Annotators-in-the-loop: Testing a Novel Annotation Procedure on Italian Case Law

Emma Zanoli[1], **Matilde Barbini**[1], **Davide Riva**[2], **Sergio Picascia**[2], **Emanuela Furiosi**[3], **Stefano D'Ancona**[3], **and Cristiano Chesi**[1]

[1]IUSS Pavia - School for Advanced Studies, NETS Lab
[2]Università degli Studi di Milano, Department of Computer Science
[3]IUSS Pavia - School for Advanced Studies
{emma.zanoli, matilde.barbini, emanuela.furiosi, stefano.dancona, cristiano.chesi}@iusspavia.it
{sergio.picascia, davide.riva1}@unimi.it

Abstract

The availability of annotated legal corpora is crucial for a number of tasks, such as legal search, legal information retrieval, and predictive justice. Annotation is mostly assumed to be a straightforward task: as long as the annotation scheme is well defined and the guidelines are clear, annotators are expected to agree on the labels. This is not always the case, especially in legal annotation, which can be extremely difficult even for expert annotators. We propose a legal annotation procedure that takes into account annotator certainty and improves it through negotiation. We also collect annotator feedback and show that our approach contributes to a positive annotation environment. Our work invites reflection on often neglected ethical concerns regarding legal annotation.

1 Introduction

Despite the success of self-supervised deep learning approaches (Jaiswal et al., 2021), accurate human annotation remains essential for NLP research, and it is no different for its applications to the legal domain. The increasing availability of corpora of legal documents has given a tremendous boost to legal NLP (Zhong et al., 2020), but this comes with serious ethical implications given the potential uses of systems trained on the annotated data (see Tsarapatsanis and Aletras (2021) for a brief overview). Legal annotation is a complex task, where even expert annotators may fail to come to straightforward conclusions (Wyner et al., 2013). This warrants particular reflection on the definition of legal annotation guidelines and on making sure they are appropriate and consistently agreed upon among annotators (Santosuosso and Pinotti, 2020).

To address the aforementioned issues, we present an annotation procedure that involves a group of legal experts in the very process of creating and negotiating the annotation guidelines. We also anonymously collect annotators' feedback and show that our procedure makes them more certain of and satisfied with their work. We believe this to be an important step towards a better treatment of annotators in the field of legal NLP.

The Italian legal system is currently undergoing significant changes in an effort to digitally transform and overall improve legal processes at all levels. At this stage, gathering high quality data is crucial to make sure that any downstream applications do not perpetuate errors and biases. The annotation procedure we describe is a preliminary step in the framework of the *Next Generation UPP* (*NGUPP*) project, funded by the Italian Ministry of Justice, and aimed at improving the efficiency of the judicial system in Italy. Specifically, we intend to empower judges with advanced information management tools to facilitate the drafting of court judgements. Such a tool would be used both retroactively, for legal search of case law, and proactively, for the creation of new judgments.

The paper is organized as follows. In Section 2 we discuss the relevant literature. In Section 3 we present the experimental design. Section 4 presents the annotation procedure. Section 5 is dedicated to the discussion of the results. Finally, in Section 6 we provide concluding remarks and ideas for future developments.

2 Related work

Corpora of legal texts are increasingly available and accessible. This is especially true of legislation (Chalkidis et al., 2019; Váradi et al., 2020), but it also applies to court judgments (Grover et al., 2004; Poudyal et al., 2020; Feng et al., 2022; Kapoor et al., 2022) and other types of legal texts (e.g. con-

The paper was jointly conceived by the authors. However, Section 4.1 was written by Emanuela Furiosi and Section 4.2 was written by Stefano D'Ancona.

Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII), pages 118–128
July 13, 2023 ©2023 Association for Computational Linguistics

tracts, Funaki et al., 2020). There have already been several annotation efforts in the legal domain (Wyner, 2010; Duan et al., 2019; Glaser et al., 2021a; Kalamkar et al., 2022), with a particular interest towards arguments (see Zhang et al., 2022 for an overview).

While some types of annotation are relatively straightforward, obtaining consistent and accurate annotations in law is extremely challenging (Walker, 2016). Nonetheless, legal annotation tasks often leverage law students as domain experts (Wyner et al., 2013; Chalkidis et al., 2017; Soavi et al., 2022; Correia et al., 2022; Kalamkar et al., 2022). We invite caution in using this approach due to a) ethical concerns on adequate annotator compensation and b) difficulty in ascertaining their domain expertise.

Legal annotation tasks may entail another potentially problematic aspect. It is not uncommon to involve a small group of annotators who initially annotate the same text, which is subsequently revised by a more expert annotator tasked with solving any discrepancies (Wyner et al., 2013; Poudyal et al., 2020; Galli et al., 2022). Although this is a widely accepted method used to obtain gold standard annotations in the legal domain, we will not be using this technique; rather, we embrace the line of research that sees variation in human annotation as something that may naturally arise due to, e.g., ambiguity, uncertainty of the annotator, genuine disagreement, or simply the fact that multiple options are correct (Plank, 2022). Specifically, we follow Basile et al. (2021), who argue that "removing the disagreement might lead to better evaluation scores, but it fundamentally hides the true nature of the task we are trying to solve".

To address the aforementioned issues, we propose an annotation procedure that promotes guideline negotiation. Previous work on legal annotation has featured modifications of annotation guidelines over time, either in a top-down manner or within small groups (Teruel et al., 2018; Correia et al., 2022; Galli et al., 2022). Lee et al. (2022) experiment with collaborative guideline creation among pairs of annotators, albeit not in the legal domain. They show that negotiation leads to improved annotator agreement within the pair, but the performance decreases dramatically among annotators of different pairs. Our group of annotators was not split into pairs for the negotiation; we are not aware of previous work that frames legal annotation as a peer-to-peer negotiation process among an entire group of legal professionals.

In an effort to contribute to a positive annotation environment, we collect feedback from the annotators. Following Nedoluzhko and Mírovský (2013) and Andresen et al. (2020), we collect measures of annotator certainty, checking whether they improve after the negotiation process. We also collect data on the overall satisfaction of the annotators.

The dataset we obtain from our annotation will be used for the development of text segmentation models. Segmenting court judgments into relevant sections can improve legal search and information retrieval; this has already been investigated by Savelka and Ashley (2018), Aumiller et al. (2021) and Glaser et al. (2021b). Licari and Comandè (2022) segment Italian civil judgements with simple regular expressions for bench-marking purposes.

We operate in the Italian legal context, which has been amply explored in previous literature (Lenci et al., 2009; Venturi, 2013; Tagarelli and Simeri, 2021; Galli et al., 2022). However, our proposed annotation procedure is language-agnostic.

3 Experimental design

This section describes our experimental design, aimed at developing an annotation procedure for the legal domain. We briefly present the dataset, the task, the annotators, the annotation tool, and the agreement metrics.

3.1 Dataset

The dataset consists of 50 Italian case law judgments, retrieved from 12 different Courts. The documents all concern first degree civil law judgments regarding the matter of unfair competition.

The selected case law judgments were available in PDF files, from which text was extracted using the Python implementation of `MuPDF`, an open source software framework for viewing and converting PDFs. The documents are very heterogeneous in terms of length: the number of tokens ranges from 1,368 for the smallest document, to more than 8,000 for the largest one, with a mean length of 4,387 tokens and a standard deviation of 1,798.

3.2 Annotation task

Given the collection of documents described in 3.1, the annotators were required to perform a "struc-

tural annotation", i.e. to recognize the distinct sections that compose the structure of a court judgement. Thus, the task was to identify sections and sub-sections (text segmentation) and to label those segments (segment labeling).

The annotators were presented with text in free form; they had to underline the segments of interest and assign a label to them choosing from a predefined set. This set of labels (called the "annotation scheme") and its development are described in depth in Section 4. The annotation is aimed at creating a dataset for legal text segmentation. The general expectation was that the entire text would be segmented and labeled (i.e. with no gaps between different sections), although this was not made explicit in the annotation guidelines.

The annotators were given the possibility to review and change their own annotations over time, provided that such modifications were made independently from other annotators.

3.3 Annotators

The annotation task was carried out by 9 law professionals, all of whom have relevant experience as both academics and practitioners: all but one of them hold a PhD and they are all licensed lawyers, having passed the Italian bar exam. While seniority varies on an individual basis (years of professional experience: min 3, max 23), they all have significant expertise in either civil law (4 annotators), criminal law (1 annotator), or a mix of different areas (4 annotators). As such, they are all familiar with Italian legal language and did not require ad hoc linguistic training. However, none of them had ever annotated before. For this reason, three law professionals with prior annotation experience were consulted for an initial draft of the annotation guidelines, but they did not perform the actual annotation task.

The annotators were asked to fill out a feedback questionnaire after the annotation process (see 5.4).

3.4 Annotation tool

Technical constraints and privacy issues prompted us to use proprietary annotation software. We use the Ellogon language engineering platform (Ntogramatzis et al., 2022), since it supports the task as defined in 3.2. The platform had to be customized to introduce the annotation labels of interest.

3.5 Agreement metrics

We evaluate agreement among annotators (Inter-Annotator Agreement, IAA) in order to provide a quantitative assessment of (1) the complexity of the annotation task, and (2) the homogeneity of the results. The annotation was carried out and analysed in the absence of a gold standard; we consider appropriate annotations to be an incrementally realised goal rather than a given (see also 5.3).

IAA has to account for 3 factors: a) presence of labels; b) alignment of annotated segments; c) agreement of labels assigned to segments. In order to cover all of these characteristics, we employ the γ coefficient (Mathet et al., 2015). It is computed as the average of all local disagreements, referred to as disorders, between units from different annotators:

$$\forall s \in c, \gamma = 1 - \frac{\delta(s)}{\delta_e(c)} \tag{1}$$

with $\delta(s)$ being the disorder of the annotation set s and $\delta_e(c)$ being the expected disorder of the corpus c. Maximum agreement is represented by $\gamma = 1$, while $\gamma < 0$ corresponds to the worst case, where annotator agreement is worse than annotating at random. Following this methodology, units of annotation are aligned to minimize the overall disorder. We compute γ scores not only for each document, but also for each label defined in the annotation scheme in order to identify the most and the least disputed structural segments.

Finally, since annotators had the possibility of going back to the documents assigned to them and review their own annotations, we store periodic dumps of the annotation database and estimate self-agreement, i.e. the extent to which annotators maintain the segments and labels they had already selected. To this aim, we introduce the metric δ, calculated as:

$$\delta = \frac{1}{T}\sum_{t=2}^{T}\frac{1}{|K|}\sum_{k\in K}\frac{|S_k^{(t-1)} \cap S_k^{(t)}|}{|S_k^{(t-1)}|} \tag{2}$$

where T is the total number of periodic dumps of the annotation database, K is the label set, and S_k^τ is the set of segments labelled with k at time τ. Notice that δ takes into account only the intersection of segment sets at consecutive times, and $\delta \in [0; 1]$.

4 Annotation scheme

In this section we summarize the development of the annotation scheme: first, we describe the initial scheme as designed by a restricted pool of experts; then, we recount its subsequent negotiation; finally, we report the resulting annotation scheme.

4.1 Initial development of the annotation scheme

The initial structural annotation scheme was developed through a reflection carried out by a small group of legal experts with specific and complementary expertise in both legal practice and in the digitization of justice. Specifically, these were: a university professor, former judge at the Court of Appeal of Milan; two legal professionals with previous annotation experience; and a researcher who also has around 7 years of experience as a lawyer and who is among the 9 annotators who carried out the annotation task.

The annotation experts involved had previously worked on complex structural annotation schemes. By contrast, it was unanimously decided to keep the structural annotation scheme simple, for two reasons. First, the structural segmentation was, at least initially, primarily aimed at distinguishing the reasoning part of the judgment from the other sections. Second, the more basic structural analysis was to be complemented and enriched by a further layer of more detailed argumentative annotation.

The initial annotation scheme featured 5 sections, specifically:

- the section **"Corte e parti"** included the indication of the **court**, the panel of **judges**, and the **parties** in the trial (i.e., the plaintiffs, the defendants, and any intervening third parties);
- the section **"Antefatto"** included **background information**, specifically on a) the proceedings of the trial, and b) the reconstruction of the facts involved in the case;
- the section **"Domande"** identified the **claims and arguments** brought forward by the parties (i.e., claims made by the plaintiff(s) and any counterclaims made by the defendant(s)). Each claim would be labelled individually;
- the section **"Motivazione"** identified the part of the judgment in which the **reasoning** for the decision of an individual claim is explained. Each line of reasoning would be labelled individually;
- the section **"Decisione"** identified the **final decision(s)** on each individual claim. If there are multiple decisions, each would be labelled individually.

In the presence of multiple claims, lines of reasoning and decisions, they would be numbered to link the three elements to one another.

The content of Italian court judgments is regulated by Article 132, c. 2 of the Italian Civil Procedure Code (CPC), which stipulates that each judgment *"must contain: 1) an indication of the judge who pronounced it; 2) an indication of the parties and their attorneys; 3) the conclusions of the prosecutor and those of the parties; 4) a concise statement of the reasons of fact and law of the judgment; 5) the ruling, the date of the deliberation and the signature of the judge."* Nonetheless, the exact outline and structure of the judgments may vary in practice (e.g. some judges may wish to add section headings to structure their decisions, while others may not; some may provide this information into clearly separated sections, while others may not; etc.). The initial annotation scheme was thus developed taking into consideration not only the structure of the judgments as currently regulated by the CPC, but also as applied in practice by judges.

As one can see, the sections of this initial scheme, while encompassing the essential elements of the judgment outlined in the CPC, are not exactly overlapping. Specifically, the contents of items (1) and (2) are grouped in the "Corte e parti" section; the contents of item (3) can be found in the "Domande" section, the contents of item (4) correspond to the "Motivazione" section, and the contents of item (5) correspond to the "Decisione" section. Additionally, the annotation scheme includes the "Antefatto" section[1]. Previous experimentation with legal search models revealed that they would sometimes retrieve judgments based on content which was presented as background information in the case, even when the expected outcome would relate to the reasoning section. This segmentation was thus meant to aid the models in excluding potentially irrelevant information by focusing on specific sections.

[1]This element was actually mandatory in a previous version of the CPC, but it has not been since 2009; in practice, a lot of judges still use it.

Please note that, at the time this annotation scheme was devised, the technical specifics of how the annotation would be carried out had not yet been defined.

4.2 Negotiation of the annotation scheme

The initial annotation scheme was modified through three meetings involving the entire group of annotators. The need for discussion and negotiation first became evident upon starting to apply the initial annotation guidelines within the constraints of the provided annotation tool. Specifically, there was an interest in mapping the overarching structural relationships between claims, reasoning and final decisions.

It was decided that individual claims, lines of reasoning and decisions would be considered subsections of more broadly defined sections. Furthermore, it was noted that the annotations of sub-sections could benefit from the definition of "chains" of reasoning, practically consisting of pairwise relationships between a claim, the reasoning on it, and the corresponding final decision.

After extensive discussion, it was further specified in the guidelines that the aforementioned "chains" should simply reflect lines of reasoning, without specifications on the nature of the reasoning itself (e.g. premise vs. conclusion, support vs. contrast). It was concluded that these would be left for a further layer of argumentative annotation, to be performed at a later time. This integration of the guidelines was considered necessary to prevent annotators from labeling text segments based on an "argumentative" and not a "structural" evaluation of their content.

Another point that required a collaborative discussion was related to the distinction between reasoning and decision. As previously mentioned, Italian court judgements are required to feature a specific section, at the very end, where the main decisions of the case are summarized: it is the so-called "Dispositivo" (final ruling), typically placed after the heading PQM, which translates to "For These Reasons". However, judges often "anticipate" their own decision within the body of the reasoning, as it may come naturally to conclude a given line of thought. The annotators thus concluded that within the "reasoning" section there could be "decision" sub-sections attributed to specific text segments.

4.3 The resulting annotation scheme

As a result of the collaborative (re)negotiation of earlier annotation schemes, the annotators came to agree on a set of guidelines, which were then used to annotate the dataset. We call these guidelines the "resulting annotation scheme", summarized in Table 1.

This annotation scheme is meant to segment Italian court judgements of civil proceedings at two levels: sections and sub-sections. The sections correspond to the ones presented in 4.1. Sub-sections are possible for the last three sections. These are meant to distinguish between different claims (e.g. <dom1>, <dom2>), different lines of reasoning (e.g. <mot1>, <mot2>), and different decisions (e.g. <dec1>, <dec2>). The sub-sections can be put in relationships of the type (<dom>,<mot>) or (<mot>,<dec>) if a motivation for decision <dec> on claim <dom> is explicit in the document, otherwise a (<dom>,<dec>) relation could be specified.

5 Analysis and discussion of the results

The results of our work include the annotation scheme as well as the output of the annotation activity. We evaluate Inter-Annotator Agreement from both a quantitative and qualitative perspective and we report annotator feedback.

5.1 Appraising the resulting annotation scheme

Given the somewhat unusual nature of our procedure, does the resulting annotation scheme reflect what we might expect?

Considering the provisions of the Italian CPC (see 4.1), it is not surprising that a similar 5-part subdivision can be found in other works on Italian legal NLP (Galli et al., 2022; Licari and Comandè, 2022). Contrary to what one may expect, though, Italian judgments often do not conform to a strict standard, with some sections (<ANT> and <MOTSEZ>) being presented in different orders or not being clearly distinguished from one another. Text segmentation of Italian judgements is therefore not a trivial task, which motivates the need for text segmentation models to be carefully evaluated.

The scheme is also comparable to other works in the literature that have, within a variety of legal contexts, outlined a structural segmentation of court judgements (see e.g. Wyner et al., 2013 for the UK, Poudyal et al., 2020 for the European Court of Human Rights, Glaser et al., 2021b for Germany).

Sections	Sub-sections	Italian	Explanation
<COR>		Corte e parti	Court, judicial panel, parties
<ANT>		Antefatto	Background information
<DOMSEZ>	<dom1>,<dom2>,...	Domande	Claim(s) and argumentation of the parties
<MOTSEZ>	<mot1>,<mot2>,...	Motivazione	Reason(s) for the final decision(s)
<DECSEZ>	<dec1>,<dec2>,...	Decisione	Final decision(s)

Table 1: The resulting annotation scheme for Italian court judgements of civil proceedings.

5.2 Annotator agreement

We report the results obtained for the metrics introduced in 3.5.

Before computing the agreement metrics, some cleaning operations were applied to the section annotation results. In particular, since sections are meant to be as contiguous as possible, quasi-consecutive segments with the same label were merged into a single segment. For practical purposes, segments within a distance of 25 characters from one another were considered quasi-consecutive. Duplicates and quasi-duplicates (i.e. segments that share at least 90% of another segment underlined later) were deleted, since they likely result from technical difficulties with the annotation tool. Finally, documents with partial annotations (i.e. with segments labelled with less than half of the labels in the annotation scheme) are not considered in the agreement evaluation. This is motivated by the expectation that each document contains at least one segment fulfilling each function, and does not undermine the results, resulting in the exclusion of only 3 documents for the section annotation and 6 documents for the subsection annotation.

Table 2 reports the γ score statistics for both sections and sub-sections. High standard deviation suggests that some documents were more complex to annotate than others.

	Segments per doc.	Mean γ	Std.Dev. γ	Max. γ
Sections	9.92	0.635	0.225	0.996
Sub-sections	13.79	0.483	0.260	0.995

Table 2: Average per-document number of annotated segments and γ score statistics over retained documents. Notice $\gamma < 0$ on a document if the annotation agreement is worse than the null case of random annotations, whereas $\gamma = 1$ on a document if annotations perfectly agree.

As the table shows, the number of labeled segments in each document exceeds the cardinality of the label set, which indicates that the sections tend to be discontinuous and sparse inside the document. Indeed, Figure 1 shows that both in a document with well-aligned and in another document with poorly aligned annotations, some sections are interrupted by others and re-appear later in the text.

Figure 2 shows the confidence intervals of γ scores for each section type, indicating that some sections are more difficult to locate than others. While the <COR> section is usually located at the beginning and is therefore widely agreed upon, the location of the <ANT> section varies depending on the judge and the specific case. Agreement on the <DECSEZ> section is among the lowest. As discussed in 4.2, although the final decisions typically conclude the judgement, anticipations of the decisions can be found in previous sections, leading to interpretative differences as to what constitutes a final decision. Although we do not have a baseline we can compare our results against, our findings are consistent with those reported by Wyner et al. (2013).

To gain a deeper understanding of the causes for disagreement, we calculated how frequent it was for the annotators to label the same segment differently, i.e. the categorial dissimilarity d_{cat} between aligned annotators units. As expected, the label pairs (a, b) that showed the highest disagreement, i.e. the highest number of segments that were annotated with a by one annotator and with b by another annotator, were (<ANT>, <DOMSEZ>) and (<MOTSEZ>, <DECSEZ>).

Figure 3 shows the confidence intervals for γ scores for each subsection type. Agreement drops significantly for these more fine-grained labels. <dec> segments are the ones that raise the highest disagreement, while segments of the other two types are comparable in terms of agreement. The higher numerosity of <dec> segments likely plays a role in their higher variability.

To calculate the metric we introduce, namely self-agreement over time, we made 4 dumps of the database, one before each negotiation meeting

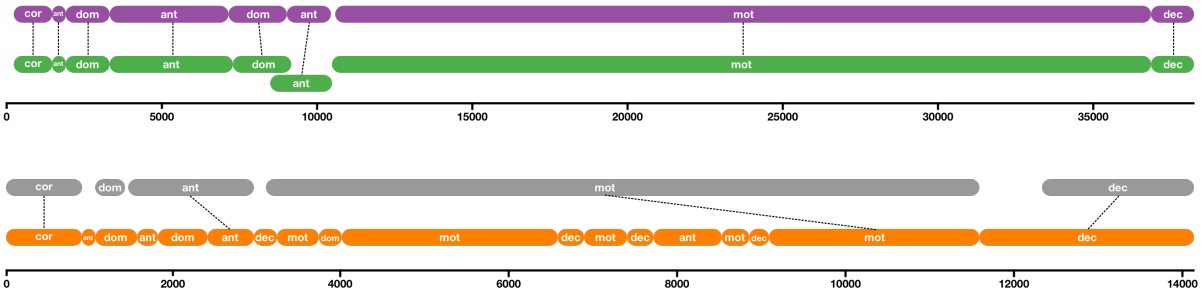

Figure 1: Example of alignment (top image) and misalignment (bottom image) of segments selected and labelled by two annotators for two documents. The horizontal axis represents the position of characters in the document.

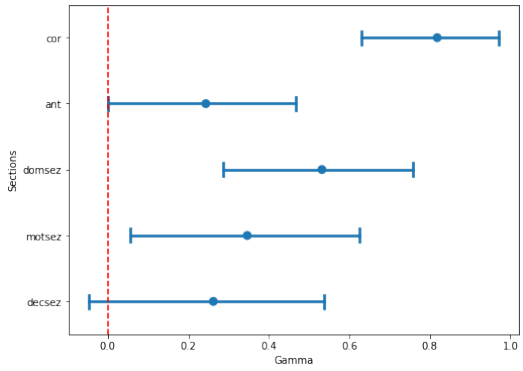

Figure 2: γ score mean and 95% confidence interval for each section label.

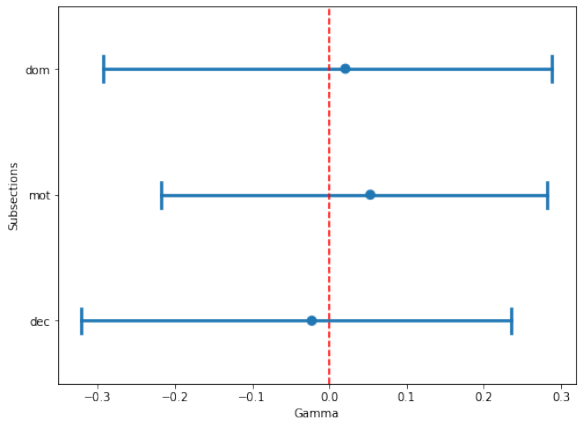

Figure 3: γ score mean and 95% confidence interval for each subsection label.

and one at the end of the annotation process. We found an average δ of 0.963, with a standard deviation of 0.151, which indicates that few changes were made to annotations over time. In particular, the mean δ reveals that few changes occurred as a consequence of the meetings, but its standard deviation suggests some annotators made much more extensive modifications than others.

5.3 Qualitative analysis of annotator disagreement

While agreement metrics are important in the evaluation of annotation, the investigation of disagreement can reveal important considerations which can greatly improve the annotation process (Lee et al., 2022; Plank, 2022). This section presents a brief but illustrative qualitative analysis of some outputs of the annotation: the aim is to highlight where the agreement between the annotators proved to be weak, leading us to reflect on what might be the primary causes of the disagreements.

From a legal standpoint, unfair competition is a rather complex matter and the judgments tend to exhibit a convoluted structure, with the judges having to address a large number of claims brought forward by the parties. This complexity is certainly a challenge for the annotators, who need to deduce and combine non-trivial information to arrive at the label (Malik et al., 2022). As reported in 5.2, the label pairs that exhibited a higher level of disagreement between annotators were (<ANT>, <DOMSEZ>) and (<MOTSEZ>, <DECSEZ>). We now review an example for each label pair.

Background information may be presented and evaluated throughout the entire judgement; an annotator might therefore be uncertain as to which label to apply. For example, the facts of the case can contribute to the argumentation of the reasoning section (see Figure 6 in the Appendix). Additionally, the judge may reference the claims of the parties in their summary of the facts (see Figure 4). Given the ambiguity, Annotator 1 (left) decided to remark the presence of the claims (<DOMSEZ>, in green), while Annotator 2 (right) chose to label the entire section as background information (<ANT>, in purple).

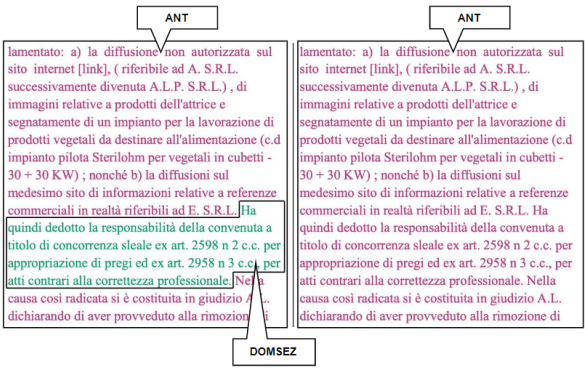

Figure 4: Excerpt showing disagreement between two annotators (<ANT> - purple, <DOMSEZ> - green).

In addition to the inherent difficulty of the subject matter, there is potential ambiguity in the annotation guidelines: as can be seen from Figure 5, Annotator 1 (left) identified parts of the decision (in orange) also within the section containing the legal reasoning (in blue), whereas Annotator 2 (right) labeled the entire segment as legal reasoning (see Figure 7 in the Appendix for another example). Although the negotiation meetings featured extensive discussion on the use of the <DECSEZ> and <dec> labels, some ambiguity remains, leading annotators to different interpretations.

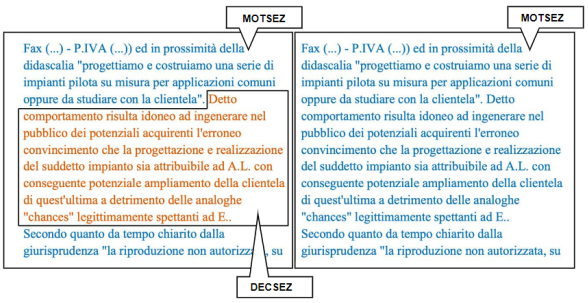

Figure 5: Excerpt showing disagreement between two annotators (<MOTSEZ> - blue, <DECSEZ> - orange).

It is evident that <MOTSEZ> is a complex section whose content interacts with other sections through complex textual realizations; as such, it is difficult to annotate in an unanimous fashion. Let us reiterate that we do not intend to conflate this complexity into an aggregated "ground truth"; rather, we are actively experimenting with methods that can capture and appreciate interpretative differences.

5.4 Annotator feedback

As we have extensively discussed, the annotators were encouraged to (re)negotiate the annotation scheme and guidelines over several meetings. Given the difficulty of legal annotation, we believe this to be crucial in making annotators feel supported. Not only did we believe that this process would improve annotator certainty (Nedoluzhko and Mírovský, 2013; Andresen et al., 2020), but we also hoped it would help annotators be satisfied with their work. To measure this, we asked the annotators to fill out a questionnaire to provide anonymous feedback on the annotation process. Based on the feedback we gathered, it appears that annotator certainty increased slightly after the meetings (35%). Additionally, all respondents but one[2]: a) express satisfaction with the work they have done; b) report that the meetings facilitated a more thorough comprehension of the annotation process; c) indicate that the meetings were instrumental in revisiting guidelines that were not sufficiently clear or appropriate.

6 Conclusion and Future Work

This paper introduces a novel annotation procedure based on the active participation of an entire group of domain experts in the process of creating and negotiating the annotation guidelines. An interdisciplinary research team, involving experts from the legal, linguistic and computer science fields, has actively collaborated in order to address the common issues faced in the annotation of legal documents. The result of this procedure is an annotation scheme tailored to Italian case law judgments, which provides a unifying structure to integrate the sections mandated by the law and the ones used in practice by judges. We consider these to be preliminary results in the ongoing development of a reliable procedure that will be extended in future work. We are currently experimenting with the annotation of more fine-grained phenomena: the structure outlined by our annotation scheme serves as the basis for the annotation of legal arguments. Since the work presented here is still ongoing, we are unable to release the annotated dataset and the annotation guidelines at present; however, the annotation scheme is presented in Table 1 and its development is documented in Section 4.

Our project comes at a crucial time in the process of re-thinking how the judicial system works

[2]In the questionnaire these were presented as statements that the annotators could either agree, partially agree, or disagree with. The same individual disagreed with all of them; regrettably, since the feedback is anonymous, we can not reach out to them directly to understand what may have gone wrong.

in Italy. The work of law professionals is changing due to the introduction of increasingly sophisticated technological tools. The annotations we collect will be used to build corpora that represent the structure and argumentation of Italian court judgments. Leveraging segmented case law judgments can improve both keyword-based and semantic-based search of legal precedents. We are actively experimenting with different techniques, including few-shot learning, that can leverage this data to improve the efficiency of legal search. The long-term goal is to integrate these tools into a document builder that supports Italian judges in the drafting of court judgments.

The annotation of a small set of 50 judgments was used to elaborate, apply and evaluate a novel annotation procedure, capable of taking into account the nuances that the legal subject matter brings, especially when applied to complex cases, while also allowing domain experts to be adequately valued in their specific expertise. Discussions on the ethics of legal NLP abound, with emphasis on the data and its potentially harmful uses. While crucial, these discussions would benefit from further reflection on how the data is being annotated. We hope that our results can inspire researchers and practitioners to carefully consider these issues in future work.

Ethics Statement

Our work is meant to inspire reflection on the treatment of annotators in the field of legal NLP. Specifically: a) we make it a point of involving legal professionals, not law students; b) the annotators involved in the project won a public selection competition to participate in a project aimed at the digitalization of the Italian judicial system; c) the annotators are all hired to work on the project and receive adequate pay; d) we make sure that their specific expertise is valued by involving them in the creation and negotiation of the annotation guidelines; e) we take measures to track whether they are happy with the work they are doing.

Limitations

Although our annotation procedure envisions a negotiation process among an entire group of legal experts, due to time constraints each document was eventually annotated by either 2 or 3 annotators. Having the entire group annotate every document might have yielded more interesting and fruitful discussions for the negotiation process and allowed for a deeper analysis of annotator (dis)agreement. We also have to point out that several annotators lamented technical difficulties in using the annotation tool (due to the limitations of the tool itself); this may have severely impacted annotation quality. We wish to address these limitations in future work.

Acknowledgements

This work is supported by the Next Generation UPP project (May 2022 - September 2023) within the PON program of the Italian Ministry of Justice. We are thankful to Amedeo Santosuosso for generously sharing his valuable expertise to the benefit of this project: he has been a long-time advocate for the creation of a document builder that supports Italian judges in the drafting of court judgments; he also contributed to the initial draft of the annotation guidelines presented here. We would also like to thank the anonymous reviewers for their insightful comments.

References

Melanie Andresen, Michael Vauth, and Heike Zinsmeister. 2020. Modeling ambiguity with many annotators and self-assessments of annotator certainty. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 48–59, Barcelona, Spain. Association for Computational Linguistics.

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 2–11, New York, NY, USA. Association for Computing Machinery.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. ACM.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy. Association for Computational Linguistics.

Fernando A Correia, Alexandre AA Almeida, José Luiz Nunes, Kaline G Santos, Ivar A Hartmann, Felipe A Silva, and Hélio Lopes. 2022. Fine-grained legal entity annotation: A case study on the brazilian supreme court. *Information Processing & Management*, 59(1):102794.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, Heng Wang, and Zhiyuan Liu. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. In *Chinese Computational Linguistics*, pages 439–451, Cham. Springer International Publishing.

Yi Feng, Chuanyi Li, and Vincent Ng. 2022. Legal judgment prediction: A survey of the state of the art. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5461–5469.

Ruka Funaki, Yusuke Nagata, Kohei Suenaga, and Shinsuke Mori. 2020. A contract corpus for recognizing rights and obligations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2045–2053, Marseille, France. European Language Resources Association.

Federico Galli, Giulia Grundler, Alessia Fidelangeli, Andrea Galassi, Francesca Lagioia, Elena Palmieri, Federico Ruggeri, Giovanni Sartor, and Paolo Torroni. 2022. Predicting outcomes of italian vat decisions 1. In *Legal Knowledge and Information Systems*, pages 188–193. IOS Press.

Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021a. Generation of legal norm chains: Extracting the most relevant norms from court rulings. In *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021b. Improving legal information retrieval: Metadata extraction and segmentation of german court rulings. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. SCITEPRESS - Science and Technology Publications.

Claire Grover, Ben Hachey, and Ian Hughson. 2004. The HOLJ corpus. supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54, Geneva, Switzerland. COLING.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies*, 9(1).

Prathamesh Kalamkar, Aman Tiwari, Astha Agarwal, Saurabh Karn, Smita Gupta, Vivek Raghavan, and Ashutosh Modi. 2022. Corpus for automatic structuring of legal documents. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4420–4429, Marseille, France. European Language Resources Association.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun T H, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. HLDC: Hindi legal documents corpus. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3521–3536, Dublin, Ireland. Association for Computational Linguistics.

Seunggun Lee, Alexandra DeLucia, Ryan Guan, Rubing Li, Nikita Nangia, Shalaka Vaidya, Lining Zhang, Zijun Yuan, Praneeth Ganedi, Britney Ngaw, et al. 2022. Common law annotations: Investigating the stability of dialog annotations. In *The Tenth AAAI Conference on Human Computation and Crowdsourcing*. Association for the Advancement of Artificial Intelligence.

Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, and Giulia Venturi. 2009. Ontology learning from italian legal texts. In *Law, Ontologies and the Semantic Web*, pages 75–94. IOS Press.

Daniele Licari and Giovanni Comandè. 2022. Italian-legal-bert: A pre-trained transformer language model for italian law. In *EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*.

Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya, and Ashutosh Modi. 2022. Semantic segmentation of legal documents via rhetorical roles. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 153–171, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Anna Nedoluzhko and Jiří Mírovský. 2013. Annotators' certainty and disagreements in coreference and bridging annotation in Prague dependency treebank. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 236–243, Prague, Czech Republic. Charles University in Prague, Matfyzpress, Prague, Czech Republic.

Alexandros Fotios Ntogramatzis, Anna Gradou, Georgios Petasis, and Marko Kokol. 2022. The ellogon web annotation tool: Annotating moral values and arguments. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3442–3450, Marseille, France. European Language Resources Association.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and

evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Amedeo Santosuosso and Giulia Pinotti. 2020. Bottleneck or crossroad? problems of legal sources annotation and some theoretical thoughts. *Stats*, 3(3):376–395.

Jaromir Savelka and Kevin D Ashley. 2018. Segmenting us court decisions into functional and issue specific parts. In *JURIX*, pages 111–120.

Michele Soavi, Nicola Zeni, John Mylopoulos, and Luisa Mich. 2022. Semantic annotation of legal contracts with ContrattoA. *Informatics*, 9(4):72.

Andrea Tagarelli and Andrea Simeri. 2021. Unsupervised law article mining based on deep pre-trained language representation models with application to the italian civil code. *Artificial Intelligence and Law*, pages 1–57.

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. The MARCELL legislative corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. European Language Resources Association.

Giulia Venturi. 2013. Investigating legal language peculiarities across different types of italian legal texts: an nlp-based approach. In *Bridging the Gap (s) between Language and the Law. Proceedings of the 3rd European Conference of the International Association of Forensic Linguistics*, pages 138–156.

Vern R Walker. 2016. The need for annotated corpora from legal documents, and for (human) protocols for creating them: the attribution problem.

Adam Z Wyner. 2010. Towards annotating and extracting textual legal case elements. *Informatica e Diritto: special issue on legal ontologies and artificial intelligent techniques*, 19(1-2):9–18.

Adam Z Wyner, Wim Peters, and Daniel Katz. 2013. A case study on legal case annotation. In *JURIX*, pages 165–174.

Gechuan Zhang, Paul Nulty, and David Lillis. 2022. A decade of legal argumentation mining: Datasets and approaches. In *Natural Language Processing and Information Systems*, pages 240–252, Cham. Springer International Publishing.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Appendix: additional examples

Additional examples of annotator disagreement, discussed in 5.3.

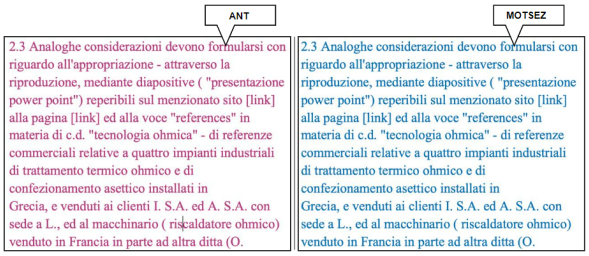

Figure 6: Excerpt showing disagreement between two annotators (<ANT> - purple, <MOTSEZ> - blue)

.

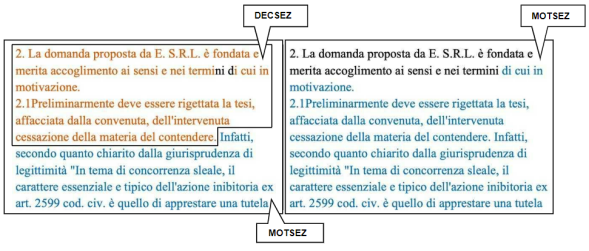

Figure 7: Excerpt showing disagreement between two annotators (<MOTSEZ> - blue, <DECSEZ> - orange). The unlabeled segments (in black) are an example of the quasi-consecutive segments referenced in 5.2, which were likely caused by technical difficulties with the annotation tool.