Hard is the Task, the Samples are Few: A German Chiasmus Dataset

Felix Schneider*, Sven Sickert*, Phillip Brandes†, Sophie Marshall‡, Joachim Denzler*

*Computer Vision Group, ‡Institut für Germanistische Literaturwissenschaft
Friedrich Schiller University Jena
{firstname}.{lastname}@uni-jena.de

†Eberhard Karls University Tübingen
phillip.brandes@uni-tuebingen.de

Abstract

In this work we present a novel German language dataset for the detection of the stylistic device called chiasmus collected from German dramas. The dataset includes phrases labeled as chiasmi, antimetaboles, semantically unrelated inversions, and various edge cases. The dataset was created by collecting examples from the GerDraCor dataset. We test different approaches for chiasmus detection on the samples and report an average precision of 0.74 for the best method. Additionally, we give an overview about related approaches and the current state of the research on chiasmus detection.

1. Introduction

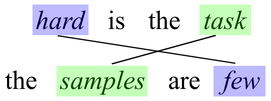

Figure 1: Example of a chiasmus

In this paper, we present a novel dataset containing annotated chiasmus examples. We make this dataset available online [1]. The chiasmus is a figure of speech which comprises an inversion of semantically related concepts, as depicted in Figure 1. This stylistic device can have various uses, such as expressing contrasting concepts. However, it is only sparsely used in literature, making it hard to gather enough instances to train a modern classifier or to conduct meaningful statistical analysis. Brandes et al. (2022) show that the automated detection of chiasmi can provide a useful tool in the field of literary studies. In their case, it is used to find differences between different genres of medieval texts.

The chiasmus dataset can also be used to test the general capabilities of language models. It is no trivial task to fully define from which point on inverted words are semantically related enough to be considered a chiasmus. Consistently detecting chiasmi is a task that requires a nuanced semantic knowledge of the underlying language included in whatever model is used for this task. Thus, this dataset is not only of interest for reseachers in the field of stylistic device detection, but also as a benchmak for the development of language models in general. The dataset consists of German chiasmi, antimetaboles and random inversions without semantic meaning. It was compiled by using a chiasmus detection method (Schneider et al., 2021) on the GerDraCor (Fischer, 2019) dataset while annotating the most highly rated chiasmus candidates. Furthermore, it is enriched by random inversions taken from the GerDraCor corpus.

In the following, we give an overview over related work on chiasmi in Section 2.. After that, we provide the details of the chiasmus dataset in Section 3.. In Section 4., we present and discuss the results of a baseline classifier on the dataset. Finally, Section 5. concludes the paper with a summary of our work.

2. Related Work

Research on chiasmi and their detection originates mainly from two areas of science. Many methodical contributions can be found in the area of computer science, while works from the field of literary studies focus on their historical occurence and meaning in the context of texts.

2.1. Computer Science

A first method to detect antimetaboles was introduced by Gawryjolek (2009). The author presents an approach which searches for repetition of words to find possible antimetaboles. While this approach uses no filtering steps, it can be considered the first step towards the automatic detection of these stylistic devices. Analogously, Java (2015) provides a method for finding general chiasmus candidates based on the inversion of syntax trees. This method also has no filtering step to remove random inversions – instead, the requirement for a full inversion of the syntax tree already narrows down the number of detected candidates. The method potentially misses candidates that can be found by looking at part-of-speech tag repetitions, but its strong requirement also removes false positives. Lim (2016) also search for chiastic structures by locating repeating words without additional filtering. This approach is capable of finding also longer structures like *A B C ... C' B' A'*. As a limitation, it lacks the means to remove false positives. However, the method is suitable to find more deeply stacked chiastic structures, which can not be reliably detected by other methods. Especially with very deeply stacked structures, the likelihood of such often repeated random inversions should be lower than with the two-level chiastic structures presented in this work.

[1] https://github.com/cvjena/chiasmus-annotations

A method to detect antimetaboles using inversions of repeated lemmata was created by Dubremetz and Nivre (2018). In their work, they also searched for inverted repetitions of lemmata, but added a filtering step afterwards. It ranks the chiasticity of the sample based on a machine learning model, which uses several features. This approach is part of a series of works, where they first introduced a manual ranking system (Dubremetz and Nivre, 2015). Later on, they introduced a machine learning approach and added new sets of features (Dubremetz and Nivre, 2016). They test their approach on a dataset of antimetaboles in English created from the Europarl dataset. Schneider et al. (2021) extend this method to find general chiasmi using inversions of repeated part-of-speech tags instead of antimetaboles by inversions of repeated lemmata. They also rank them by a machine learning model. To cope with large amounts of false positives in their part-of-speech tag based approach, they add cosine distances between the word embeddings of the supporting tokens and lemma repetition information to their set of features.

2.2. Literary Studies

In literary studies, the chiasmus is studied as a stylistic device and as a wider structure in text. Welch (2020) gives a broad overview of chiasmus used in antique texts. The work covers a great timespan, beginning from sumero-akkadian literature, covering the Old and New Testament, until the era of ancient greek and latin texts. More recent texts are covered by Brandes et al. (2022). They analyze the use of chiasmus in Middle High German texts in the *Trois Matières*. In their work, they use automatic chiasmus detection techniques to compare the styles of different texts between genres, times and authors.

3. The Dataset

In the following, we describe our proposed dataset. Before we get to the details, we will summarize what a chiasmus is based on definitions found in the literature.

3.1. Chiasmus Definition

There are many definitions of the word chiasmus, which can include very informal descriptions by literary scholars. Those also include purely semantic chiasmi, that comprise opposing concepts, but are not represented by a certain syntactical structure (Welch, 2020). We acknowledge this diversity in the use of the term *chiasmus*. However, we need to set some constraints to operationalize the term for the use in a dataset and benchmark.

In the context of this dataset a chiasmus is considered a cross-wise inversion of semantically related words which can be read as a stylistic device. We further define this as a cross-wise repetition of part-of-speech tags in an *A B B' A'* pattern. An example for this would be the sentence ***narrow*** *is the* ***world*** *and the* ***brain*** *is* ***wide***, comprising the supporting tokens *narrow, world – brain, wide* and the inverted part-of-speech tags *ADJ, NOUN – NOUN, ADJ*.

A special case of this chiasmus definition is the antimetabole, which consists of an inverted repetition of part-of-speech tags and lemmata. An example for this would be the sentence ***one*** *for* ***all***, ***all*** *for* ***one***, with the repeated supporting tokens *one, all – all, one*. Further, we only consider chiasmi comprising two word pairs. More complex chiasmi like *A B C C' B' A'* patterns are not part of this work.

Type		Samples
chiasmus	(c)	31
antimetabole	(a)	39
negative examples	(x)	242
random negative examples	(xr)	4000
parallelism	(fp)	76
antithetic parallelism	(ap)	23
false antimetabole	(fa)	3
false chiasmus	(fc)	7
false antithetic parallelism	(fap)	2
false parallelism	(ffp)	2
synthetic parallelism	(fsp)	22

Table 1: The number of samples per class in our proposed chiasmus dataset.

3.2. Chiasmus Annotations

The samples in our dataset are annotated in different manners. In additon to the base classes of chiasmus, antimetabole and inversion without special semantic meaning, we also include parallelisms. The parallelisms are instances, where the part-of-speech tags of all four supporting tokens are similar, leading to the *A A' A" A"'* structure matching the *A B A' B'* structure of a parallelism. There are also special cases, where the sample constitutes a chiasmus, but the main supporting tokens are not marked. Table 1 gives an overview over the different samples in the data. In the following, we quickly summarize and define the classes used in the dataset.

Chiasmus (c) represents the standard case of a chiasmus with repeating inverted part-of-speech tags, excluding antimetaboles.

Antimetabole (a) stands for antimetaboles, defined as an inverted repetition of lemmata.

Parallelism (fp) stands for parallelisms in the form of *A B A' B'*. Since the candidates are generated by searching for an *A B B A* pattern in the part-of-speech tags, also *A A A A* candidates are found, resulting in some parallelism examples.

Synthetic parallelism (sp) stands for a form of parallelism, where the stylistic device gets its meaning by a related series of statements together. An example would be: *do the research, write the paper, submit the work.*

Antithetic parallelism (ap) describes a form of parallelism with a chiastic semantic meaning, representing a form of opposite.

Near misses (f.∗) are examples that contain a chiasmus or an antimetabole in their scope, but the supporting tokens detected by the algorithm are not the real supporting tokens of the chiasmus. An example would be: *narrow* ***is the*** *world, and* ***the*** *brain* ***is*** *wide*, which is a chiasmus with

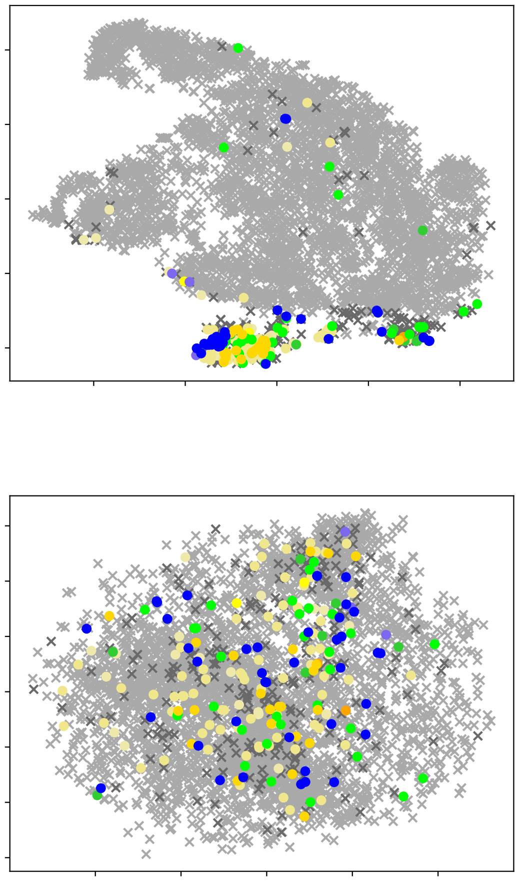

Figure 2: Plots of the T-SNE projected chiasmus datapoints. The upper plot shows the DLE features, the lower plot shows the features from the last hidden state of the DistilBERT model. Non-chiasmi are drawn as crosses, dark for annotated ones and bright for the random inversions. Chiastic samples and parallelisms are repesented as filled circles. The green colored dots are chiasmi, blue colored dots are antimetaboles and yellow colored dots are parallelisms. This figure is best viewed in color.

the real supporting tokens *narrow, world – brain, wide*, but the detected supporting tokens *is, the – the, is*.

Negative Examples (x) are all other part-of-speech tag inversions without special semantic meaning.

Randomly Chosen Negative Examples (xr) are the same as the other negative examples, but were randomly chosen and automatically annotated without human supervision.

For a more detailed descriptions of the above mentioned stylistic devices, we refer to works by (Braungart et al., 2010; Burdorf et al., 2007) and Ueding (1998).

3.3. Dataset Visualization

To give a visual overview of the dataset, we show in Figure 2 a scatterplot of a T-SNE (van der Maaten and Hinton, 2008) projection of the data. The features that were used for this projection were the DLE features by Schneider et al. (2021), which will be explained in Section 4.. The labels for our proposed dataset were created by annotating the top results from the application of the chiasmus detection algorithm by Schneider et al. (2021). It can be seen that many of the annotated negative examples lie close to the annotated positive examples, while the rest of the random inversions can easily be separated. This introduces examples which cannot be easily separated by the existing methods and need further research. On the other hand, it can be seen that the simple DistilBERT (Sanh et al., 2019) approach does not result in features that are easily pre-separated, even though it is already fine-tuned on the dataset.

3.4. Creation

The dataset was created on the basis of the experiments of Schneider et al. (2021). In their work, the authors searched for inverted repetitions of part-of-speech tags in texts from GerDraCor and then ranked the obtained results by different ranking methods. The top 100 results were then annotated by a domain expert. The dataset comprises these annotated results as well as 4000 randomly sampled inversions from the corpus, annotated as negative examples. Since the chiasmus is such a scarce phenomenon, this can be safely done. However, every example has an annotation indicating whether it is annotated by a domain expert or sampled randomly as a negative example.

3.5. Source and Dataset Format

The underlying data is drawn from the GerDraCor corpus, which comprises plays in the German language published between 1650 and 1940. As a result, some words might be spelled differently in the texts than they would be in modern German. Also, when taking into account the long timespan, some words may have gone through semantic changes (Schlechtweg et al., 2017; Koch, 2016).

The search window for the part-of-speech tag inversions for creating the dataset had a size of 30. That is, the distance from the first to the last supporting token spans at most 30 tokens. The phrases contained in the dataset have additionally 5 tokens before the first and after the last supporting token as context. The mean of the phrase length from the first to the last supporting token is 22 tokens, with a standard deviation of 6 tokens.

The data format is a JSON file. Every entry consists of seven different fields. The field *ids* contains the offset of the four supporting tokens in the source files, *cont_ids* describes the offset of the whole phrase. The different tokens are recorded in *tokens*, the lemmas in *lemmas* and the part-of-speech tags in *pos*. The results of the dependency parsing can be found in *dep*. Finally, the annotations are contained in *annotation*.

For tokenizing, lemmatizing, part-of-speech tagging, and for the dependency trees we used the spaCy library (Honnibal et al., 2020). However, the word embeddings which we used in the experiments were not created with spaCy, since the models included there only create embeddings for words contained in their dictionary. Instead, we used the German FastText (Bojanowski et al.,

	D	**DL**	**DE**	**DLE**	**DistilBERT**
full	0.49 ± 0.26	0.65 ± 0.32	0.73 ± 0.35	0.72 ± 0.35	0.10 ± 0.07
fp removed	0.61 ± 0.30	0.69 ± 0.33	0.74 ± 0.35	0.74 ± 0.35	0.07 ± 0.08

Table 2: Baseline results for the chiasmus detection. The values represent the mean average precision and their standard deviation.

POS Tag	**Percentage**	**Parallelisms**
NOUN	45.7%	92.1%
PRON	34.3%	03.9%
VERB	13.6%	01.3%
PROPN	04.3%	01.3%
DET	01.4%	01.3%
ADJ	00.7%	00.0%

Table 3: Percentage of the different part-of-speech tags in the positive examples.

2017) model available on the FastText website (Grave et al., 2018).

3.6. Special Cases and Biases

One source of bias is that all positive examples were found by using a single algorithm with various sets of features, all trained on the same dataset. While every positive example was annotated manually, this means that positive examples which were not fitting the criteria in the ranking algorithm may be left out of the datset. The whole number of chiasmi in the GerDraCor corpus is not known. Thus, no quantitative statement can be made about this potential limitation.

Table 3 shows the distribution of different part-of-speech tags in the positive examples. We used positive annotations for chiasmi and antimetaboles, as well as the parallelisms. It can be seen that most chiasmi are based on repeated nouns, followed by pronouns. Determiners and especially adjectives make up the least part of the dataset. This may be a source of bias, since some better known chiasmus examples like ***narrow** is the **world** and the **brain** is **wide*** are also based on adjectives.

Another potential bias source is the annotation. Since the annotations were done by a single domain expert, metrics like inter-annotator agreement can not be reported.

3.7. Examples

In the following we show some examples of chiasmi and antimetaboles in the dataset. The examples are first given in German, followed by their English translation. The last line contains the class of the example. The examples were chosen since they carry obviously semantically related meaning between their main words and are thus very prototypical and unambiguous examples. The last three examples got very low scores with all feature combinations.

- O **Augen** ohne **Kopf**, o **Kopf** ohne **Augen**
  O **eyes** without **head**, o **head** without **eyes**
  Antimetabole

- **Dir** widert **Landluft**, **Seeluft** widert **dir**.
  You dislike **country air**, the **sea air** disgusts **you**.
  Chiasmus

- . . . der Menschen **belustigen mich** lange, eh sie **mich reizen**.
  . . . of the humans **amuse me** for long before they **irritate me**.
  Chiasmus

- **Ich** bin nicht, was ich **scheine**, und **scheine** auch nicht, was ich bin, Und wenn ich das wäre, was **ich** sein möchte
  I am not what I **seem** and do not **seem** what I am, and if I would be that, what **I** want to be
  Chiasmus

- Ja **ich** hab einen **Sohn** gequält, und ein **Sohn** mußte **mich** wieder quälen
  Yes **I** have tortured a **son**, and a **son** had to torture **me** again
  Chiasmus

- Meinst **du** damit etwa **mich**? Mein **ich** damit etwa **dich**?
  Do **you** happen to mean **me**? Do **I** happen to mean **you**?
  Antimetabole

4. Baseline Benchmark

For our baseline benchmark experiments, we use the approach by Schneider et al. (2021). Following their evaluation, we compare different feature sets presented in this work, including the features proposed by Dubremetz and Nivre (2018).

4.1. Methods

Chiasmus candidates are ranked by their chiasticity using a support vector machine (SVM) with an RBF kernel (Schölkopf and Smola, 2001). The regularization parameter for the SVM model is 1 with a maximum of 1000 iterations for the fitting of the model. The features are preprocessed to a mean of 0 and scaled to unit variance.

In the following we summarize the features. For a full explanation, we refer to the respective works:

Dubremetz features (D) include various basic features that are already useful for the detection of antimetaboles. These features include the usage of certain words like negations, the usage of punctuation, the repetition of words in certain parts of the example as well as the repetition of syntax tree tags. For a complete summary, please see the work of Dubremetz and Nivre (2018).

Lexical features (L) are binary features that indicate in a pairwise manner whether the main words constituting the chiasmus comprise repeating lemmata. They were proposed by Schneider et al. (2021).

Embedding features (E) describe the pairwise cosine distance of the word embeddings of the main words constituting the chiasmus and thereby indicate their relation. They were proposed by Schneider et al. (2021).

In addition, we also conducted experiments using a BERT-like language model (Devlin et al., 2019). We used *distilbert-base-german-cased*, a DistilBERT (Sanh et al., 2019) model trained on German texts, which is available for download on the huggingface website [2] with the pytorch implementation of the language model. We presented the data as one text string per example, as the tokens appear in the dataset, with the tokens separated by single spaces. The finetuning was conducted for 5 epochs with a batch size of 128, using the AdamW optimizer with a lerning rate of 0.001. We trained the model for 20 epochs with a weight decay of 0.01.

4.2. Results

Table 2 shows the results of the chiasmus detection. The experiments were conducted by using 5-fold cross-validation with 80% of the data used for training and 20% used for testing. For evaluation, we report the *average precision* metric. This information retrieval criterion describes the area under the precision-recall curve. This metric was chosen because the main interest of chiasmus detection is to extract chiasmi from a longer corpus instead of just classifying instances. The two experiments presented, *full* and *fp removed* show a different choice of samples. In *full*, all positive samples were annotated with *a, c, fa*, and *fc*. In *fp removed*, the positive examples were *a* and *c*, while *fa* and *fc* were removed from the dataset. It can be seen from the results that both sets of choices result in similar differences between average precision values. *D*, *DL*, *DE*, and *DLE* stand for different combinations of the features explained above. Please see Sec. 3. for an explanation of the acronyms. Additionally, the results of the DistilBERT experiment are included.

The combination of all features yields either the best or the second best results on our dataset. For the *full* experiment, the results deviate slightly. However, the standard deviation of the results is also very high, which implies that the small improvement of the *DL* combination over the *DLE* combination may be attributed to random noise. Since the repeated lemmas should also have very similar word embeddings, even if the tokens themselves differ, a lot of the information from the lemma repetition features is already included in the word embedding features. In comparison, the DistilBERT experiment performs worse in our case. As BERT-like models have shown superior performance in many NLP applications, this unexpected result needs further investigation.

To find out whether the improvements of the feature combinations compared to the baseline *D* were statistically significant, we ran a 5x2cv (Dietterich, 1998) evaluation on the full dataset. *DLE* showed an improvement in average precision from 0.2 to 0.51 with a *p* value of 0.06 compared to only using *D* features. *DL* increased average precision to 0.41 with a *p* value of 0.18, while *DE* also yielded 0.51 (like *DLE*), but with a higher *p* value of 0.09 instead of 0.06. While the similar average precision of the *DL* and the *DLE* combinations makes the lexical features look less useful, the lower *p* value of the *DLE* combination indicates with its lower *probability of the improvement being random* the importance of these features.

5. Conclusions

In this paper, we presented a dataset containing instaces of chiasmi and antimetaboles, as well as parallelisms and part-of-speech tag inversions without special semantic meaning. The data is annotated in a way that opens up different possibilities on how to use it in the future, including which subset to use and what parts to exclude. It is difficult to define how exactly the supporting tokens of a chiasmus candidate need to be related to constitute a real chiasmus. Thus, chiasmus detection is a hard problem that needs further research.

At the same time, it is suitable to evaluate different kinds of language models on it. Our baseline experiments show that the task in principle is solvable. However, there is still much room for improvement with respect to performance. We hope that this dataset will encourage new research in that field and will be used to improve both the detection of chiasmi and the understanding of language models in general.

References

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov, 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Brandes, Phillip, Felix Schneider, and Sophie Marshall, 2022. Stilfiguren aus der Distanz gelesen. Zur automatischen Detektion von Wortstellungsfiguren und deren Nutzen für die qualitative Analyse.

Braungart, Georg, Harald Fricke, Klaus Grubmüller, Jan-Dirk Müller, Friedrich Vollhardt, and Klaus Weimar (eds.), 2010. *Reallexikon der deutschen Literaturwissenschaft*. Berlin, Boston: De Gruyter.

Burdorf, Dieter, Christoph Fasbender, and Burkhard Moennighoff, 2007. *Metzler Lexikon Literatur*.

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

[2] https://huggingface.co/distilbert-base-german-cased

Dietterich, Thomas G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923.

Dubremetz, Marie and Joakim Nivre, 2015. Rhetorical figure detection: the case of chiasmus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. Denver, Colorado, USA: Association for Computational Linguistics.

Dubremetz, Marie and Joakim Nivre, 2016. Syntax Matters for Rhetorical Structure: The Case of Chiasmus. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*. San Diego, California,USA: Association for Computational Linguistics.

Dubremetz, Marie and Joakim Nivre, 2018. Rhetorical figure detection: Chiasmus, epanaphora, epiphora. *Frontiers in Digital Humanities*, 5:10.

Fischer, Frank, 2019. Programmable corpora: Introducing dracor, an infrastructure for the research on european drama.

Gawryjolek, Jakub Jan, 2009. Automated annotation and visualization of rhetorical figures.

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

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd, 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Java, James, 2015. *Characterization of Prose by Rhetorical Structure for Machine Learning Classification*. Ph.D. thesis, Nova Southeastern University.

Koch, Peter, 2016. *Meaning change and semantic shifts*. Berlin, Boston: De Gruyter Mouton, pages 21–66.

Lim, SeungJin, 2016. An algorithm for detection of chiastic structures in text databases. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Schlechtweg, Dominik, Stefanie Eckmann, Enrico Santus, Sabine Schulte im Walde, and Daniel Hole, 2017. German in flux: Detecting metaphoric change via word entropy. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics.

Schneider, Felix, Björn Barz, Phillip Brandes, Sophie Marshall, and Joachim Denzler, 2021. Data-driven detection of general chiasmi using lexical and semantic features. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Punta Cana, Dominican Republic (online): Association for Computational Linguistics.

Schölkopf, Bernhard and Alexander J. Smola, 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press.

Ueding, Gert, 1998. *Historisches Wörterbuch der Rhetorik*. Tübingen: Niemeyer.

van der Maaten, Laurens and Geoffrey E. Hinton, 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Welch, John W, 2020. *Chiasmus in antiquity*. Wipf and Stock Publishers.