Experiments in different training settings with and without lexical overlap show that it is possible to learn attribute selection patterns based on semantically related adjectives and nouns: abstract attributes such as `perception`, `time`, or `society` can be learned and predicted for new, unseen data.

The results of the experiments with different lexical overlap settings are in line with previous research: partial lexical overlap leads to better results on this task. However, this is not only due to lexical memorization. The models are still able to decide which attribute to select for an ambiguous adjective in the test set if it appears in training with all its possible meanings, based on the nouns combined with.

The experiments confirm that attributes are more difficult to predict for collocations than for free phrases. However, not all types of collocations are equally difficult. Attributes can be learned correctly for collocations when the meaning shift occurs systematically. Strongly lexicalized collocations cannot benefit from these regularities.

As future work it would be interesting to investigate attribute-selection in other languages, e.g., in Russian. Compounding in Russian is not as productive as in German and the function of compounds is often taken over by adjective-noun phrases, so a higher degree of lexicalization would be expected. This could result in an even greater difference between collocations and free phrases. Secondly, it would be interesting to investigate how using a full sentence as context impacts the results, especially in ambiguous cases. For instance, the phrase *stürmischer Tag* ‘stormy day’ can either express the attribute `climate` when the adjective is used in its literal sense or the attribute `manner` when *stormy = chaotic*. For such phrases, disambiguation is only possible in context. Finally, it would be useful if a model could learn a general intuition about whether a phrase is a collocation or a free phrase and which attributes are selected by an adjective in its literal and collocational senses.

Acknowledgments

We would like to thank our student assistants Daniela Rossmann, Alina Leippert and Mareile Winkler for their help with the annotations. We are also very grateful to the anonymous reviewers for their insightful and helpful comments that helped us to improve the paper. Financial support of the research reported here has been provided by the grant *Modellierung lexikalisch-semantischer Beziehungen von Kollokationen* awarded by the Deutsche Forschungsgemeinschaft (DFG).

References

Lawrence W. Barsalou. 1992. Frames, concepts, and conceptual fields. In *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pages 21–74. Lawrence Erlbaum Associates, Inc.

Melanie J. Bell and Martin Schäfer. 2013. Semantic transparency: challenges for distributional semantics. In *Proceedings of the IWCS 2013 Workshop Towards a Formal Distributional Semantics*, pages 1–10, Potsdam, Germany. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Corina Dima. 2016. On the compositionality and semantic interpretation of English noun compounds. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 27–39, Berlin, Germany. Association for Computational Linguistics.

Luis Espinosa Anke, Steven Schockaert, and Leo Wanner. 2019. Collocation classification with unsupervised relation vectors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5765–5772, Florence, Italy. Association for Computational Linguistics.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Hanshin Publishing Co., Seoul, South Korea.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jeffrey S. Gruber. 1965. *Studies in Lexical Relations*. Ph.D. thesis, MIT. Distributed by: Indiana University Linguistics Club, Bloomington, Indiana.

Birgit Hamp and 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*.

Matthias Hartung. 2015. *Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns*. Ph.D. thesis, Heidelberg University, Germany.

Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 54–64, Valencia, Spain. Association for Computational Linguistics.

Verena Henrich and Erhard Hinrichs. 2010. GernEdiT - The GermaNet Editing Tool. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 2228–2235.

Franz Hundsnurscher and Jochen Splett. 1982. *Semantik der Adjektive des Deutschen*. VS Verlag für Sozialwissenschaften.

Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.

Abhik Jana, Dima Puzyrev, Alexander Panchenko, Pawan Goyal, Chris Biemann, and Animesh Mukherjee. 2019. On the compositionality prediction of noun phrases using poincaré embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3263–3274, Florence, Italy. Association for Computational Linguistics.

Daniël de Kok and Sebastian Pütz. 2019. *Stylebook for the Tübingen Treebank of Dependency-parsed German (TüBa-D/DP)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Omer Levy, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 970–976, Denver, Colorado. Association for Computational Linguistics.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Pavel Rychly. 2008. A Lexicographer-Friendly Association Score. In *Sojka, Petr /Horák, Aleš (Hg.): Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008*, pages 6–9, Brno.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, pages 28–34, Lancaster, UK. UCREL, IDS.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Vered Shwartz and Chris Waterson. 2018. Olive oil is made *of* olives, baby oil is made *for* babies: Interpreting noun compounds using paraphrases in a neural model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 218–224, New Orleans, Louisiana. Association for Computational Linguistics.

Yana Strakatova, Neele Falk, Isabel Fuhrmann, Erhard Hinrichs, and Daniela Rossmann. 2020. All that glitters is not gold: A gold standard of adjective-noun collocations for German. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4368–4378, Marseille, France. European Language Resources Association.

Ton van der Wouden. 1997. *Negative Contexts. Collocation, polarity, and multiple negation*. Routledge.