# A Dataset for the Legal Argument Reasoning Task based on The Glannon Guide to Civil Procedure

The book “The Glannon Guide To Civil Procedure” by Joseph Glannon contains questions with answers and a corresponding analysis to educate students on U.S. civil procedure. We automatically extracted the questions and their answers and formed a new dataset for legal Natural Language Processing (NLP) experiments with the aim to enhance legal NLP models and show their current incapability of reasoning. For example in the area of deep learning, datasets like this are a necessity for further exploration. Despite its necessity, datasets for legal reasoning are sparse and their lack impedes further research.

## Dataset origin and size

The book consists of 27 chapters. For our dataset, we extracted questions and contextual information from Chapters 2 - 25. We formulated our problem as a binary classification task which means that each single entry in the dataset (a data point) contains at least a question, an answer, and a label, which states whether the given answer is correct or not.

As the book frames the problems as multiple-choice questions, we created a separate data point for each answer candidate. So if there are four possible answers to one question, four data points were created.

In addition to the minimum requirements for a binary classification dataset, the book offers further content.

Each subchapter contains an introductory text that is relevant to the question that follows. Furthermore, an analysis part is offered after each question, which goes deeper into the reasoning why an answer is right or wrong. The analysis is a continuous text. However, most of the time the individual answer analysis can be easily differentiated. These additional pieces of information are also entries of the data points. For our dataset, we manually separated the analysis sections from each other, but since this was not always possible, we also added the complete analysis of all possible answers to the question as a data point entry.

In the current dataset 918 data points are collected by an automatic, rule-based extraction method. Overall, each data point consists of: (question, answer, label, analysis, complete analysis, explanation). An example is in the Appendix.

## Purpose of sharing the dataset

The main purpose of sharing this data set for research purposes is to enable full reproducibility. For reproducibility in research, it is very important to compare the results of methods obtained with exactly the same data sets. The ability to compare the results of different scientific approaches is also critical and requires the use of the same data for each approach. By making this dataset accessible for other research teams, research in legal argumentation in natural language processing can be advanced.

## Terms of use

We wish to be able to share the data set with the research community (“users”) under the following conditions:

* Users of the dataset may use it for noncommercial research only.
* The dataset may not be distributed further and must be deleted after completing the experiments.
* For each publication based on the dataset, credits should be given to the author of the book, the authors of the dataset, and the publisher.

To ensure the user agrees to these terms, we require users to submit an application. This application includes signing the license agreement, personal information, and a description for which experiment the dataset should be used.

## Acquiring the dataset and dataset lifecycle

To get access to the dataset, the request form (see below) must be filled out and sent via e-mail to Dr. Ivan Habernal ([ivan.habernal@tu-darmstadt.de](mailto:ivan.habernal@tu-darmstadt.de)).

After a manual check of the requester’s credibility by Dr. Ivan Habernal, the dataset will be shared by e-mail to the applicant.

All requests can be transparently shared with the book author and the publisher, if desired.

## 

## Request form

The dataset is available under the following conditions:

1. The dataset is to be used for non-commercial purposes only.
2. All publications on research based on the dataset should give credit to the author (Joseph Glannon) and the publisher (Aspen Publishing).
3. No part of the dataset may be shared with third parties. The dataset may only be used by the person who agrees to the terms of the license.
4. The dataset must be deleted after finishing the experiments.

For access to the dataset accept the above conditions and fill out the following form.

Fields marked with an \* are required

|  |  |
| --- | --- |
| Name\* |  |
| E-mail\* |  |
| Institution\* |  |
| Department/Research center\* |  |
| Address\* |  |

I agree to the above-mentioned conditions\*: Yes/No