Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4368–4378
Marseille, 11–16 May 2020

All That Glitters is Not Gold: A Gold Standard of Adjective-Noun Collocations for German

Yana Strakatova[1], **Neele Falk**[1], **Isabel Fuhrmann**[2], **Erhard Hinrichs**[1], **Daniela Rossmann**[3]
[1,3]University of Tübingen, [2]Berlin-Brandenburg Academy of Sciences
[1]firstname.lastname@uni-tuebingen.de
[2]fuhrmann@bbaw.de
[3]daniela.rossmann@student.uni-tuebingen.de

Abstract

In this paper we present the GerCo dataset of adjective-noun collocations for German, such as *alter Freund* 'old friend' and *tiefe Liebe* 'deep love'. The annotation has been performed by experts based on the annotation scheme introduced in this paper. The resulting dataset contains 4,732 positive and negative instances of collocations and covers all the 16 semantic classes of adjectives as defined in the German wordnet GermaNet. The dataset can serve as a reliable empirical basis for comparing different theoretical frameworks concerned with collocations or as material for data-driven approaches to the studies of collocations including different machine learning experiments. This paper addresses the latter issue by using the GerCo dataset for evaluating different models on the task of automatic collocation identification. We compare lexical association measures with static and contextualized word embeddings. The experiments show that word embeddings outperform methods based on statistical association measures by a wide margin.

Keywords: MultiWord Expressions & Collocations, Semantics, Statistical and Machine Learning Methods

1. Introduction

Lexical collocations (further simply collocations) such as *golden memory* and *old friend* are multi-word units that consist of a *base*, chosen freely by a speaker, and a *collocate*, the choice of which is restricted depending on the base (Mel'čuk, 2012). In that aspect, they differ from free phrases, e.g. *golden crown* or *old lady*, where neither of the constituents is lexically constrained. However, collocations are not fully lexicalized as opposed to semantically opaque idiomatic expressions such as *golden ticket* 'good opportunity' or *old flame* 'former romantic partner'. In the recent decades, collocations have been extensively studied with the focus predominantly on their statistical properties and methods of automatic collocation extraction (Church and Hanks, 1990; Smadja, 1993; Evert, 2004; Pecina, 2008a; Bouma, 2009; Evert et al., 2017; Garcia et al., 2019). Identifying and extracting collocations, either manually or automatically, is a challenging task that requires clear definitions of concepts and reliable tools and resources. In spite of the growing interest in collocations, there exist only a few resources that can serve as gold standards in collocation research.

This paper reports on the construction of the dataset annotated by experts that comprises 4,732 instances of collocations and non-collocations (free phrases, idioms, named entities, terms). The dataset includes only adjective-noun phrases as they have been studied less extensively than verbal collocations. In our current research, we focus on the German language since digital resources and tools are available for German that identify relevant adjective-noun co-occurrences. The DWDS (short for Digitales Wörterbuch der deutschen Sprache) (DWDS, 2019) and its collocation extraction application the Wortprofil[1] serve as our empirical basis, and the German wordnet GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) provides information about the semantic classes of adjectives and nouns. The data cover all the 16 semantic classes of adjectives defined in GermaNet. The dataset serves as the empirical basis for the lexicographic work on extending the Wortprofil and for the enrichment of GermaNet with new lexical relations. The dataset is intended to be used not only for linguistic studies of collocations, but also in computational linguistics. As the dataset contains both positive and negative instances of collocations, it can serve as a suitable resource for evaluating different models of automatic collocation extraction and/or classification. This paper presents experiments based on two different types of models: we compare models based on association measures with models that build on word embeddings. The experiments show that embeddings outperform methods based on statistical association measures by a wide margin.

The remainder of this paper is structured as follows: Section 2. presents the related work on collocations and the existing data collections. In Section 3., we introduce the annotation scheme adopted for our collocation dataset and discuss the results of the annotation and the inter-annotator agreement. In Section 4., we report on the results of a series of machine learning experiments on the dataset. Section 5. describes the two ongoing studies on the further semantic enrichment of the dataset. We conclude the paper with a brief summary of the presented work and discuss the planned future work.

2. Related work

Collocations are numerous in every language, and due to their idiosyncratic nature, they pose considerable problems for non-native speakers. Thus, there is a strong need to include them in dictionaries, and they have received considerable attention in lexicography. There are a few specialized collocation dictionaries for different languages. For En-

[1] `https://www.dwds.de/wp/`, last accessed November 22, 2019

glish, there are the Oxford Collocations Dictionary for students of English (McIntosh et al., 2009) and the Macmillan Collocations Dictionary for Learners of English (Rundell, 2010). For German, the Wörterbuch der Kollokationen im Deutschen (Quasthoff, 2011), for Spanish - an Online Collocation Dictionary of Spanish (Vincze et al., 2011). The latter one includes semantic grouping of collocations based on Mel'čuk's Lexical Functions (Mel'čuk, 1996), but covers only the semantic class of emotions. Lexical Functions have also been extensively used in lexicographic projects on French collocations and semantic derivations (Polguere, 2000).

In recent decades, collocations have received much attention in lexical and computational linguistics. Most of the studies in computational linguistics are concerned with the statistical properties of collocations and focus on improving the methods for automatic collocation extraction, based on different association measures (AMs). Such studies require gold standard evaluation datasets. Evert (2004) uses the dataset of 21,796 German PP-verb combinations (`German_PNV_Krenn`) in his experiments, manually annotated as lexical collocations or non-collocations by Brigitte Krenn (Krenn, 2000). However, Evert (2004) emphasizes that it is not possible to generalize the results of experiments on verbal collocations to other types of collocations, i.e. with different syntactic relations. Thus, to investigate the properties of adjective-noun collocations, a database of German adjective-noun collocations has been created (Evert, 2008). The database (codenamed `Lallt`) is a collection of 1,252 collocation candidates annotated by lexicographers and classified into six subgroups that can be further generalized to two classes: *collocations* vs *non-collocations*. This dataset is one of the two collections known to us which features both positive and negative instances of adjective-noun collocations.

A second collocation database with positive and negative instances we are aware of, the Czech PDT-DEP dataset, is presented in Pecina (2008a). It comprises 12,232 dependency bigrams including verbal, nominal, and adjectival ones. All the candidates have been manually annotated, the true collocations in the dataset include stock phrases, named entities, support verb constructions, technical terms, and idiomatic expressions. The Czech PDT-DEP dataset and the two German datasets (Krenn, 2000; Evert, 2008) serve as training and evaluation material in the experiments on automatic collocation extraction reported in Pecina (2008b). In addition to the approach adopted by Evert (2004), where individual association measures are evaluated, Pecina (2008b) combines 55 association measures and uses them as features in different kinds of classifiers.[2] The results of the experiments illustrate how challenging the task of automatic collocation extraction is and that the performance differs depending on the data and the task. This issue is closely examined in the large-scale evaluation study by Evert et al. (2017). The study concludes that individual AMs yield dramatically different results depending on the gold standard and thus on the definition of the studied phenomenon; other parameters that influence the AMs' performance are the size and the quality of the corpus (Evert et al., 2017). Similar conclusions are drawn in Garcia et al. (2019), where 12 AMs are compared in the experiments on collocation extraction in English, Portuguese, and Spanish. They also show that the performance of individual AMs is similar for these languages.

The approach for constructing the above described datasets is to randomly extract a certain amount of dependency bigrams, filter them based on their frequencies, and give the list of candidates to annotators. The databases created in this way contain a wide variety of bases and their collocates, but no information about the semantics of the phrases.

A different kind of a collocation dataset is described in Espinosa-Anke et al. (2019). The `LexFunc` dataset comprises 10,077 English collocations annotated with relations in terms of Mel'čuk's Lexical Functions (Mel'čuk, 1996) where each keyword is disambiguated. `LexFunc` includes only positive instances of collocations and has been used in multi-class machine learning experiments for classifying the semantic relations that hold between the constituents of collocations.

Our main motivation for creating the GerCo dataset was to provide an annotation scheme that can be used for the following classification tasks: for the binary classification between collocations and non-collocations, or for a multi-class classification of free phrases, collocations, terms, idioms, and named entities.

3. Dataset Construction

In our work, we build upon the knowledge about statistical properties of collocations and use a sketch-engine-like platform the Wortprofil (Geyken et al., 2009) to select a list of collocation candidates. The Wortprofil provides lists of statistical co-occurrences for a given word based on the frequencies obtained from the DWDS corpora. In the Wortprofil, the co-occurrences are classified according to their grammatical relations (*is an attribute of, is a subject of, etc.*) and are sorted according to their logDice scores (Rychly, 2008). A high score for a phrase serves as an indication that the phrase may be lexically restricted.

3.1. Annotation

We follow a systematic approach to collocation analysis and aim at covering different semantic classes of the German lexicon. We rely on the German wordnet GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) in choosing the adjectives for analysis: there are 16 semantic classes for adjectives in GermaNet. From each class we randomly selected three adjectives. Table 1 gives an overview of the 48 adjectives selected for investigation together with their semantic classes as defined in GermaNet. The co-occurring nouns were obtained from the Wortprofil that provides maximum 100 candidates for each adjective (relation *"is an attribute of"*). This resulted in a collection of 4,732 adjective-noun pairs that were given to two annotators. The annotation of the dataset has been performed by two native speakers of German: Annotator 1 is an expert in lexical semantics, Annotator 2 is an advanced student of Computational Linguistics with a solid background in

[2] Linear Logistic Regression, Linear Discriminant Analysis, Neural Networks with 1 and 5 units in the hidden layer.

semantics. The decision tree in Figure 1 illustrates our approach to identifying lexical collocations.

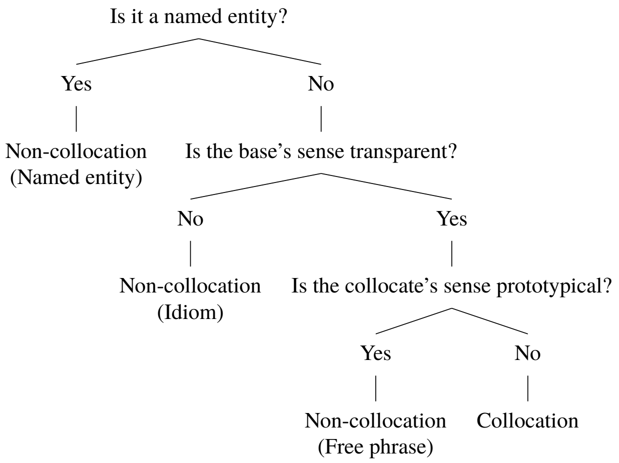

Figure 1: Annotation scheme for classification of the collocation candidates.

A word pair is classified as a *collocation* based on two criteria: (1) the meaning of the base is transparent; (2) the meaning of the collocate is not *prototypical*. The prototypical meaning of an adjective is its basic, most literal sense (German: “Grundbedeutung"). We rely on the DWDS dictionary in assigning the prototypical meaning to each adjective in the dataset. Consider the adjective *tief* ‘deep’: its prototypical sense is ‘reaching from top to bottom’. However, in the phrase *tiefe Liebe* ‘deep love’ the adjective is not used in its basic meaning and means ‘intense’. The noun ‘love’ is used in its literal sense, thus, this pair is classified as a collocation. Another example of a collocation is the phrase *alte Regierung* ‘old government‘: the adjective ‘old’ in this case is interpreted as ‘previous’ and is not used in its prototypical sense ‘indicates the age’.

We further distinguish collocations from other lexically restricted phrases in which the meaning of the adjective is not prototypical:

- Idioms: phrases where the meaning of the noun (base) is figurative. For instance, in the combination *alter Hase* (lit. ‘old rabbit’) the noun *Hase* ‘rabbit’ is not used in the sense ‘animal’, the phrase as a whole is used figuratively and means ‘old hand’, ‘having great expertise’. Other examples of idioms are *leichten Herzens* ‘with a light heart’ or metonymic expressions such as *offenes Ohr* lit.‘open ear’, fig. ‘sympathetic ear’ where ‘ear’ refers to a person.
- Named entities: e.g. *Alte Försterei* (lit. ‘old forester’s house’) is the name of a football stadium located in Berlin, *Blinder Schacht* ‘Blind shaft’ is a movie title.

The above desribed categories are opposed to lexically free phrases in which the meaning of the adjective (collocate) is prototypical. Such expressions as *alter Mann* ‘old man’, *tiefes Wasser* ‘deep water’, and *offenes Fenster* ‘open window’ are considered free phrases.

Semantic class	**Adjectives**
General	*herrlich*‘wonderful’, *knapp* ‘scarce’, *stark* ‘strong’
Movement	*sanft* ‘mellow’, *starr* ‘rigid’, *wild* ‘wild’
Feeling	*bitter* ‘bitter’, *süß* ‘sweet’, *zart* ‘delicate’
Cognition	*dumm* ‘stupid’ , *hell* ‘bright’, *schlau* ‘smart’
Society	*arm* ‘poor’ , *blank* ‘broke’, *deftig* ‘solid’
Body	*blind* ‘blind’, *dick* ‘fat’, *zäh* ‘tough’
Quantity	*prall* ‘full’, *reich* ‘rich’, *teuer* ‘expensive’
Phenomenon	*karg* ‘sparse’ , *mild* ‘mild’, *stürmisch* ‘stormy’
Location	*rund* ‘round’, *steil* ‘steep’, *tief* ‘deep’
Pertonym	*barock* ‘baroque’, *historisch* ‘historical’, *steinig* ‘stony’
Perception	*dunkel* ‘dark’, *scharf* ‘sharp’, *schwarz* ‘black’
privative	*frei* ‘free’, *tot* ‘dead’, *windig* ‘windy’
Relation	*leicht* ‘light’, *mächtig* ‘powerful’, *sicher* ‘safe’
Substance	*grob* ‘coarse’, *hölzern* ‘wooden’, *offen* ‘open’
Behaviour	*frech* ‘bold’, *hart* ‘tough’, *rau* ‘rough’
Time	*alt* ‘old’,*frischgebacken* ‘recent’, *spät* ‘late’

Table 1: Adjectives in the dataset and their semantic classes according to GermaNet.

3.2. Inter-Annotator Agreement

To assess the consistency of the annotation, we calculate the inter-annotator agreement (IAA) for the initial non-adjudicated data. The established practice is to use a standard measure that takes into account the probability of random agreement between the annotators. Cohen’s Kappa (κ) (Cohen, 1960) is a suitable measure in a binary classification task with two annotators. The IAA yields Cohen’s κ of 0.80. It indicates that in spite of the complexity of the task and the general vagueness of the concept of collocation, the annotation guidelines provide enough information to make the performance of the annotators consistent.

However, there were a number of disagreement cases. Table 2 gives an overview of agreement for each adjective in the dataset after the initial annotation. The annotators reach perfect agreement when an adjective has very few senses and they can be clearly distinguished from one another. For example, the adjective *hölzern* has only two senses: the prototypical one ‘wooden’ which belongs to the domain of concrete concepts and the non-prototypical abstract one ‘awkward’, and both annotators correctly identified one case of the non-prototypical usage in *hölzerner Dialog* ‘awkward dialog’. In contrast, all the senses of the adjective *historisch* ‘historical’ are very abstract and the interpretation of each adjective-noun pair is highly subjective which reflects in the agreement of only 66%. The same is true for the other adjectives with lower agreement: *sicher* ‘secure, safe’, *frei* ‘free’, *herrlich* ‘wonderful’ - they all have only abstract interpretations. Consider the adjective *sicher*: the definition from the DWDS dictionary assigned to be prototypical is ‘not threatened by danger, safe’. The prototypical meaning is conveyed in phrases such as *sicherer Abstand* ‘safe distance’ or *sichere Zone* ‘safe area’. However, there are combinations, such as *sicherer Arbeitsplatz* ‘secure job’ where *sicher* rather means ‘stable’, which can be also interpreted as ‘safe from the danger of getting fired’.

All the disagreement cases have been discussed and resolved by the two annotators in the process of adjudication. The most problematic cases are the ones that allow for more than one interpretation due to the ambiguity of the noun or the adjective. For instance, the word pair *old friend* has two readings depending on the context: it either refers to the `age` of the friend or it emphasises the `duration` of the

Adjective	N coll	Synsets	IAA %
deftig ‘savoury’	59	2	100%
frischgebacken ‘freshly baked’	73	1	100%
hölzern ‘wooden’	1	2	100%
stürmisch ‘stormy’	67	3	99%
teuer ‘expensive’	1	2	99%
blank ‘shiny’	74	4	98%
reich ‘rich’	32	3	98%
windig ‘windy’	66	2	98%
blind ‘blind’	57	4	97%
grob ‘coarse’	74	4	97%
rau ‘rough’	79	4	97%
steinig ‘stony’	3	1	96%
hell ‘bright’	21	4	96%
knapp ‘scarce’	79	4	96%
leicht ‘light’	90	7	96%
schlau ‘smart’	16	1	96%
süß ‘sweet’	54	4	96%
tief ‘deep’	71	6	96%
hart ‘hard’	86	10	95%
spät ‘late’	58	2	95%
starr ‘stiff’	92	2	94%
dumm ‘stupid’	22	3	93%
karg ‘sparse’	72	2	93%
offen ‘open’	80	5	93%
sanft ‘mellow’	87	3	93%
stark ‘strong’	90	5	93%
tot ‘dead’	12	3	93%
zäh ‘viscous’	90	2	93%
bitter ‘bitter’	81	4	92%
scharf ‘sharp’	94	10	92%
steil ‘steep’	17	2	92%
arm ‘poor’	28	3	91%
mächtig ‘powerful’	29	4	90%
dunkel ‘dark’	29	5	89%
prall ‘firm’	51	2	89%
wild ‘wild’	61	6	87%
rund ‘round’	29	4	86%
schwarz ‘black’	30	8	86%
barock ‘baroque’	0	2	85%
zart ‘soft’	79	5	85%
alt ‘old’	37	5	83%
frech ‘bold’	29	2	79%
mild ‘mild’	84	3	74%
herrlich ‘wonderful’	28	1	72%
frei ‘free’	57	9	71%
dick ‘thick’	40	6	70%
historisch ‘historical’	38	2	66%
sicher ‘safe’	58	4	63%

Table 2: The number of collocations and the raw agreement between the two annotators for each adjective in the dataset, sorted by IAA. The translations are given for the *prototypical* meaning of the adjectives.

friendship (in the sense of ‘longtime friend’). The Wortprofil links the word combinations to the corpus contexts where they occur. Relying on the provided context sentences, the annotators chose the most salient reading and assigned the labels accordingly. In the case of *old friend*, the final decision was to annotate it as a collocation.

The adjudicated dataset comprises 2,505 positive and 2,227 negative instances of collocations. The adjective-noun pairs identified as non-collocations have been further annotated by Annotator 2 as free phrases, idiomatic expressions, named entities, and terms. Free phrases make up the largest group of non-collocations in the dataset: 1,979 instances. Idiomatic expressions comprise 145 pairs; named entities 43 combinations. Apart from that, 18 pairs were annotated as terms (e.g. *dunkle Materie* ‘dark matter’). There are 42 cases, where the status of the expression depends on the context, for example the phrase *runder Tisch* ‘round table’ can be used either symbolically or literally, or *wilde Maus* ‘wild mouse’ can either be a free phrase or refer to a name of a roller coaster.

4. Application: Building Classifiers to Detect Collocations

This section exemplifies how the presented gold-standard collection can be used to develop and evaluate different models for collocation classification. The dataset could be used for multi-class classification using the annotations described in Subsection 3.2. (free phrases, collocations, idioms, named entities, terms). However, since not all phrase types have a similar amount of instances in our dataset (e.g. there are only 145 idioms) and our main interest focuses on collocations, we define a binary classification task and distinguish between two classes: lexically free phrases vs. collocations. Detecting and modelling collocations is important for tasks like natural language generation (because the generated language should sound natural) or machine translation (because the collocation cannot be translated literally). We investigate what feature representations provide a useful source of information to solve the task. In the first experiment, we test whether lexical association measures contain enough distinctive information to discriminate between the two classes. As the meaning of the adjective diverges from the prototypical meaning in collocations, representations based on co-occurrence frequencies alone are not likely to perform well. Thus, in a second line of experiments, we examine different types and combinations of word embeddings which contain more information about the meaning of words and consequently are more likely to be applicable for this task. We are interested in finding out whether static word embeddings of the base and the collocate suffice for the classification or whether deep-contextualized word representations are more useful because of their ability to model different meanings of the same word depending on the context.

4.1. Data

In order to obtain context-aware representations of meaning, we extracted sentences containing the adjective-noun pairs from the GerCo dataset. The source corpora stem from different domains (encyclopedia, newspaper, blogs) and are publicly available. These corpora include Wikipedia (dumps from 2017, 2018, 2019), the One Million Posts Corpus (Schabus et al., 2017; Schabus and Skowron, 2018), the German proceedings from the EuroParl corpus (Koehn, 2005; Tiedemann, 2012) and the German Political Speeches Corpus (Barbaresi, 2018). We were able to extract context sentences for 3,652 free phrases and collocations. We randomly selected one context sentence per phrase with a sentence length between 15 and 30 words. To be able to analyse the performance for each of the 48 adjectives, we created six splits, each split containing a different set of adjectives in the test set. On top of that, the adjectives in the test set, are not present in the training data, thus it can be investigated whether the models can generalize above the word level. Figure 2 gives an overview of the class portion of each adjective. It indicates that the class portion per adjective varies, some adjectives occur only in the prototypical sense (*steinig* 'stony'), others are almost only used as a collocation (*frischgebacken* 'freshly baked'). In order to examine whether additional information about the meaning of the adjective helps to classify, we added the sense definitions for the basic literal sense for each adjective. Table 4 displays two training set instances for the adjective 'old' – each instance comes with a context sentence and a sense definition.

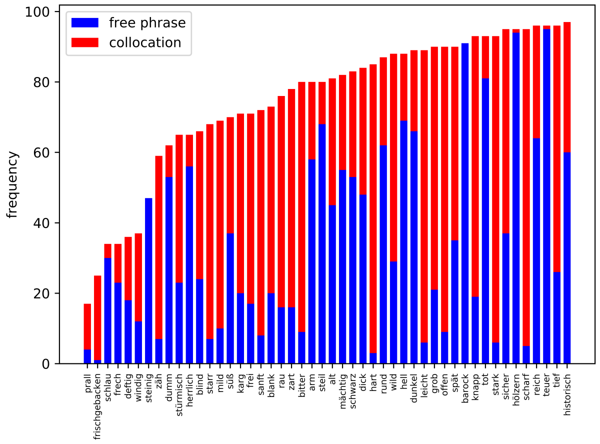

Figure 2: Class distribution for each adjective in the test set.

4.2. Association Measures

Previous work in the field of automatic collocation identification makes use of a variety of lexical association measures (AMs) to extract and rank a list of collocation candidates. Generally, these measures test whether the occurrence of the adjective-noun pair is statistically significant. They are computed based on the joint and individual frequencies of the base and collocate. The values produced by AMs can be viewed as a measure of the strength of association between base and collocate. Consequently higher values can indicate that an observed bigram is actually a collocation.

For the classification of collocations, any association measure can be used as a binary classifier by setting a threshold. Phrases that exceed a certain threshold can be classified as collocations, phrases with a score below the threshold fall into the class of free phrases. (Pecina and Schlesinger, 2006) As an alternative the association measures can be used as features for training a linear or non-linear classifier.

Previous work has revealed that the type of co-occurrence chosen as the basis for the computation of the association measures has an impact on the quality of the collocation extraction and classification, and that syntactic co-occurrence outperforms window-based approaches (Evert et al., 2017). For that reason, we extracted adjective-noun pairs[3] with an attributive dependency relation from three large, automatically annotated treebanks (Wikipedia 2017 and Wikipedia 2018 (de Kok and Pütz, 2019) , decow16ax (Schäfer and Bildhauer, 2012; Schäfer, 2015)). We used the UCS-toolkit[4] by Stefan Evert (Evert, 2004) to compute 22 different AMs, including standard measures, such as mutual information or log-likelihood.[5] In order to be able to use the measures as features for machine learning, we applied normalization and scaled each measure independently between 0 and 1.

We tuned the threshold for each AM on the training set and classified each test instances based on the best threshold. We then classify every instance in the test set based on a combination of the individual predictions of each AM classifier and use the majority (Bishop, 2006) as the final prediction.

Similarly to Pecina and Schlesinger (2006), we use the AMs as input to train a linear and a non-linear classifier that predicts the class based on a combination of AMs for a given adjective-noun pair. The idea is that the neural classifier learns a good internal feature representation given the available AMs as input. The weights of the classifier are optimized for minimizing the cross-entropy loss on the training set. An adjective-noun pair is represented by a vector of 22 dimensions, each dimension associated with a different type of association strength. For the linear classifier, we use a Support Vector Machine with a Radial Basis Function (RBF) kernel and l2 regularization[6]. We use a feed-forward neural network with one hidden layer of size 4 (hidden layer size was tuned on validation, see Appendix B) and a `ReLU` non-linearity and apply early stopping. We train each type of classifier on each split and measure the overall performance by taking the average of all splits.

model	validation accuracy	test accuracy
majority baseline	0.5408	**0.5344**
majority vote of threshold-classifiers	0.5350	0.5003
linear classifier (SVM)	0.5319	0.5028
combined AMs with nonlinear classifier	0.545	0.5256

Table 3: Average validation and test accuracy on all splits with association measures

The results in Table 3 show that association measures alone are not able to detect collocations. Even combining the association measures and mapping them into a new feature space does not lead to a performance better than that of the majority baseline. This might be due to the fact that

[3] 42,445,060 adjective-noun phrases in total

[4] `http://www.collocations.de/software.html`, last accessed November 22, 2019

[5] The complete list of AMs used in this study can be found in the Appendix A

[6] We use the standard implementation of sklearn (Pedregosa et al., 2011)

context sentence	phrase	sense definition	label
Im Zoo Rostock sind naturnah gestaltete Anlagen verbunden mit einer Parklandschaft mit alten Bäumen und Gehölzen . At Rostock Zoo, near-natural facilities are connected to a park landscape with old trees and shrubs.	*alten Bäumen* old trees	*gibt das Alter, die Lebensjahre an* indicates the age, the years of life	0
Sie muss jedoch bald feststellen , dass der Theaterclub aufgelöst wurde bzw. keine Mitglieder mehr hat , nachdem alle ihre alten Freunde die Schule abgeschlossen hatten . However, she soon has to realize that the theater club was dissolved or has no more members after all her old friends had finished school.	*alten Freunde* old friends	*gibt das Alter, die Lebensjahre an* indicates the age, the years of life	1

Table 4: Positive– (collocation, label = 1) and negative example (free phrase, label = 0) of training set instance for the adjective 'old'.

the instances in the dataset were already extracted using the logDice association measure, consequently all the pairs in the dataset have a relatively strong association. The results confirmed the hypothesis that that lexical association measures might be a good approach for extracting a number of collocation candidates, but that the feature representations based on these measures alone do not suffice for a more fine-grained and more semantically restricted classification task.

4.3. Word Embeddings

Since the findings from the first experiment indicate that additional semantic information is necessary, the second set of experiments makes use of a richer source of semantic information, namely word representations. Word embeddings encode semantic information about words as they are designed to capture information about similar words and words they co-occur with. Static word embeddings, such as GloVe (Pennington et al., 2014) or Word2Vec (Mikolov et al., 2013), represent the meaning of a word based on its distribution in language (large corpora) but suffer from the meaning conflation deficiency: all the possible senses of a word are represented by the same vector. Recent work in natural language processing has revealed that this issue can be overcome by computing dynamic representations of words conditioned on local context (Peters et al., 2018; Devlin et al., 2019). These representations are not only dynamic in the sense that they are able to capture different meanings of a word depending on the context, but they are designed to work well for predicting other words in contexts, making them general enough to be applicable for a wide range of natural-language applications out-of-the-box.

In the following line of experiments we would like to find out whether semantic representations of words in general are useful for detecting collocations and how much contextual information is needed to classify the examples correctly. If the sense of the adjective is mainly restricted by the noun, static word representations might suffice to solve the task. If the phrase in isolation is ambiguous or other words from the local context have a great impact on the meaning of the adjective, contextualized embeddings should be more helpful. We also examine whether definitions of the prototypical meaning of the adjectives are a valuable source of information for the classification.

We experiment with different setups. In each setup, a nonlinear classifier with one hidden layer and a `ReLU` nonlinearity is trained given different feature representations as input:

- adj + noun
- adj + noun + context sentence
- adj + noun + sense definition of prototypical sense
- adj + noun + context sentence + sense definition

With these experiments, we hope to gain insights about how much additional information is needed to correctly disambiguate the adjective. In the experiments with static embeddings, we used pretrained word embeddings[7] that were trained with the finalfrontier utility[8] on subcorpora (Wikipedia, Taz, EuroParl) of Tüba D/DP (de Kok and Pütz, 2019). The embeddings were trained with the structured skip-gram algorithm (Ling et al., 2015) and have a dimension of 300. This algorithm uses the architecture of skip-gram (Mikolov et al., 2013) and predicts context words given a target but preserves the structure of the context words during optimization. On top of that, the pretrained vectors were trained with subword embeddings (Bojanowski et al., 2016) and are thus capable of modeling out-of-vocabulary words.

For extracting contextualized embeddings, we use the bidirectional transformer (BERT) introduced by Devlin et al. (2019). This model was trained on a masked language modelling objective (randomly masked tokens have to be predicted from context) and can either be used for finetuning the model parameters on any classification task or extracting representations for different contexts that can be used for further processing. We use the bert-base-german-cased model, trained by deepset.ai[9] on corpora of different domains.

We represent an adjective-noun phrase as a concatenation of the word embeddings of the two individual words. Because BERT divides some words into smaller subwords

[7] `https://finalfusion.github.io`, last accessed November 22, 2019

[8] `https://github.com/finalfusion/finalfrontier`, last accessed November 22, 2019

[9] `https://deepset.ai/german-bert`, last accessed November 22, 2019

(word pieces), we extract a single embedding for a word by computing the centroid of the corresponding word piece embeddings. We encode a sequence (context sentence and sense definition) with a bi-directional LSTM that takes the corresponding word– or word piece embeddings as input. We tuned the size of the hidden layer of the feed-forward classifier and the hidden dimension of the LSTM on the validation set. We applied dropout and early stopping[10] As BERT comes with 12 hidden layers, we experimented with different pooling methods.[11] Taking the mean of all layers worked best for all setups. Similar to the first experiment, we measure the overall performance by taking the average of the performance of each classifier on all splits. We also computed the *human performance* on a random sample of the test sets. Two students of linguistics (native speakers) were first trained on the annotation task and then asked to annotate a sample of 200 instances. The accuracy is computed based on the average number of correctly annotated instances of both annotators.

setup	embeddings	validation accuracy	test accuracy
majority baseline		0.5457	0.5087
phrase	static	0.8445	0.7138
phrase	**contextualized**	**0.8857**	**0.7415**
phrase + context	static	0.8357	0.7123
phrase + context	contextualized	0.8608	0.7337
phrase + definition	static	0.8426	0.71
phrase + definition	contextualized	0.8705	0.7390
phrase + context + definition	static	0.8414	0.7112
phrase + context + definition	contextualized	0.8645	0.7338
phrase + context + definition	*human*		*0.83*

Table 5: Averaged results on the validation and test splits for different setups with static vs. contextualized embeddings.

The results in Table 5 indicate that using contextualized representations slightly outperforms static representations for correctly classifying collocations and free phrases. Neither adding a representation of the context nor adding a representation for the prototypical sense definition helps to improve the results in general. However, it is possible that the additional information improve the classification for some adjectives but reduce it for others. This indicates that the type of information might not be robust enough to generalize for different adjectives. For example, sense definition often contain negations and antonomy, which is known to pose a problem for such models. Even though the classifier cannot reach human performance, the results indicate that word representations are a useful source for detecting collocations.

Table 7 shows the per-class accuracy for some adjectives that were either hard to classify or that got results with larger differences for different embedding types or setups. For both types of word representations similar adjectives were difficult to classify (*frischgebacken* ‘freshly baked’, *karg* ‘sparse’, *mächtig* ‘powerful’), even though‘freshly baked’ was one of the easiest adjectives for human annotators. With a corpus frequency of only 273, this adjective is rare and thus hard to model . Additional information about the word meaning in form of a context representation helps to improve the class accuracy for both embedding types (+4% for static embeddings, +15% for contextualized). Adding the sense definition improved the results for contextualized embeddings even more (+26%). Different adjectives are hard to model for one embedding type, while being easier for the other (e.g. *steinig* ‘stony’ is a difficult adjective for static embeddings, *dumm* ‘stupid’ and *schlau* ‘smart’ are difficult for contextualized embeddings.) In general, contextualized embeddings gain greater improvements through additional context and sense representations, while the accuracy for models based on static embeddings often reduces. These representations tend to rely more on the noun itself. Even though one might expect that the static embeddings would have problems to model adjectives with very different senses due to the meaning conflation deficiency, they are able to detect both, prototypical and non-prototypical meaning.

5. Further Semantic Annotation

The presented dataset is suitable not only for evaluating statistical measures and conducting machine learning experiments, but also for a more fine-grained semantic classification of adjective-noun phrases. There are two ongoing studies to find the right level of granularity for such a semantic classification.

In the first approach, we describe the relations that hold between the base and its collocate, similarly to the idea of Lexical Functions (Mel’čuk, 1996). We have examined different theoretical frameworks that can serve as a basis for semantic modelling of adjective-noun phrases (Strakatova and Hinrichs, 2019). In the current study, we rely on the information about the semantic subclasses of adjectives provided in GermaNet to define the relations between the adjective and the noun. Consider the phrase *alte Frau* ‘old lady’: here, the adjective *alt* ‘old’ expresses the value for the noun’s attribute `age`. However, in the collocation *alter Freund* ‘old friend’ the adjective ‘alt’ in most cases does not make a reference to the age of a person, but rather describes the `duration` of a friendship. Table 6 presents further examples of such attribute relations. Almost all the adjectives in the dataset are polysemous and are highly likely to express different attributes depending on the noun they modify. Two annotators are currently working on that task and the preliminary results are promising. In the process of annotation, all the adjectives and nouns are disambiguated by the annotators according to their senses in GermaNet: for each lexical unit, its ID number from GermaNet is provided. In that way, we will be able to integrate the attribute information into GermaNet.

adjective	noun	relation
alt ‘old’	*Frau* ‘woman’	`age`
alt ‘old’	*Freund* ‘friend’	`duration`
tief ‘deep’	*Wasser* ‘water’	`dimension`
tief ‘deep’	*Stimme* ‘voice’	`sound`
grob ‘coarse’	*Korn* ‘grain’	`texture`
grob ‘rough’	*Schätzung* ‘estimate’	`precision`

Table 6: Examples of attribute relations in adjective-noun pairs from the dataset.

In the second study, we add fine-grained semantic information about the nouns. For this purpose, we utilize the

[10] A table with tuned parameters can be found in Appendix B

[11] top layer, mean, sum

	static embeddings per-class Accuracy			**contextualized embeddings per-class Accuracy**		
adjective	phrase	+context	+sensedef	phrase	+context	+sensedef
steinig ‘stony’	0.08	0.12	0.15	0.85	**0.9**	0.79
karg ‘sparse’	0.21	0.21	0.15	**0.25**	0.25	0.25
frischgebacken ‘freshly baked’	0.27	0.31	0.31	0.12	0.27	**0.38**
mächtig ‘powerful’	0.3	**0.39**	0.27	0.25	0.25	0.25
windig ‘windy’	0.45	0.45	**0.47**	0.31	0.31	0.4
frech ‘bold’	**0.46**	0.4	0.34	0.39	0.39	0.42
schlau ‘smart’	**0.66**	0.63	0.54	0.11	0.11	0.11
dumm ‘stupid’	**0.7**	0.54	0.59	0.17	0.22	0.25
mild ‘mild’	0.7	0.77	0.69	**0.8**	0.74	0.69
prall ‘firm’	**0.67**	0.61	0.61	0.5	0.61	0.56
deftig ‘savoury’	0.86	0.84	0.84	0.81	0.86	**0.95**
herrlich ‘wonderful’	0.83	0.86	**0.91**	0.5	0.52	0.59
tot ‘dead’	0.52	0.49	0.6	**0.88**	0.78	0.85
teuer ‘expensive’	0.65	0.64	0.74	0.89	0.8	**0.92**

Table 7: Sample of adjectives with test set accuracy.

LexikoNet, a large lexical ontology of German nouns developed by the DWDS team (Geyken and Schrader, 2006). Each noun from the dataset is manually assigned a corresponding semantic label from LexikoNet. This information is integrated into the Wortprofil application, which allows its users to do semantic queries. The result of a query in this extended tool corresponds to a list of co-occurrences ordered by statistical salience and grouped not only by the syntactic relations, but also by the semantic classes. Thus, the adjective *tief* (‘deep’, ‘low’, ‘profound’) co-occurs with nouns from 43 semantic classes: e.g. `[feeling]` *tiefe Trauer* ‘deep sorrow’, `[social relation]` *tiefe Freundschaft* ‘close friendship’, `[colour]` *tiefes Blau* ‘deep blue’. Table 8 illustrates the query for the adjective *tief* ‘deep’ and the semantic class `[feeling]`: the top-10 co-occurring nouns are shown sorted by the logDice value. All the nouns that have the label `[feeling]` are highlighted in bold (*Trauer* ‘sorrow’, *Mißtrauen* ‘distrust’). This semantically enriched version of the Wortprofil is currently used internally to support the lexicographic work at the DWDS.

Grundwort	**Lemma**	**logDice**	**Freq**
Krise ‘crisis’	Krise ‘crisis’	9.59	3929
Einblicke ‘insights’	Einblick ‘insight’	8.91	1783
Loch ‘hole’	Loch ‘hole’	8.73	1767
Einschnitte ‘cuts’	Einschnitt ‘cut’	8.67	1524
Sinn ‘sense’	Sinn ‘sense’	8.54	2157
Graben ‘grave’	Graben ‘grave’	8.49	1255
Trauer ‘grief’	**Trauer ‘grief’**	**8.24**	**1050**
Spuren ‘traces’	Spur ‘trace’	8.23	1240
Rezession ‘recession’	Rezession ‘recession’	8.21	1070
Misstrauen ‘mistrust’	**Mißtrauen‘mistrust’**	**8.18**	**1039**

Table 8: Extended version of Wortprofil. The top-10 results for the query “*tief* ‘deep’ + `[feeling]`". The nouns from the semantic class `[feeling]` are in bold.

6. Conclusion and Future Work

In this paper we have presented the GerCo dataset of adjective-noun collocations for German[12] that provides a broad coverage of different semantic classes of adjectives as defined in GermaNet. The proposed dataset has been annotated by experts. The collection contains 4,732 positive and negative instances of collocations, both of which exhibit a strong statistical association between their constituents. The collection also includes context sentences for a subset of 3,888 instances that can be used to examine collocations in context or conduct machine learning experiments with deep-contextualized representations.

A number of such machine learning experiments has been conducted in Section 4. The experiments revealed that using word embeddings as features outperforms approaches based on association measures. Feature representations that capture more semantic information about the candidates are more powerful and applicable for differentiating restricted word combinations from free phrases if both exhibit a strong statistic association. Contextualized embeddings or some representation of the context help when the static word representation is dominated by one sense. If the context is not sufficient to provide information about the meaning of the adjective, a representation of a sense definition can improve the results. There is no optimal setup for all adjectives, the quality of the context sentences and the sense definitions has an impact on the classification performance, but it would also be interesting to investigate whether the ambiguity and frequency of the noun play a role in the success of the models. Neither static, nor contextualized embeddings can achieve human performance, and although these features are much stronger than association scores, a simple non-linear classifier is not able to generalize perfectly on adjectives it has not encountered in training.

Further research questions can be investigated with the presented data set in the future. The representation of the adjective-noun phrase can be constructed with the aim of capturing more interaction between the individual parts of the word combinations. This can be achieved by using a composition model (Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010; Dima et al., 2019), that takes the individual word vectors as input and combines them in a way such that the combined representation is more suitable for a specific task. Additional semantic information (sense gloss or semantic class from GermaNet) may contain information that is not implicit in the representations of context and phrases and can further provide information that is necessary to solve such tasks.

7. Acknowledgements

We would like to thank our student assistants Eva Huber, Alina Leippert, and Mareile Winkler for their help with the annotations. We would also like to acknowledge Corina Dima for valuable feedback and discussions about the experiments. We are also very grateful to the anonymous reviewers for their insightful and helpful comments. The financial support of the research reported here is provided by the grant Modellierung lexikalisch-semantischer Beziehungen von Kollokationen awarded by the Deutsche Forschungsgemeinschaft (DFG).

Appendix A: Lexical AMs

log.likelihood, Dice, Jaccard, MI, MI2, MI3, MS, Poisson.Stirling, average.MI, chi.squared, chi.squared.corr, frequency, gmean, local.MI, odds.ratio, odds.ratio.disc, random, relative.risk, simple.ll, t.score, z.score[13]

[12] The dataset is available at `http://hdl.handle.net/11022/0000-0007-DABA-2`, last accessed November 22, 2019

[13] A detailed explanation of the measurements can be found here: `http://www.collocations.de/UCS/`

Appendix B: Hyperparameters non-prototypical classifier

Drop out rates between 0.0 and 0.8 in 0.2 increments were tuned on the validation set for all non-linear classifier. Hidden Layer sizes were for all non-linear classifier and the LSTM were tuned within the following range: (50, 100, 200, 300, 400, 500). The hidden layer size for the classifier trained on association scores was tuned within the following range: (1, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 25, 50, 100).

setup	embeddings	hidden layer	dropout	sentence hidden dim LSTM
phrase	combination of AMs	6	-	-
phrase	static	400	0.2	-
phrase	contextualized	500	0.8	-
phrase + context	static	400	0.2	200
phrase + context	contextualized	500	0.8	50
phrase + sense def	static	400	0.2	50
phrase + sense def	contextualized	500	0.8	500
phrase + context + sense def	static	400	0.2	100
phrase + context + sense def	contextualized	500	0.8	200

Table 9: Hyperarameters for all models tuned on the validation set.

8. References

Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 792–797, Paris, France. European Language Resources Association (ELRA).

Baroni, M. and Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1183–1193, Cambridge, MA, October. Association for Computational Linguistics.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer Science+ Business Media.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *CoRR*, abs/1607.04606.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, March.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Dima, C., de Kok, D., Witte, N., and Hinrichs, E. (2019). No word is an island—a transformation weighting model for semantic composition. *Transactions of the Association for Computational Linguistics*, 7:437–451.

DWDS. (2019). DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart, hrsg. v. d. Berlin-Brandenburgischen Akademie der Wissenschaften, <https://www.dwds.de/>.

Espinosa-Anke, L., Schockaert, S., and Wanner, L. (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, July. Association for Computational Linguistics.

Evert, S., Uhrig, P., Bartsch, S., and Proisl, T. (2017). E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification. In Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference, Leiden, The Netherlands, pages 531–549.

Evert, S. (2004). The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Evert, S. (2008). A lexicographic evaluation of German adjective-noun collocations. In Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008), Marrakech, Morocco.

Garcia, M., García Salido, M., and Alonso-Ramos, M. (2019). A comparison of statistical association measures for identifying dependency-based collocations in various languages. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 49–59, Florence, Italy, August. Association for Computational Linguistics.

Geyken, A. and Schrader, N. (2006). LexikoNet - a lexical database based on type and role hierarchies. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, May. European Language Resources Association (ELRA).

Geyken, A., Didakowski, J., and Siebert, A. (2009). Generation of Word Profiles for Large German Corpora. *Corpus Analysis and Variation in Linguistics*, 1:141–157.

Hamp, B. and Feldweg, H. (1997). GermaNet - a lexical-semantic net for German. In Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid.

Henrich, V. and Hinrichs, E. (2010). GernEdiT - the GermaNet Editing Tool. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, May. European Languages Resources Association (ELRA).

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of the Tenth Machine Translation Summit (MT Summit X), pages 79–

UCS-Perl-html/UCS/AM.html, last accessed November 22, 2019

86, Phuket, Thailand.

Krenn, B. (2000). The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations. Ph.D. thesis, German Research Center for Artificial Intelligence and Saarland University, Saarbrücken, Germany.

Ling, W., Dyer, C., Black, A. W., and Trancoso, I. (2015). Two/too simple adaptations of Word2Vec for syntax problems. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1299–1304, Denver, Colorado, May–June. Association for Computational Linguistics.

McIntosh, C., Francis, B., and Poole, R. (2009). Oxford Collocations Dictionary for students of English. Oxford University Press.

Mel'čuk, I. (1996). Lexical functions: a tool for the description of lexical relations in a lexicon. *Lexical functions in lexicography and natural language processing*, 31:37–102.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer, 3(1). In *Yearbook of Phraseology*. De Gruyter, pp. 31–56.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, et al., editors, Advances in Neural Information Processing Systems 26, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates, Inc.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Pecina, P. and Schlesinger, P. (2006). Combining association measures for collocation extraction. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 651–658. Association for Computational Linguistics.

Pecina, P. (2008a). Lexical Association Measures: Collocation Extraction. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic.

Pecina, P. (2008b). A machine learning approach to multiword expression extraction. In Proceedings of the LREC 2008 Workshop Towards a Shared Task for Multiword Expressions, pages 54–57, Marrakech, Morocco. European Language Resources Association.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237. Association for Computational Linguistics.

Polguere, A. (2000). Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In Egbert Lehmann Christian Rohrer Ulrich Heid, Stefan Evert, editor, Proceedings of the 9th EURALEX International Congress, pages 517–527, Stuttgart, Germany. Institut für Maschinelle Sprachverarbeitung.

Quasthoff, U. (2011). Wörterbuch der Kollokationen im Deutschen. De Gruyter, Berlin, Boston.

Michael Rundell, editor. (2010). Macmillan Collocations Dictionary for Learners of English. Macmillan Education.

Rychly, P. (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.

Schabus, D. and Skowron, M. (2018). Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC), pages 1602–1605, Miyazaki, Japan, May.

Schabus, D., Skowron, M., and Trapp, M. (2017). One million posts: A data set of german online discussions. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pages 1241–1244, Tokyo, Japan, August.

Schäfer, R. and Bildhauer, F. (2012). Building Large Corpora from the Web Using a New Efficient Tool Chain. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 486–493, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1):143–177.

Strakatova, Y. and Hinrichs, E. (2019). Semantic modelling of adjective-noun collocations using FrameNet. In Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 104–113, Florence, Italy, August. Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pages 2214–2218, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Vincze, O., Mosqueira, E., and Alonso Ramos, M. (2011). An online collocation dictionary of Spanish. In Proceedings of the 5th International Conference on Meaning-Text Theory. Barcelona, pages 275–286.