Choosing Suitable Text Corpora for Identifying Collocations – A Case Study of a Large Reference Dictionary of Contemporary German

Luise Köhler[1], Gregor Middell[1], Alexander Geyken[1]

[1]Berlin-Brandenburg Academy of Science and Humanities, Jägerstraße 22/23, 10117 Berlin, Germany

E-mail: luise.koehler@bbaw.de, gregor.middell@bbaw.de, geyken@bbaw.de

Abstract

This paper investigates the impact of corpora for extracting collocation candidates from large text corpora. We compare a variety of corpora, including best-selling fiction and non-fiction literature, a contemporary German reference corpus, the German Wikipedia, a curated web page monitor corpus, and a large newspaper corpus. All corpora undergo processing with an NLP pipeline for morphological and syntactic annotations. Collocation candidates are then extracted using dependency parse tree patterns.

For our evaluation, we utilized three gold standard collocation datasets for contemporary German with a total amount of appr. 200,000 collocations. Our findings confirm that well-curated, medium-sized corpora, diverse in text types, offer superior coverage compared to opportunistically collected corpora of equivalent size. We could also confirm that well curated newspaper texts perform better than pure web corpora of the same size. Additionally, we observed significant variation in coverage among the examined corpora, depending on specific syntactic relations.

Future work will focus on training models to classify collocation candidates, leveraging the advancements in LLMs. Further research into larger, more diverse datasets for both training and evaluation could improve collocation ranking and candidate discovery.

Keywords: collocations; corpora; evaluation

1. Introduction

Collocations have long been a focal point within linguistic and lexicographical research (e.g. Firth, 1957; Hausmann, 1984; Sinclair, 1991). The emergence of evidence-based lexicography, coupled with the increasing availability of extensive text corpora, spurred a significant research interest in computational approaches (e.g. Church & Hanks, 1990; Kilgarriff et al., 2004). These methods aimed to extract typical co-occurrences from such corpora and subsequently assist lexicographers in the complex task of identifying genuine collocations among them.

Numerous studies have explored the statistical characteristics of collocations across a variety of languages, different collocation types, and with the application of multiple gold standards and diverse corpora. This extensive research often involved the evaluation of various association measures to quantify the strength of co-occurrence. Despite the undisputed heuristic value of statistical methods for the task at hand, the overall results of such studies do not provide clear conclusions for practical lexicography, including

the workflow of dictionary compilation. More recently, machine learning methods have been applied to detect and classify collocations. Taking its point of departure from these findings and recent advances in applying machine learning methods to detect and classify collocations, this study examines a research question that prior research has addressed only in passing. This question becomes more critical as the focus shifts from the precision of association measures to the recall required when constructing representative datasets for training classifiers: Which type of corpora should form the basis of lexicographic work for compiling dictionaries?

To provide answers to this question, we proceed as follows. Section 2 briefly reviews related work on collocation extraction. Section 3 details the NLP method employed for processing corpora and extracting collocation candidates. Section 4 describes the data: three German collocation datasets totaling approximately 200,000 collocations are introduced in 4.1, and seven distinct German corpora, varying in textual stratification and size, are used to evaluate the coverage of these datasets (section 4.2). Section 5 presents and discusses key findings. Finally, Section 6 offers a conclusion and outlines directions for future research.

2. Related Work

Computational lexicography has a long tradition of investigating the impact of corpus characteristics (type and size), association measures, and linguistic preprocessing on retrieving collocation candidates. Such investigations are informed not only by linguistic aspects like language or collocation type but also by technical and infrastructural constraints like the availability of representative evaluation datasets, large corpora, or the state-of-the-art in processing and annotating such corpora.

Bartsch & Evert (2014) use a subset of an English collocation dictionary comprising around 3,000 collocations. Collocation candidates are obtained from syntactic relations, the co-occurrence of collocates within three span sizes, and the sentence context. They observe that larger corpora and broader co-occurrence contexts improve coverage while the syntactic dependency criterion decreases coverage. Considering precision, they find that corpus size does not play the same role as textual balance and quality. In a later study, Evert et al. (2017) investigate the statistical properties of collocations by applying different association measures. They note the focus of earlier studies on a singular syntactic relation for collocation retrieval. In their research, they use subsets of two English collocation dictionaries as a gold standard that cover different syntactic relations and compare different association measures on a range of corpora. Their results show that large (web) corpora perform better than smaller corpora. For the comparison of similar-sized corpora, they conclude that a balanced and curated corpus outperforms an unbalanced one. Concerning association measures, they note that the "optimal choice of an [association measure] depends strongly on the particular gold standard used". Uhrig, Evert & Proisl (2018) compare the impact of different parsers on collocation candidate selection. Besides parsing schemes, they study different association measures for six collocation types. They find that the performance of parsers and the precision of association measures vary considerably between relations. They note that collocation candidate rankings produce better results on a balanced corpus than on a large web corpus, e.g., due to boilerplate text. In contrast, total coverage improves with large corpora.

Recently, the use of language models and word embeddings for collocation retrieval has been studied in various experimental setups (cf. Strakatova et al., 2020; Falk et al.,

2021; Espinosa-Anke, Codina-Filbá & Wanner, 2021; Ljubešić, Logar & Kosem, 2021). Espinosa-Anke, Codina-Filbá & Wanner (2021) apply a transformer-based language model to collocation prediction and categorization in English. They use a masking setup to study the predicted collocates and, in a second experiment, compare the classification of collocations according to lexical function categories. Ljubešić, Logar & Kosem (2021) compare the performance of statistical and machine-learning methods in collocation ranking. As the attention of their research is on use in lexicography, they focus on the logDice score, which is "lexicographically highly popular" (Ljubešić, Logar & Kosem, 2021) as a statistical measure. They use *fastText* embeddings and a support-vector machine regressor for the machine-learning approach. The ML collocate ranking outperforms the logDice ranking on both tested datasets.

3. Method

The approach to corpus processing that we apply in this study aligns with the methodology of the *Digitales Wörterbuch der deutschen Sprache* ('Digital Dictionary of the German Language', *DWDS*) (**?**). Each corpus undergoes initial processing by an NLP pipeline, which enriches the texts with morphological and syntactic annotations. This pipeline (**?**) is designed for *DWDS*' lexicographic applications. It offers two versions: a BERT-based version and a version utilizing static word vectors, with the latter prioritizing processing speed on CPU architectures over accuracy. Table 1 provides a detailed breakdown of the expected accuracy for both versions, specifically for the linguistic features and syntactic dependency relations that are subsequently used to extract collocation candidates.

Category	**Word Vectors**	**BERT**
Part-Of-Speech	98.05%	98.32%
Morphological Feature – Case	83.07%	85.32%
Dependency Relations (LAS)	93.01%	95.26%
Dependency Relation – advmod	89.90%	92.18%
Dependency Relation – amod	99.33%	99.49%
Dependency Relation – case	99.09%	99.49%
Dependency Relation – cop	89.06%	93.55%
Dependency Relation – conj	82.83%	88.79%
Dependency Relation – nmod	89.20%	92.24%
Dependency Relation – nsubj	93.27%	96.84%
Dependency Relation – nsubj:pass	94.82%	97.44%
Dependency Relation – obl	88.86%	92.84%
Dependency Relation – obl:arg	86.35%	88.23%
Dependency Relation – obj	89.25%	94.68%

Table 1: Accuracy of a vector-based and BERT-based pipeline on different annotation tasks

For this study, we applied the word-vector-based version to speed up the annotation process, with a minor but acceptable decline in accuracy as the survey compares the coverage provided by different corpora, not the performance of the underlying annotation process.
In addition to the grammatical category of words and their syntactic relations[1], measuring collocation coverage also depends on the accurate lemmatization of texts. A rule-based German morphology, together with a comprehensive lexicon derived from the *DWDS* dictionary, is used for this task. Named *DWDSmor* for its dual heritage – the *DWDS* as its lexicon source and *SMOR* (Schmid, Fitschen & Heid, 2004) as its source for the finite-state morphology – it covers 98-100% of German lexemes, depending on the grammatical category of a word.[2]
Once all corpora under evaluation are tagged and lemmatized, collocation candidates are extracted by filtering the dependency parse trees for patterns. Grouped by corpus, lemma, and collocation relation type, co-occurrences are counted and stored in a database for lookup during evaluation. Types of collocation relations between lemmata thus identified are noun-noun relations (with and without prepositions), attribute-noun/verb relations, noun-verb relations, and coordinating structures (see section 4.2 for a more detailed description of the investigated relations).

4. Data

4.1 Collocation Datasets

For evaluation, we use a gold standard comprised of three collocation resources for contemporary German: a dataset of German noun-adjective collocations (Strakatova et al., 2020), a dictionary of German collocations (**?**), and all the collocations compiled for and integrated into the *DWDS* dictionary.
The smallest collocation resource used is the *GerCo* dataset (Strakatova et al., 2020). The authors use the co-occurrence data from the *DWDS-Wortprofil* (Didakowski & Geyken, 2014) for 48 adjectives, which cover the 16 semantic classes of *GermaNet* (Hamp & Feldweg, 1997). The dataset contains 4,732 adjective-noun phrases; around half of these were classified by annotators as positive instances of collocations, the rest consist of free phrases. In this study, we use a subset of *GerCo* representing proper collocations.
The *Wörterbuch der Kollokationen im Deutschen* (**?**), hereafter referred to as *Quasthoff*, is a dictionary of collocations for 3,200 articles (collocation bases). It comprises approximately 192,000 collocations that were extracted based on text corpora of the *Deutscher Wortschatz | Leipzig Corpora Collection* (Goldhahn, Eckart, Quasthoff et al., 2012). The most frequent words in the corpus were selected as bases, and left and right neighbors were considered collocate candidates. Lists of candidates were filtered for significance and manually edited. The base lemmas are predominantly nouns (2,346 nouns, 617 verbs, 290 adjectives). Collocates can be verbs or adjectives for noun bases. For verb and adjective bases, only adverb collocates are available. The base lemma can be in a subject or object (direct or indirect) relation for noun-verb collocations.
The *DWDS* dictionary, a comprehensive documentary dictionary, offers extensive lexicographic information for each headword. This includes grammatical form, spelling, word

[1] We use a dependency parser (**?**) trained on Universal Dependencies.
[2] As the sole exception, named entities are not well-covered by *DWDSmor* because they are not included in the underlying dictionary/lexicon. Fortunately named entities are also not part of the evaluated collocation datasets, so this constraint of *DWDSmor* has no impact in this context.

formation, and meaning, encompassing word senses, definitions, corpus examples, and collocations. Collocations are explicitly annotated within the meaning component, organized by sense number and syntactic categories (Lemnitzer et al., 2025). At the time of dataset generation (late February 2025), around 19,000 *DWDS* dictionary articles featured collocations. Due to the *DWDS*'s revision strategy, which prioritizes new entries, the collocation database is biased towards more recently included vocabulary. The *DWDS-Wortprofil*, a co-occurrence analysis tool (Didakowski & Geyken, 2014;**?**), aids in the selection of collocations. Further details on the lexicographic selection process of collocations can be found in Lemnitzer et al. (2025). For easier readability for human users, the collocations often consist of inflected forms or small phrases to illustrate their typical usage in natural language. We used regular expression rules to extract the collocates and *DWDSmor* (with the Python package *simplemma* (**?**) as a fallback) for their lemmatization to bring the collocations into a processable format. We filtered out duplicates stemming from different word senses or different bases, as some collocations can be found in the entries of the head and the dependent. The resulting dataset, hereafter referred to as *DWDScoll*, consists of approximately 177,000 collocations.

The syntactic relations covered by our datasets are as follows: adjective attribute (*ATTR*), adverb (*ADV*), attributive genitive (*GMOD*), coordination (*KON*), direct/accusative object (*OBJ*), indirect (dative/genitive) object (*OBJO*), prepositional phrases (*PP*), predicative (*PRED*), active subject (*SUBJ*), and passive subject (*SUBJP*). Table 2 gives an overview of the used datasets, detailing the number of collocations by type and dataset and the coverage of those collocation sets when measured against all corpora (see next section) consulted in this paper.
Since the noun-verb collocations in the *Quasthoff* dataset are not labeled in as much detail as in *DWDScoll*, we decided to group them all into one separate category (*SUBJ/OBJ*). As a result, and as can be seen in Table 2, some relations are only available in the *DWDScoll* dataset. The *GerCo* dataset focuses only on one commonly evaluated collocation type, attribute-noun collocations, while the other two provide more variety. Some relations, e.g., noun-noun (*GMOD*), noun-noun with preposition, and coordinating structures, are unavailable in the *Quasthoff* dataset.

When looking at all datasets and corpora, the total coverage of all collocations is about 92%. The *GerCo* dataset exhibits the best coverage (97.5%). This is not surprising since *GerCo* is much smaller than the other two datasets and focuses only on one syntactic relation. Of the two larger datasets in this study, the *Quasthoff* dataset shows significantly higher coverage, probably owing to the mode of its construction: Collocation candidates had to pass a rather strict threshold of statistical significance before being even considered for manual approval and incorporation in the dataset. Also, the base lemmas in the *Quasthoff* dataset were selected for their frequency and thus have a higher probability of occurring in the tested corpora.
The *DWDScoll* dataset, on the other hand, has a much larger lexical coverage. If all collocation bases and collocates are collected in a set, this results in 41,000 unique lemmas. The *Quasthoff* dataset only contains 16,000 unique lemmas, while the total number of collocations in both datasets is comparable. The two datasets share 10,880 lemmas.
Even though some of the lemmas in the *DWDScoll* dataset are not shared by the frequency-based *Quasthoff* dataset, this dataset reflects the needs of lexicographic work: a lexicogra-

Relation	GerCo	%	Quasthoff	%	DWDScoll	%	∑	%
∑	1,870	97.49	192,029	96.84	164,920	87.50	358,819	92.55
ADV	0	0.00	26,615	86.50	4,433	95.17	31,048	87.74
ATTR	1,870	97.49	105,363	98.51	56,411	95.83	163,644	97.57
GMOD	0	0.00	0	0.00	24,581	44.16	24,581	44.16
KON	0	0.00	0	0.00	10,471	91.94	10,471	91.94
OBJ	0	0.00	0	0.00	30,659	96.72	30,659	96.72
OBJO	0	0.00	0	0.00	1,440	58.06	1,440	58.06
PP	0	0.00	0	0.00	35,246	95.03	35,246	95.03
PRED	0	0.00	0	0.00	1,043	92.33	1,043	92.33
SUBJ/OBJ	0	0.00	60,051	98.50	0	0.00	60,051	98.50
SUBJP	0	0.00	0	0.00	636	94.97	636	94.97

Table 2: Collocation coverage by all corpora, per grammatical relation and dataset

pher might be interested in determining which corpus is more diverse to also be able to find collocations and examples for less frequent lexemes.

4.2 Corpora

To test the coverage of our collocation datasets, we utilized a range of corpora used in *DWDS*'s daily work. We chose a mix of newspapers, web-based text, and contemporary literature from the *DWDS* corpus collection.

Corpus	Tokens	Sentences	Co-occurrences
bestseller	87.92	6	15.88
DWDS-base	315.04	14.86	61.55
wikipedia	1,305.68	62.08	234.7
webmonitor	5,022.86	309.95	1,031.57
DWDS-newspaper	5,741.78	348.92	1,135.86
DWDS-meta	6,056.82	363.79	1,197.41
web-meta	6,328.54	372.03	1,266.27

Table 3: Summary of corpus statistics (counts in millions)

The *bestseller* corpus is a collection of over 600 contemporary fiction and nonfiction books on the canonical bestseller list in Germany between 2018 and 2022. It contains texts written in German and books translated into German, has a token count of 87.92 million,

and is thus the smallest of the tested corpora.
The *DWDS-base* corpus (Geyken, 2007) is a large corpus of 20th-century German language. It comprises four text types: fiction, newspaper, science, and functional literature, and it consists of 315.04 million tokens.
For web-based texts, we included two corpora: the *wikipedia* corpus, which is based on a dump of the German Wikipedia (July 1st, 2024) and contains 1.3 billion tokens, and the *webmonitor* corpus with 5 billion tokens. The latter is a corpus of several hundred popular German-language web pages with diverse topics. They include websites of companies or associations as well as the online texts of newspapers. Its sources were curated in 2021. The *webmonitor* corpus is updated daily using the Python library *trafilatura* (Barbaresi, 2021) that includes boilerplate cleaning of the scraped websites.
The *DWDS-newspaper* corpus is a collection of major regional and nationwide daily and weekly newspapers. It covers the period from 1946 until today, with a focus on recent publications (since 2005). The corpus is updated monthly and contains 5.7 billion tokens.
We formed two meta corpora for analysis purposes, the *web-meta* and the *DWDS-meta* corpus. The *web-meta* corpus combines the two internet-based corpora (*webmonitor* and *wikipedia*). The *DWDS-meta* corpus consists of the corpora used for the co-occurrence analysis tool in the *DWDS* dictionary context (combining the corpora *DWDS-base* and *DWDS-newspaper*). Table 3 shows the counts of tokens, sentences, and co-occurrences for all corpora.

5. Results and Discussion

Table 4 illustrates the percentage of total coverage of the collocation datasets across various corpora.
It is worth noting that even the smallest corpora, *bestseller* and *DWDS-base*, demonstrate significant coverage for *GerCo*. The *bestseller* corpus covers 70.6% of *GerCo* collocations, while *DWDS-base* achieves 87.3%. However, for the *Quasthoff* and *DWDScoll* datasets, the *bestseller* corpus performs considerably worse, covering only 51.38% and 22.86%, respectively. In contrast, the *DWDS-base* corpus achieves 76.84% for *Quasthoff* and 45.67% for *DWDScoll*. It is noteworthy that the *DWDS-base* corpus's coverage for *GerCo* and *Quasthoff* is only marginally inferior to the *wikipedia* corpus, despite the latter being four times larger, supporting the conclusions of Evert et al. (2017).

Significant findings emerged from the three largest corpora: *web-meta*, *DWDS-newspaper*, and *DWDS-meta*. The *GerCo* dataset exhibited very high coverage across all three, at approximately 97%. This coverage slightly decreased for the *Quasthoff* dataset, with *web-meta* at 94.9%, and *DWDS-newspaper* and *DWDS-meta* at 96.6%. A more substantial difference was observed for the *DWDScoll* dataset, where only 80.9% of collocations were found in the *web-meta* corpus. Coverage improved slightly to 85.5-86% with the *DWDS-newspaper* and *DWDS-meta* corpora.
For the *Quasthoff* dataset, increasing the amount of data yields only a slight improvement. The larger subcorpora already provide extensive coverage. For instance, the web-meta corpus, with 94.24% of *Quasthoff* collocations found in the *webmonitor*, sees only marginal improvement from including the *wikipedia* corpus. Similarly, for the *DWDS-newspaper* and *DWDS-base* corpora, adding the *DWDS-base* corpus to the newspaper corpus (as in the

Corpus	**GerCo**	**Quasthoff**	**DWD-Scoll**	$\sum$
bestseller	70.64%	51.38%	22.86%	38.37%
DWDS-base	87.27%	76.84%	45.67%	62.57%
wikipedia	92.89%	83.09%	61.62%	73.27%
webmonitor	96.36%	94.24%	77.33%	86.47%
DWDS-newspaper	97.27%	96.56%	85.49%	91.47%
DWDS-meta	97.27%	96.61%	86.15%	91.80%
web-meta	97.01%	94.94%	80.93%	88.51%

Table 4: Coverage of datasets by corpora, the last column aggregates the three datasets

DWDS-meta corpus) does not enhance collocation coverage. However, for the *DWDScoll* dataset, these meta corpora show a slight improvement. Web-based corpora, for example, demonstrate a 3.5% better coverage when using a larger corpus (*webmonitor* and *wikipedia*). In contrast, the improvement for the *DWDS-newspaper* and *DWDS-meta* corpora is only marginal.

Table 5 demonstrates the varying corpus coverage across different syntactic relations. A significant disparity in coverage is observable between the corpora. For instance, the *ATTR* phrase relation, which constitutes approximately half of the collocations in the gold datasets, shows coverage ranging from 40% (*bestseller*) to 97.05% (*DWDS-meta*). Similarly, for the verbal relation *OBJ*, *bestseller* offers poor coverage (31.15%), while *DWDS-meta* achieves the best coverage at 95.79%. The lowest coverage is found for *GMOD* and *OBJO*, where even the best corpus (*DWDS-meta*) only reaches 42.98% and 53.26% respectively, and the smaller bestseller corpus barely exceeds 6% for both relations.

In summary, evaluating against datasets containing collocations of multiple grammatical relations is essential, and small datasets are not sufficient to answer the question which corpora are best suited for collocation retrieval. For small datasets or datasets containing frequency-based collocations (such as our *Quasthoff* dataset), the corpus base does not need to be very large to reach sufficient coverage. However, larger corpora with diverse sources are necessary if the dataset is more varied, such as the *DWDScoll* dataset. Here, we find that newspaper-based corpora are better suited than web-based corpora.

6. Conclusion and Future Work

In our work, we have investigated the impact of corpus selection for extracting collocation candidates from large text corpora. Our findings on a large dataset of approximately 200,000 collocations from three datasets confirm previous studies: well-curated, medium-sized corpora, diverse in text types, offer superior coverage compared to opportunistically collected corpora of equivalent size. We could also confirm that well-curated newspaper texts outperform pure web corpora of the same size. Moreover, the findings indicate that large corpora are useful for practical lexicography when compiling large general language dictionaries. Unlike learner dictionaries, these should encompass collocations for less common words and various grammatical relationships.

Relation	best-seller	DWDS-base	wiki-pedia	web-monitor	DWDS-newsp.	web-meta	DWDS-meta
$\sum$	38.37%	62.57%	73.27%	86.47%	91.47%	88.51%	91.80%
ADV	48.84%	68.84%	72.91%	84.85%	87.06%	85.64%	87.20%
ATTR	40.17%	68.13%	78.42%	91.80%	96.75%	93.84%	97.05%
GMOD	6.16%	19.55%	28.14%	34.77%	42.38%	37.95%	42.98%
KON	18.08%	38.69%	60.99%	75.12%	87.98%	81.88%	89.19%
OBJ	31.15%	54.83%	70.57%	90.27%	95.30%	92.20%	95.79%
OBJO	6.32%	19.93%	26.60%	38.89%	51.81%	43.26%	53.26%
PP	22.81%	45.24%	65.30%	84.34%	93.08%	87.76%	93.56%
PRED	17.64%	41.71%	58.01%	84.37%	90.60%	86.86%	90.99%
SUBJ/OBJ	58.98%	81.68%	87.55%	96.49%	98.15%	97.00%	98.19%
SUBJP	15.09%	41.51%	64.94%	83.02%	91.35%	88.05%	92.14%

Table 5: Coverage of gold collocations by corpora, grouped by relations

A worthwhile extension of this research is examining the degree to which our collocation dataset's coverage diminishes when utilizing less meticulously curated web corpora instead of the *webmonitor* (employed in the present study). Such a study would also facilitate a comparison of German findings with existing English results (cf. Evert et al., 2017; Uhrig, Evert & Proisl, 2018), thereby providing valuable cross-language insights.
Given the current interest in large language models and the proven utility of available corpora for generating collocation candidates, future work should also focus on training models to classify these candidates. As Falk et al. (2021) demonstrated for German adjective-noun phrases, this is achievable even with relatively small training datasets.
Moreover, Ljubešić, Logar & Kosem (2021) show that such classifiers can outperform frequency-based rankings even using more traditional machine learning methods and static vector representations of classified phrases. Exploring larger, more diverse datasets for training and evaluation is expected to yield interesting results for collocation ranking, similar to its impact on collocation candidate discovery.

7. References

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 122–131.

Bartsch, S. & Evert, S. (2014). Towards a Firthian Notion of Collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2(1), pp. 48–61.

Church, K. & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), pp. 22–29.

Didakowski, J. & Geyken, A. (2014). From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions. *OPAL–Online publizierte Arbeiten zur Linguistik*, 2(2014), pp. 39–47.

Espinosa-Anke, L., Codina-Filbá, J. & Wanner, L. (2021). Evaluating Language Models for the Retrieval and Categorization of Lexical Collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. pp. 1406–1417.

Evert, S., Uhrig, P., Bartsch, S. & 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 the eLex 2017 Conference*. pp. 531–549.

Falk, N., Strakatova, Y., Huber, E. & Hinrichs, E. (2021). Automatic Classification of Attributes in German Adjective-Noun Phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics (ACL), pp. 239–249.

Firth, J.R. (1957). *Papers in Linguistics, 1934-1951*. Oxford University Press.

Garcia, M., García Salido, M. & 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)*. Association for Computational Linguistics (ACL), pp. 49–59.

Geyken, A. (2007). The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. In C. Fellbaum (ed.) *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. Continuum Press London, pp. 23–42.

Goldhahn, D., Eckart, T., Quasthoff, U. et al. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. pp. 31–43.

Hamp, B. & Feldweg, H. (1997). GermaNet - A Lexical-Semantic Net for German. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources*.

Hausmann, F.J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31(4), pp. 395–406.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL*, volume 2013. pp. 125–127.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Lemnitzer, L., Ermakova, M., Palmes, L., Roll, B., Siebel, K. & Geyken, A. (2025). Kollokationen im DWDS-Wörterbuch und ihr Mehrwert für DaF/DaZ. *Deutsch als Fremdsprache*, 62(2), pp. 67–80.

Ljubešić, N., Logar, N. & Kosem, I. (2021). Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 9(2), pp. 41–70.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *LREC 2004: Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 1263–1266.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Strakatova, Y. (2024). German Adjective-Noun Co-occurrences with Attributes [Dataset]. University of Tübingen. URL https://doi.org/10.57754/FDAT.76krc-egt63.

Strakatova, Y., Falk, N., Fuhrmann, I., Hinrichs, E. & Rossmann, D. (2020). All that Glitters is not Gold: A Gold Standard of Adjective-Noun Collocations for German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, pp. 4368–4378.

Uhrig, P., Evert, S. & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. *Lexical Collocation Analysis: Advances and Applications*, pp. 111–140.

Dictionaries

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 122–131.

Bartsch, S. & Evert, S. (2014). Towards a Firthian Notion of Collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2(1), pp. 48–61.

Church, K. & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), pp. 22–29.

Didakowski, J. & Geyken, A. (2014). From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions. *OPAL–Online publizierte Arbeiten zur Linguistik*, 2(2014), pp. 39–47.

Espinosa-Anke, L., Codina-Filbá, J. & Wanner, L. (2021). Evaluating Language Models for the Retrieval and Categorization of Lexical Collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. pp. 1406–1417.

Evert, S., Uhrig, P., Bartsch, S. & 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 the eLex 2017 Conference*. pp. 531–549.

Falk, N., Strakatova, Y., Huber, E. & Hinrichs, E. (2021). Automatic Classification of Attributes in German Adjective-Noun Phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics (ACL), pp. 239–249.

Firth, J.R. (1957). *Papers in Linguistics, 1934-1951*. Oxford University Press.

Garcia, M., García Salido, M. & 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)*. Association for Computational Linguistics (ACL), pp. 49–59.

Geyken, A. (2007). The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. In C. Fellbaum (ed.) *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. Continuum Press London, pp. 23–42.

Goldhahn, D., Eckart, T., Quasthoff, U. et al. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. pp. 31–43.

Hamp, B. & Feldweg, H. (1997). GermaNet - A Lexical-Semantic Net for German. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources*.

Hausmann, F.J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31(4), pp. 395–406.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL*, volume 2013. pp. 125–127.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Lemnitzer, L., Ermakova, M., Palmes, L., Roll, B., Siebel, K. & Geyken, A. (2025). Kollokationen im DWDS-Wörterbuch und ihr Mehrwert für DaF/DaZ. *Deutsch als Fremdsprache*, 62(2), pp. 67–80.

Ljubešić, N., Logar, N. & Kosem, I. (2021). Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 9(2), pp. 41–70.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *LREC 2004: Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 1263–1266.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Strakatova, Y. (2024). German Adjective-Noun Co-occurrences with Attributes [Dataset]. University of Tübingen. URL https://doi.org/10.57754/FDAT.76krc-egt63.

Strakatova, Y., Falk, N., Fuhrmann, I., Hinrichs, E. & Rossmann, D. (2020). All that Glitters is not Gold: A Gold Standard of Adjective-Noun Collocations for German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, pp. 4368–4378.

Uhrig, P., Evert, S. & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. *Lexical Collocation Analysis: Advances and Applications*, pp. 111–140.

Software

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 122–131.

Bartsch, S. & Evert, S. (2014). Towards a Firthian Notion of Collocation. *Vernetzungsstrategien, Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern*, 2(1), pp. 48–61.

Church, K. & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1), pp. 22–29.

Didakowski, J. & Geyken, A. (2014). From DWDS Corpora to a German Word Profile – Methodological Problems and Solutions. *OPAL–Online publizierte Arbeiten zur Linguistik*, 2(2014), pp. 39–47.

Espinosa-Anke, L., Codina-Filbá, J. & Wanner, L. (2021). Evaluating Language Models for the Retrieval and Categorization of Lexical Collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. pp. 1406–1417.

Evert, S., Uhrig, P., Bartsch, S. & 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 the eLex 2017 Conference*. pp. 531–549.

Falk, N., Strakatova, Y., Huber, E. & Hinrichs, E. (2021). Automatic Classification of Attributes in German Adjective-Noun Phrases. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics (ACL), pp. 239–249.

Firth, J.R. (1957). *Papers in Linguistics, 1934-1951*. Oxford University Press.

Garcia, M., García Salido, M. & 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)*. Association for Computational Linguistics (ACL), pp. 49–59.

Geyken, A. (2007). The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century. In C. Fellbaum (ed.) *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. Continuum Press London, pp. 23–42.

Goldhahn, D., Eckart, T., Quasthoff, U. et al. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. pp. 31–43.

Hamp, B. & Feldweg, H. (1997). GermaNet - A Lexical-Semantic Net for German. In *Proceedings of ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources*.

Hausmann, F.J. (1984). Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31(4), pp. 395–406.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL*, volume 2013. pp. 125–127.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pp. 105–116.

Lemnitzer, L., Ermakova, M., Palmes, L., Roll, B., Siebel, K. & Geyken, A. (2025). Kollokationen im DWDS-Wörterbuch und ihr Mehrwert für DaF/DaZ. *Deutsch als Fremdsprache*, 62(2), pp. 67–80.

Ljubešić, N., Logar, N. & Kosem, I. (2021). Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 9(2), pp. 41–70.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *LREC 2004: Fourth International Conference on Language Resources and Evaluation*. European Language Resources Association, pp. 1263–1266.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Strakatova, Y. (2024). German Adjective-Noun Co-occurrences with Attributes [Dataset]. University of Tübingen. URL https://doi.org/10.57754/FDAT.76krc-egt63.

Strakatova, Y., Falk, N., Fuhrmann, I., Hinrichs, E. & Rossmann, D. (2020). All that Glitters is not Gold: A Gold Standard of Adjective-Noun Collocations for German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, pp. 4368–4378.

Uhrig, P., Evert, S. & Proisl, T. (2018). Collocation Candidate Extraction from Dependency-Annotated Corpora: Exploring Differences across Parsers and Dependency Annotation Schemes. *Lexical Collocation Analysis: Advances and Applications*, pp. 111–140.