Design and Implementation of German Legal Decision Corpora

Stefanie Urchs[1][a], Jelena Mitrović[2][b] and Michael Granitzer[2][c]
[1]*Chair of Software Engineering for Business Information Systems, Technical University of Munich, Garching, Germany*
[2]*Chair of Data Science, University of Passau, Passau, Germany*

Keywords: Datasets, German Law, German Legal Writing Styles, Machine Learning, Natural Language Processing.

Abstract: Law professionals are wordsmiths, their main tool is language. Therefore, the field of law produces a vast amount of written text. These texts have to be analysed, summarised, and used in the creation of new text, which is a task that reaches the limits of what is humanly possible. However, it is possible to automate this analysis by using Natural Language Processing techniques. To perform these techniques (annotated) text corpora are required. Unfortunately, publicly available (annotated) legal text corpora are rare. Even scarcer is the availability of (annotated) German legal text corpora.
To meet this need for publicly available German legal text corpora this paper presents two German legal text corpora. The first corpus contains 32,748 decisions from 131 German courts, enriched with metadata. The second one is a subset of the first corpus and consists of 200 randomly chosen judgements. In these judgements a legal expert annotated the components conclusion, definition and subsumption of the German legal writing style Urteilsstil. Furthermore, the paper presents experiments on these corpora.

1 INTRODUCTION

In the field of law a vast amount of written text is produced every day. It is not humanly possible to read every text and process all these information. Therefore, law professionals are in need of computational help. Computers can process thousands of documents in mere seconds. By utilising machine learning it is even feasible to go beyond pure digitisation of a document. It is possible to analyse its semantics. However, these machine learning methods need legal text corpora for training purposes. Unfortunately, there are only few annotated legal corpora publicly available, even less annotated German legal text corpora. Non-annotated content is easily available, however, this content has to be crawled and processed for Natural Language Processing usage.

In this paper we present two German legal text corpora. The first one consists of 32,748 decisions of 131 German courts that are enriched with meta data. The second corpus is formed from a subset of these decisions. It consists of 200 randomly chosen judgements, which are annotated with the components *conclusion*, *definition* and *subsumption* of the German legal writing style *Urteilsstil* (appraisal style). These annotations can be utilised to train machine learning models to automatically detect parts of German legal writing styles.

The remainder of this paper is structured as follows. Chapter 2 introduces the related work of this paper. The first corpus is presented in chapter 3, the annotated judgement corpus in chapter 4. In chapter 5 possible use cases for the corpora are discussed and chapter 6 concludes the paper.

2 RELATED WORK

This chapter introduces a selection of already available legal corpora. Furthermore, German legal writing styles are discussed, especially the *Urteilsstil*.

2.1 Legal Corpora

In 2006 Reed et al. (Reed, 2006) presented the first corpus containing analysed legal argumentation. This corpus contains text from various domains, including judicial summaries and discussion. Due to a security breach this corpus is no longer available. Researchers from the same group presented another legal corpus: a collection of 47 documents from the Euro-

[a] https://orcid.org/0000-0002-1118-4330
[b] https://orcid.org/0000-0003-3220-8749
[c] https://orcid.org/0000-0003-3566-5507

Urchs, S., Mitrović, J. and Granitzer, M.
Design and Implementation of German Legal Decision Corpora.
DOI: 10.5220/0010187305150521
In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) - Volume 2*, pages 515-521
ISBN: 978-989-758-484-8

pean Court of Human Rights (Palau and Moens, 2009; Mochales and Moens, 2011), annotated with argumentation components and argumentation schemes. Unfortunately, they did not publish this corpus.

The US courts publish their cases online. Sugathadasa et al. (Sugathadasa et al., 2017) compiled 22,776 cases from the United States supreme court into a corpus and published it in combination with other legal cases. Judicial decisions from the UK form the basis of the corpus of Rizzo et al.(Rizzo and Pérez, 2012). This corpus was created and published because most legal corpora are too small to act as normative reference or are not publicly available. The authors cite the corpus of Favretti et al. (Rossini Favretti et al., 2001) as one of the few corpora comparable to their work. It contains European legislation in English and Italian. A corpus consisting of Italian legislation in Italian and German is presented by Gamper (Gamper, 2000). However, the paper does not mention publication of this corpus.

The international research group Computer Assisted Legal Linguistics (CAL^2)[1] compile legal corpora, containing legislation from around the world. Unfortunately, these corpora are not publicly available.

These selected corpora might be valuable for research if they would have been published. To fill the gap, left open by current research, we publish our corpora on the open science platform zenodo[2].

2.2 German Legal Writing Styles

The legal way of argumentation is the so called legal syllogism (Alexy, 1983). The example in table 1 illustrates a legal syllogism. At first a general norm is cited as first premise, subsequently the real world case is stated and a conclusion is deducted from the two premises. In other words, general norms/laws are used to solve concrete problems, by logically deducting the solution from the norms/laws (Weber, 2018).

The used example shows how similar the thinking in mathematical logic and legal reasoning is. However, in the legal reasoning process one encounters some difficulties. For example: What is rain in a legal sense? It is necessary to define ”fuzzy” terms of norms/laws. When rain is defined one has to look at the concrete problem and ask if that water coming from the sky is rain in the sense of the common rules. Here a subsumption is necessary (Weber, 2018).

The German legal education teaches law students to work on legal problems in two distinct styles. First

[1]https://cal2.eu/index.php

[2]https://zenodo.org/record/3936726#.X1enLIvgomL

Table 1: Example of legal deduction with a syllogistic inferring method (Alexy, 1983).

General norm / law	Major premise	A soldier has to say the truth in official business affairs (§13 Abs. I SoldatenG).
Concrete problem	Secondary premise	Mr M. is a soldier.
Final sentence	Conclusio	Mr M. has to say the truth in official business affairs.

the *Gutachtenstil* (appraisal style), which is used almost exclusively until the first state exam. After the first state exam, the *Urteilsstil* (judgement style) is used. Both styles share the components `definition` and `subsumption`. The appraisal style is concluded in a different way than the judgement style, therefore this component does not match.

The introduced writing styles might be used in other countries in a similar way. However, this paper introduces only the German way of formal legal writing, other countries are not considered.

Urteilsstil. The *Urteilsstil* begins with the conclusion and proceeds with the reasoning. The most basic version of the *Urteilsstil* consists of three stages:

1. the concrete legal consequences
2. followed by the abstract legal facts and consequences (i.e. exact wording of the law)
3. ending in the concrete facts

Between the second and third item a *Feststellungssatz* (determination sentence, the result of the subsumption) states if the legal requirements are fulfilled or not (Danger, 2005).

The determination sentence and the concrete facts together form the subsumption. This leads to the following structure as shown in Table 2.

In practice this basic schema is mostly extended. If norms refer to other regulations, these must first be defined before a subsumption can be made.

3 GERMAN LEGAL DECISION CORPUS

This chapter introduces a legal decision corpus, enriched with metadata and saved in an easily accessible JSON format. To show a possible use-case of the corpus, an experiment on the corpus is presented.

Table 2: Example of *Urteilsstil* with named subcomponents according to (Danger, 2005).

Conclusion	Overall Result	The claim is well justified. The defendant owes the plaintiff 1500€ in damages.
Definition	Abstract Legal Facts and Consequences	If an object is damaged by an animal,then according to §833 S 1 BGB the person who keeps the animal is obliged to compensate the injured person for the resulting damage.
Subsumption	Determination Sentence	These conditions are fulfilled here.
	Concrete Facts	The defendant's dog scratched the paint of the plaintiff 's car. The repainting caused costs of 1500€.

3.1 Corpus Design

The decision corpus is a collection of the decisions published on the website www.gesetze-bayern.de[3]. At the time of the crawling the website offered 32,748 decisions of 131 Bavarian courts, dating back to 2015. The decisions are provided from the Bavarian state after the courts agreed to a publication. All decisions are processed by the publisher C.H.BECK, commissioned by the Bavarian state. This processing includes anonymisation, key-wording, and adding of editorial guidelines to the decisions. Intentionally all decisions are included to provide a complete corpus of German law. Normalisation measures can be applied depending on the task at hand.

The decisions have 22 different types, mostly resolutions (*Beschluss*, 16,711 (51%)), judgements (*Urteil*, 11,955 (37%)), or end-judgements (*Endurteil*, 2,748 (8.3%)). From the 131 courts the most decisions are made by VG Munich (8,028 (25%)), VGH Munich (7,141 (22%)) and OLG Munich (2,337 (7%)). The courts of the Bavarian capital are generally the most busy ones. The VGH Munich is the *Verwaltungsgerichtshof* (Higher Administrative Court) of Bavaria, the highest instance for all topics regarding the public administration. Below the VGH exist six VG (*Verwaltungsgericht* / Administrative Court) almost all administrative regions have their own VG, except for lower Bavaria, which fall under the jurisdiction of VG Regensburg. This type of court (VG) issues the most decisions in the corpus (22,155(68%)). Figure 1 depicts that over the year courts are mostly consistent in issuing decisions. The first three months are busier that the rest of the year. August, September and December are the months with the least number of issued decisions.

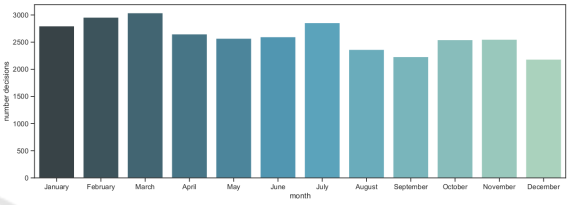

Figure 1: Number of decisions issued per month.

Every decision is saved in the following JSON format:

```
{
    "meta": {
        "meta_title": "",
        "court": "",
        "decision_style": "",
        "date": "",
        "file_number":"",
        "title": "",
        "norm_chains": ["", ""],
        "decision_guidelines": ["", ""],
        "keywords": "",
        "lower_court": ["", ""],
        "additional_information": "",
        "decision_reference": ""
    },
    "decision_text":{
        "tenor": ["",""],
        "legal_facts": ["",""],
        "decision_reasons": ["",""]
    }
}
```

The metadata are provided by the publisher C.H.BECK. The following is a short description of the fields:

- **`meta title`**
  Title provided by the website, it is used for saving the decision
- **`court`**
  Issuing court
- **`decision style`**

[3]https://www.gesetze-bayern.de/ accessed and crawled on 2020-05-13

Style of the decision

- **date**
  Date when the decision was issued by the court
- **file number**
  Identification number used for this decision by the court
- **title**
  Title provided by the court
- **norm chains**
  Norms related to the decision
- **decision guidelines**
  Short summary of the decision
- **keywords**
  Keywords associated with the decision
- **lower court**
  Court that decided on the decision before
- **additional information**
  Additional Information
- **decision reference**
  References to the location of the decision in beck-online
- **tenor**
  Designation of the legal consequence ordered by the court
- **legal facts**
  Facts that form the base for the decision
- **decision reasons**
  In depth explanation of the court decision

The corpus is published on the open research platform zenodo[4].

3.2 Experiments

As mentioned above, the corpus is intentionally not normalised towards a specific task. Therefore, the corpus can be used for many different research questions. One of possible question is: "Is it possible to detect the type of a decision"?

Overall 22 types of decisions are contained in the corpus. However, 16 of these types are represented by less than 100 decisions. Furthermore, the corpus contains nine different kinds of judgements and six different kinds of resolutions. To form a balanced dataset all judgements get the label `judgement` and all kinds of resolutions are labelled as `resolution`, everything else is labelled as `other`. Leading to 17,013 (52%) decisions labelled as `resolution`, 14,818 (45%) decisions labelled as `judgement` and 917 (3%) as `other`. As a feature for the classification term frequency - inverse document frequency (tf-idf) is calculated for the decision reasons of each decision. The *scikit learn*[5] TfidfVectorizer[6] is used. With this feature a logistic regression[7] (LR) and a linear support vector classification[8] (SVC) are trained in a one-vs-the-rest multilabel scheme. This means that one class is accepted as positive and all others as negative. For each class a classifier is trained and finally the prediction with the highest probability is returned. Table 3 shows the results of training on 80% of the data and testing with the remaining 20%. Although a very simple feature was used, the classifications are very good. The SVC slightly outperforms the LR.

This experiment is only meant to illustrate the usefulness of the corpus. For this reason, no further discussion of how the results of the classification were obtained will be given here, this will be discussed in future work.

Table 3: Results of training a logistic regression and a support vector classification on the decision corpus. The support vector classification slightly outperformed the logistic regression.

Feature/ Classifier	Precision	Recall	F1-Measure	Accuracy
tf-idf/ LR	0.96	0.80	0.85	0.96
tf-idf/ SVC	**0.97**	**0.88**	**0.92**	**0.97**

4 GERMAN LEGAL JUDGEMENT CORPUS

As stated above, the judgement corpus consist of 200 randomly chosen judgements that are annotated by a legal expert, who holds a first legal state exam. According to (Wissler et al., 2014) multiple experts are needed to create a gold standard corpus. Due to financial, staff and time reasons the presented iteration of the corpus was only annotated by a single expert. In a future version several other experts will annotate the corpus and the inter-annotator agreement will be calculated.

The decision reasons of German legal judgements

[4]https://zenodo.org/record/3936726#.X1enLIvgomL

[5]https://scikit-learn.org

[6]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[7]https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[8]https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

are written in *Urteilsstil*. Law students are first exposed to this writing style after finishing their first state exam.

In the judgement corpus the components `conclusion`, `definition` and `subsumption` of the *Urteilsstil* are annotated. All other sentences are labelled as `other`. The corpus is created as part of the master thesis of Stefanie Urchs (Urchs, 2020). The goal of the thesis is to train a model on *Urteilsstil* and use it on *Gutachtenstil*. As outlined in 2.2 the *Urteilsstil* and *Gutachtenstil* share the components `definition` and `subsumption`. Therefore, the model trained on judgements in *Urteilsstil* should be able to withstand a domain transfer to *Gutachtenstil* and detect `definitions` and `subsumptions` in legal exercise cases.

This chapter explores the 200 annotated legal judgements and presents an experiment performed on the corpus.

4.1 Corpus Design

Overall 25,075 sentences are annotated. 5% (1,202) of these sentences are marked as `conclusion`, 21% (5,328) as `definition`, 53% (13,322) are marked as `subsumption` and the remaining 21% (6,481) as `other`. The length of judgements in sentences ranges from 38 to 862 sentences. The median of judgements have 97 sentences, the length of most judgements is on the shorter side.

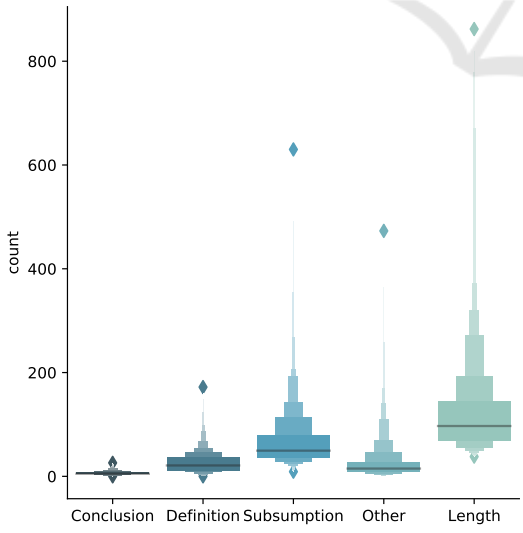

Figure 2: Number of sentence per label on a judgement basis. The last part of the figure depicts the number of sentences per judgement.

Figure 2 shows that the amount of marked `conclusions` and `definitions` is low in all documents. Furthermore, the length of a judgement only slightly increases the number of sentences marked as `definition` in a judgement. `Subsumptions` contrast this behaviour. The number of sentences labelled with `subsumption` correlated strongly with the number of sentences in a judgement. The median of sentences labelled with `conclusion` is 5 per judgement, `definition` is 21 per judgement and for `subsumption` 50 per judgement. This shows that independent of the length of a judgement the conclusion of the overall result and the description of the abstract law is short. Facts are stated without much detour. However, on average law professionals use more than twice as many sentences to subsume the case. When looking back to chapter 2.2 this seems reasonable. The subsumption is the place to argue and combine the pure listing of the law from the definition with the real world case. Law professionals can show their skills in this component. The remaining sentences are labelled as `other`. These sentences include text that does not belong to one of the writing style components. Additionally, the annotator is instructed to only label sentences as `conclusion`, `definition` or `subsumption`. If one sentence contains both it is labelled as other.

As a guiding example the annotation guide for the annotator includes the example stated in table 2.

Judgements from 22 of the 131 courts are selected for the corpus. Most judgements originate from the *VG Augsburg* (59 / 30%) followed by the *VG Ansbach* (39 / 20%) and *LSG Munich* (33 / 17%). For most courts one to four judgements are selected. However, the analysis of the base data already revealed that *Verwaltungsgerichte* (VG) contribute more to the corpus than other courts. The *LSG Munich* is in place four for contributing judgements to the base corpus. Therefore, the random selection of judgements seems to reflect the underlying distribution of decision by courts well.

29% (58) of all selected judgements are issued in the year 2016, followed by 22% (44) from the year 2017 and 21% (41) issued in the year 2015. This selection deviates slightly from the base data where 24.5% (7,663) decisions are issued in 2015, 23.9% (7,487) in 2016 and 21.7% of the decisions are issued in 2017. The percentages of selected judgements and decisions issued in 2018 and 2019 are roughly the same. No judgements from 2020 are selected. However, decisions of 2020 form only 0.4% of the base data, which makes not selecting them a valid representation of the base data.

The JSON format from 3.1 is extended in the following way:

```
{
   "meta": {
      "meta_title": "",
      "court": "",
      "decision_style": "",
      "date": "",
      "file_number":"",
      "title": "",
      "norm_chains": ["", ""],
      "decision_guidelines": ["", ""],
      "keywords": "",
      "lower_court": ["", ""],
      "additional_information": "",
      "decision_reference": ""
   },
   "decision_text":{
      "tenor": ["",""],
      "legal_facts": ["",""],
      "decision_reasons": [
      [[text, label],[text, label]]
      ]
   }
}
```

The decision reasons are segmented into sentences using the SoMaJo[9] tool. SoMaJo is a state of the art tokenisation and sentence segmentation tool for German and English. SoMaJo was chosen after testing several common segmentation tools, as it is the only one that performs acceptably on German legal texts. Each sentence is paired with a label. To preserve the paragraph structure given by the website, the sentences are saved in a list for every paragraph, resulting in a list of lists format.

The corpus is published on the open research platform zenodo[10].

4.2 Experiments

To automatically detect `conclusion`, `definition` and `subsumption` in legal text a logistic regression and a SVC are trained on the judgement corpus. The implementation of the logistic regression and the SVC in *scikit learn* offers a multi class classification option. The algorithm decomposes a multi-class problem into a binary problem. The correct prediction is determined in a one vs. the rest approach.

Two simple features are used in this experiment all unigrams[11] of the training data and a tf-idf[12] for the vocabulary of the training data.

[9]https://github.com/tsproisl/SoMaJo

[10]https://zenodo.org/record/3936490#.X1ed7ovgomK

[11]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

[12]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Table 4 shows the results of a five fold cross-validation over logistic regressions and SVM trained with either unigrams or tf-idf. The baseline is formed by a decision stump that always predicts the majority class. The baseline is always outperformed, therefore, meaningful features are chosen. Interestingly, LR with unigram features and SVC with tf-idf features perform the same. However both are outperformed by a LR with tf-idf features in terms of precision. All three share the same accuracy of 0.77.

Table 4: Results of training a Logistic Regression and a Support Vector Classification on the judgement corpus. The tf-idf outperforms the Unigram feature independent of the classification method.

Feature/ Classifier	Precision	Recall	F1-Measure	Accuracy
Unigrams/ decision stump	0.13	0.25	0.17	0.53
Unigrams/ LR	0.74	**0.67**	**0.70**	**0.77**
Unigrams/ SVC	0.67	0.66	0.66	0.74
tf-idf/ decision stump	0.13	0.25	0.17	0.53
tf-idf/ LR	**0.79**	0.63	0.68	**0.77**
tf-idf/ SVC	0.74	**0.67**	**0.70**	**0.77**

5 DISCUSSION

The presented corpora open many different opportunities to explore German law. Some of the use-cases are the presented experiments, other possible use-cases for the decision corpus are the analysis of topics in courts over time and across courts. Are there seasonal topics in the courts, that reoccur every year? Do topics change over the years? When are more or less decisions issued and does that depend on the court?

Besides the pure analysis of the meta-data an analysis of the textual parts might provide other fascinating insights. It might be interesting to explore whether the authors of decisions can be identified according to their writing style. A sentiment analysis on the textual part can lead to a deeper insight into what the author really thought about a case.

Furthermore, the German legal terminology can be extracted from the corpus and modelled into an upper level ontology like SUMO (Mitrović et al., 2019). Such an ontology is helpful to automatically explore

unseen German legal documents and analyse their contents.

The judgement corpus can be used for writing style detection in free text. Furthermore, it is possible to look into the inner workings of the judgement. Which `conclusion` is connected to which `definition` and which `definition` belongs to which `subsumption`. Additionally, judgements from different courts could be compared in order to answer whether there are big differences between the writing styles of the courts.

Based on the work presented in Mitrović et al. (Mitrović et al., 2017) writing style components can be represented ontologically, and their persuasiveness assessed based on the rhetorical elements contained therein.

6 CONCLUSION

This paper presents two novel German legal corpora based on Bavarian Court decisions between 2015 and 2020.

The first one contains 32,748 decisions from 131 German courts. A model that predicts the type of decision was trained on this corpus. Resulting in a precision of 0.97.

The second corpus is a subset of the first one. 200 judgements were randomly chosen and annotated with `conclusion`, `definition`, `subsumption` and `other`, components of the *Urteilsstil*. On this corpus several models were trained to predict to which component a sentence of a judgement belongs. The baseline is always outperformed, however no clear best approach could be determined. LR performs well with unigrams and SCV performs the same with tf-idf.

Both corpora are published on the open science platform zenodo.

In future work different legal experts will inspect the existing labels to ensure the label quality.

ACKNOWLEDGEMENTS

SPONSORED BY THE

Federal Ministry
of Education
and Research

The annotation of the judgement corpus was only possible due to the help and funding from the research centre FREDI[13].

The project on which this report is based was funded by the German Federal Ministry of Education and Research (BMBF) under the funding code 01—S20049. The author is responsible for the content of this publication.

REFERENCES

Alexy, R. (1983). *Theorie der juristischen Argumentation*. Suhrkamp Verlag AG.

Danger, D. J. (2005). Urteil und urteilsstil in der zivilrechtlichen assessorklausur: eine praktische hilfestellung. *Juristische Arbeitsblätter*, (07).

Gamper, J. (2000). A parallel corpus of italian/german legal texts. In *LREC*.

Mitrović, J., O'Reilly, C., Mladenović, M., and Handschuh, S. (2017). Ontological representations of rhetorical figures for argument mining. *Argument & Computation*, 8(3):267–287.

Mitrović, J., Pease, A., and Granitzer, M. (2019). Modeling legal terminology in sumo.

Mochales, R. and Moens, M.-F. (2011). Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Reed, C. (2006). Preliminary results from an argument corpus. *Linguistics in the twenty-first century*, pages 185–196.

Rizzo, C. R. and Pérez, M. J. M. (2012). Structure and design of the british law report corpus (blrc): a legal corpus of judicial decisions from the uk. *Journal of English Studies*, 10:131–145.

Rossini Favretti, R., Tamburini, F., and Martelli, E. (2001). Words from bononia legal corpus. *International journal of corpus linguistics*, 6(1-12):13–34.

Sugathadasa, K., Ayesha, B., de Silva, N., Perera, A. S., Jayawardana, V., Lakmal, D., and Perera, M. (2017). Synergistic union of word2vec and lexicon for domain specific semantic similarity. *IEEE International Conference on Industrial and Information Systems (ICIIS)*.

Urchs, S. (2020). Extracting definition and subsumption from german law. Master's thesis.

Weber, D. (2018). *Methodik der Fallbearbeitung*. Nomos Verlagsges.MBH + Co.

Wissler, L., Almashraee, M., Díaz, D. M., and Paschke, A. (2014). The gold standard in corpus annotation. In *IEEE GSC*.

[13] https://www.jura.uni-passau.de/fakultaet/forschungseinrichtungen/forschungsstelle-fredi/