Part Two: Formalism in Theory, on the Ground and in Between (What Formalism Is, How It Manifests and How to Catch It in the Wild)

This chapter develops a framework for empirical investigation of claims about legal reasoning in the CEE. It focuses on distinguishing and measuring formalistic and non-formalistic judicial reasoning. Based on a thorough literature review, I first identify five core tenets of formalism as described in the literature on the CEE formalism. I then establish empirical indicators that bridge these theoretical features with real world judicial decisions, revealing how formalism manifests in practice. For instance, I identify the presence of text-based argumentation as a core tenet formalism and then, based on the legal argumentation theories, describe that the notion "text-based argument" covers, e.g., linguistic interpretation which appears, e.g., when a court explicitly refers to ordinary meaning, syntax or a dictionary.

Afterwards, I suggest a twofold measurement approach to formalism: analyzing the frequency of formalistic arguments and evaluating decisions holistically as formalistic/non-formalistic.[26]

This original methodology results in a novel annotation scheme, completed with detailed guidelines and flow charts for analyzing judicial decisions. The framework addresses limitations of previous studies on formalism in the CEE. And most importantly, it lays groundwork for the quantitative analysis and its results (Chapter 3).

[26] I am thankful to Ivan Habernal for this idea of including holistic assessment.

2.1 What CEE Formalism Means and How it Manifests

This study empirically examines the anti-formalist narrative using content analysis, a core method in empirical legal research (Hall and Wright, 2006; Ovádek et al., 2025).[27]

Detemermining whether a court is formalistic requires the following:

1. **Defining Formalism:** Establish a clear definition of what formalism is and, generally, how it could manifest.
2. **Creating Annotation Guidelines:** As Ovádek et al. (2025, p. 8) emphasize, it is essential to identify indicators that connect abstract concepts (e.g. formalism) to observable facts (e.g. judicial decisions). Guidelines require detailed criteria to guide the annotation process, ensuring consistency and reliability. The aim of our guidelines was to to very concretely describe (incl. real world examples) how formalism and different argumentation practices manifest in judicial decisions.
3. **Training Annotators:** Equip annotators with the skills to identify and classify relevant arguments, including what to look for in court decisions.
4. **Data Collection and Sampling:** Gather a representative dataset of court decisions through a sampling procedure.
5. **Annotation of Decisions:** Apply the guidelines to annotate the collected decisions.
6. **Analysis:** Interpret the annotated data to assess the prevalence and nature of formalism in judicial reasoning.

2.1.1 Defining CEE Formalism: Five Core Tenets

First comes the definition of formalism. When scholars claim CEE judges are formalistic, what precisely do they mean? Is "formalism"

[27] Ovádek et al. call the elaborated version of content analysis I pursue in this study expert coding. They describe similar steps as I do. See Ovádek et al., 2025.

more than an insult?[28] Does it carry any descriptive content that can be empirically verified? Formalism concerns courts and their decision-making, but how does formalistic judgment typically look like?

Definitions of formalism vary. This study verifies claims about CEE judicial practices and thus tries to investigate formalism as described by the CEE legal scholarship. Although there is no "Bible of CEE formalism", current literature reveals some consistency in identifying five distinctive features of formalistic judicial decision-making. I call them *five core tenets of CEE formalism*:

Formalistic courts in the CEE would

1. rely heavily on a limited set of arguments derived from statutory text,[29]
2. focus on most locally applicable rule, not broader principles,[30]
3. usually exclude "external standards" in reasoning, like efficiency, justice, moral and political reasoning[31] or teleological interpretation,[32]
4. dismiss cases on formal grounds to avoid analyzing them on the merits,[33]
5. provide insufficient reasoning for their decisions.[34]

[28] Bobek (2015) argues that the term "formalism" is often used merely as a pejorative label. Similarly, Schauer (1988) observes that there is "scant agreement on what it is for decisions in law, or perspectives on law, to be formalistic, except that whatever formalism is, it is not good" (pp. 509–510). However, Schauer (1988) convincingly demonstrates that formalism has a descriptive content that can be clearly identified and, in some contexts, may even be worth pursuing.

[29] See Matczak et al., 2010, p. 86; Matczak et al., 2015; Bystranowski et al., 2022, p. 1911; Jakab et al., 2017, p. 222; Cserne, 2020, p. 881; Kühn, 2011.

[30] See Matczak et al., 2010, p. 87; Matczak et al., 2015; Bystranowski et al., 2022, p. 1911; Mańko, 2013, p. 7; Cserne, 2020, p. 881; Bencze, 2021; Kühn, 2011, p. 209.

[31] See Kühn, 2004, p. 557; Matczak et al., 2010, p. 86; Bystranowski et al., 2022, p. 1913; Malolepszy and Gluchowski, 2023.

[32] See Kühn, 2004, pp. 544, 558; Jakab et al., 2017, p. 222; Malolepszy and Gluchowski, 2023, p. 1798; Kühn, 2011, p. 210.

[33] See Mańko, 2013, p. 6; Bystranowski et al., 2022, p. 1913; Foric et al., 2020, p. 6; Uzelac, 2010, p. 383.

[34] See Kühn 2004, 557; Suteu 2023, 527, Kühn, 2011, 204.

Let me put the five core tenets (mainly the first three) in the context of existing interpretive theories both outside and inside the region:

German debate focuses on the interpretative canons. It traditionally distinguishes four interpretation methods: linguistic, systemic, historic, teleological (Alexy, 2010; Möllers, 2020). Besides, German authors have traditionally distinguished two theories of interpretation: subjective theory leaning towards the will of the legislator and objective theory leaning towards the purpose of the law (Möllers, 2020). Linguistic and systemic interpretation could be considered formalistic, while historic and teleological non-formalistic. Both subjective and objective theory of legal interpretation would be considered non-formalistic by this study.

Czech debate oscillates between authors preferring so called linguistic and systematic interpretation on one hand,[35] or teleological interpretation on the other.[36] Authors preferring linguistic and systematic interpretation would be considered formalists, authors opting for more teleological interpretation non-formalists.

[35] For instance, scholars like Gerloch or Tryzna. See Gerloch, Tryzna et al., 2012, Gerloch 2021. Wintr (2019) mentions "However, the serious problem is, among other things, the disagreement of the legal community on these rules of priority. In particular, two basic approaches clash here, one of which, in the apparent clash between a linguistic-systematic interpretation and a teleological interpretation, prefers the one and the other the other."

[36] For instance, scholar like Wintr (2019), Melzer (2011) – quite interestingly, the younger generation of legal scholars.

The US debate on statutory interpretation typically differentiates textualism,[37] purposivism, originalism and pragmatism.[38] Roughly speaking, the formalistic reasoning, as understood by this study, includes textualist reasoning and excludes purposivist, originalist and pragmatist reasoning.[39]

Thus, the core tenets of CEE formalism match with debate on interpretation both in Czechia and the CEE. They also match the debate on formalism in the US, which defines formalism very similarly to the first three core tenets described in the CEE literature.[40] This makes our

[37] For the definition, see Eskridge et al. (2022), who describe the methodology of "new textualism," as advocated by Justice Antonin Scalia and popular at the U.S. Supreme Court, as focusing on "the text, the whole text, and nothing but the text." They outline three core tenets of how modern textualists understand meaning: "(1) understood by the ordinary person, (2) applying standard rules of semantics, definitions, and grammar, (3) at the time the statute was enacted" (pp. 1612–1613). Similarly, Watson (2022) explains that "textualism limits the set of admissible arguments in hard cases: it confines judges to considering what a reasonable reader would have been most likely to infer that the legislature intended to assert rather than what the legislature 'really' intended to assert or which reading of the statute best advances its purpose or maximizes social welfare" (p. 46). Textualism is the dominant argumentation theory of the Supreme Court. For empirical studies, see, e.g., Krishnakumar (2023)

[38] For the overview of the approaches, see Watson (Forthcoming).

[39] As I will elaborate below, CEE formalism is distinct in its exclusion of originalism, categorizing historical interpretation as a non-formalistic argument. This contrasts with the U.S. context, where formalism is often characterized by a combination of textualism and originalism. For constitutional interpretation in the U.S., Thomas C. Grey notes that "the formalist approach to American constitutional law over the last half century has been embodied in the linked ideas of textualism and originalism" (Grey, 2014, p. 4). Similarly, Stiglitz and Thalken highlight that contemporary formalism emphasizes both textualism and originalism (2024).

[40] Grove (2020) characterizes "formalistic textualism" as an interpretive approach that prioritizes statutory language while deliberately downplaying policy concerns and practical consequences. Similarly, Stiglitz and Thalken (2024) distinguish between formal and grand reasoning, where formal reasoning treats law as a closed, mechanical system that excludes political, social, and economic considerations, while grand reasoning explicitly engages with these broader factors. Solum (2014) presents ideal types of formalism and realism as follows: "A perfectly formalist judge would decide entirely on the basis of the authoritative legal materials," while "a perfectly realist judge would decide entirely on the basis of policy preferences" (p. 2490). Solum further argues that this dichotomy reflects a deeper normative dispute about the proper role of authoritative sources in judicial decision-making—formalists

methodology useful for empirical research of legal reasoning in other CEE states and beyond.[41]

Ideally, after establishing the core tenets, one could prepare annotation guidelines and start analysing the case law. Since most tenets concern argumentation, pursuing further steps might seem like a straightforward task: simply read the decisions and quantify the types of arguments courts use. However, empirical analysis of Supreme Courts' argumentation is complicated for two main reasons.

First, despite most tenets of formalism concern argumentation (e.g., usage of text-based arguments), the scholarship on CEE formalism typically lacks detailed explanations necessary for annotation. For instance, the literature on CEE formalism lacks clear criteria what text-based argument is and when to consider it present in the judgment. Without such specification, reasoning practices remain unmeasurable.

Second, simply counting arguments proved insufficient. In our pilot studies, we noticed that context significantly matters; measuring argument frequency alone provides an incomplete picture of judicial formalism. Consider a Supreme Court decision that criticizes lower courts for excessive formalism or insufficient reasoning—such a decision might contain many references to statutory text or case law yet be fundamentally non-formalistic in its nature. Moreover, the last two tenets—dismissal on formal grounds and insufficient reasoning—cannot be captured through argument counting alone. These complexities demand a methodology that goes beyond simple quantification of arguments.

advocate for their primacy, while realists promote an instrumentalist approach that allows policy considerations to override the plain meaning of statutes or the doctrine of stare decisis (pp. 2490, 2492).

[41] The debate on formalism in Central and Eastern Europe (CEE) is marked by distinctive features shaped by the region's historical and socio-political context. What sets the CEE discourse apart is not necessarily the concept of formalism itself, but its framing: formalism is often regarded as a legacy of the communist era, its persistence seen as evidence of unfinished judicial reform, and the debate intertwined with the broader regional effort to confront and reconcile with the past. While these historical and political dimensions are crucial to understanding the debate, this study does not center on these aspects.

Thus, we needed to develop a new strategy to measure formalism. We came up with a dual approach. First, we developed a new annotation scheme and guidelines to analyze the frequency and distribution of different argument types, classifying them as either formalistic or non-formalistic. Second, we supplement this method by also holistically evaluating each decision with binary variable formalistic/non-formalistic, considering the context of the decision and the five core tenets of formalism.

2.1.2 New Taxonomy of Arguments for Empirical Analysis of CEE Argumentation Practices

The first way to bridge the gap between the core tenets and real world decisions is through argument types (MacCormick and Summers 2016). Legal theory traditionally uses argument types like historical interpretation, variously called "argumentation schemes", "patterns of arguments" (Walton et al., 2021) or "argument forms" (Alexy, 2010), to analyze and evaluate judicial reasoning. For instance, categorizing some set of propositions as historical interpretation enables one to assess strength and weakness of such argument (e.g. look whether explanatory note fundamentally differs from enacted provision due to a "rider") and to provide/debunk counter-arguments (e.g. to argue that the will of the legislator is/is not a chimera). Argument types enable lawyers to better understand and analyze decisions (Walton et al., 2021).

Using argument types for empirical research of formalism offers distinct advantages over binary formalistic/non-formalistic coding of sentences or paragraphs pursued by recent study on formalism by Stiglitz and Thalken (2024).[42] It better reflects how legal theorists analyze argumentation, provides more detailed information on using traditional interpretation methods (e.g. that Czech apex courts very scarcely refer

[42] Stiglitz and Thalken analyzed SCOTUS decisions by coding individual paragraphs with the variable "formalistic" and "grande," subsequently using this annotated dataset to train an argument mining model (Stiglitz and Thalken, 2024).

to the will of legislator) and benefits from well-established description of the argument schemes provided in the numerous literature on legal argumentation.

To annotate the decisions, we created a new taxonomy of arguments. Our taxonomy draws on three established taxonomies in legal argumentation theory: Alexy's theory of legal argumentation, Walton, Macagno and Sartor's taxonomy, itself a compilation of common and civil law argument taxonomies, and MacCormick and Summers' taxonomy based on a comparative study of statutory interpretation across nine jurisdictions. Based on these sources and the legal argumentation literature from the CEE context, we classify arguments into eight argument types:

I. **Formalistic Argument Types**
   1. **LIN** – Linguistic Interpretation
   2. **SI** – Systemic Interpretation (incl. CCI Constitutional Conforming Interpretation and EUCI EU Law Conforming Interpretation)
   3. **CL** – Case Law
   4. **D** – Doctrine

II. **Non-formalistic Argument Types**
   5. **HI** – Historical Interpretation
   6. **PL** – Principles of Law and Values (incl. CV – Constitutional Values, Rights, Principles and EUP – EU Values and Principles)
   7. **TI** – Teleological Interpretation
   8. **PC** – Practical Consequences

A comparison with existing taxonomies is enclosed in the annex.

Research design affects results and so does our taxonomy. One of the key research design decisions is how to classify case law. The easiest answer would be to consult the core tenets of CEE formalism, but they do not concern case law directly. Some authors argue that CEE judicial formalism is characterized by an "ideology of simple law," suggesting CEE formalism excludes case law (Kühn, 2004; Manko, 2013).[43] This perspective would suggest classifying case law as either neutral or non-formalistic type of argument. However, we find important, that case law also functions similarly to statutes, serving as an authoritative source of law generated within the legal system and constraining interpretive alternatives. Given that case law constitutes a significant portion of arguments—approximately 40 % of all arguments and 66 % of all formalistic arguments if classified as formalistic (see Part Three)—the decision on how to classify it heavily influences the results. In this study, we classify case law as a formalistic argument.[44]

Second, inflation of categories might cause inflation of presence of such category. For example, when a court references multiple constitutional principles in a single passage, the taxonomy determines whether this counts as one non-formalistic argument or multiple separate arguments. Previous studies that differentiated various fundamental rights into distinct categories would likely count multiple arguments, potentially inflating the non-formalistic score (Matczak et al., 2010; 2015). The same applies to any other argument type. To avoid such "inflation" and maintain theoretical validity, we aligned our taxonomy closely with the core tenets of formalism and the three established argumentation theories. Annex A includes a comparative table showing what our taxonomy shares with existing taxonomies and how it differs.

The third research design issue concerns linguistic interpretation. In a nutshell, the question is whether to distinguish between the application and interpretation of law. Courts often apply the law without explicitly reasoning about possible interpretative alternatives. In many cases, they just apply the law rather than interpreting it. But when they do so, courts likely rely on the text of the law as their primary reference, raising the question of whether such applications should be considered and annotated as linguistic interpretation. If we consider

[43] See, Kühn (2004): "Because any persuasive sources of law are beyond the ken of socialist scholars and judges, precedent is rather weightless. 'We have no precedent in our system of law, we are not common law judges,' is a typical answer to any objections made to the traditional refusal of precedent." (p. 560).

[44] Similarly, see, e.g. (Matczak et al., 2010; 2015).

such practices as linguistic interpretation, this will significantly inflate the frequency of linguistic interpretation, as virtually every decision would include at least one instance, often much more, given that courts routinely apply multiple statutes (e.g., procedural and substantive laws). To address this, we marked linguistic interpretation only when courts referred to standards like meaning of words, definitions, phrasing, or syntax, although the threshold was quite low.[45]

Annotating court judgments is challenging, even for domain experts (Habernal et al, 2023; Lüders 2024). Determining whether an argument type is present and identifying its category often involves interpretation. This requires analyzing context, language, the proposition's role in the text, its connection to prior sections, and the presence of argument type features. Ambiguity in decisions frequently demands clarification, one could say interpretation (of the court's interpretation). Since at least two annotators must independently agree on the presence or an absence of an argument type, the process relies heavily on expert knowledge, rigorous training, detailed guidelines and experts solving disagreements (Braun, 2024).

To tackle this issue, we created detailed annotation guidelines that include both a coding scheme and clear instructions defining each argument type and when to identify it. These guidelines were built through a step-by-step process: we started with established argumentation theories (Alexy 2010; MacCormick and Summers 2016; Walton et al., 2021), added insights from German and Czech legal scholarship to reflect Central European perspectives (Möllers 2020; Wintr 2019), tested and improved them on over 200 pilot decisions from Czech Supreme Courts (SC and SAC), and then verified the developments against the original theoretical sources. We went through these stages multiple times over several months before coding the MADON dataset.

We provide such guidelines in Annex B. For each argument type, our guidelines usually focus on five key elements to better guide the

[45] See our annotation guidelines in the annex B. Basically, any reference to "wording", "text" or phrase that "the text is clear" would very likely be considered linguistic interpretation.

annotators: the argument's core (such as purpose in teleological interpretation), subcategories (like travaux préparatoires within historical interpretation), typical phrases, so called "Triggers" (such as using the term "wording" which indicates linguistic interpretation), commentaries including tips for annotations, and examples (both borderline and typical). To enhance intercoder reliability, we developed decision flow charts to guide annotators through the annotation process. Given that existing studies share very little about how they annotated decisions (Matczak et al., 2010; 2015), the guidelines themselves are a contribution to empirical investigation of legal reasoning. We used this document heavily when annotating the decisions.

2.1.3 Holistic Assessment as Complementary Method to Measure Formalism

As mentioned above, quantifying arguments does not suffice for empirical research on formalism and argumentation.

During the pilot studies, we discovered that measuring argument frequency alone provides an incomplete picture of judicial formalism. Some decisions, while predominantly using formalistic arguments, appeared non-formalistic when evaluated holistically against the core tenets of formalism and considering the entire decision, its context, the facts of the case, and procedural history. For instance, although the decision no. 33 Cdo 1746/2022–235 of the Czech Supreme Court was rather short and relied on 1 linguistic, 2 case law arguments and 1 teleological argument, it was still considered non-formalistic, because the teleological argument was crucial.[46]

[46] The court determined that paying individual invoices cannot be considered an acknowledgment of the rest of the debt invoiced through other invoices. The court held that interpreting invoice payments as acknowledgment of other invoices would contradict the purpose of the legal provisions governing debt acknowledgment. According to the court, this provision aims to help creditors specifically in situations where a single invoice is only partially paid (and not when one invoice is fully paid while others are not), with the acknowledgment applying only to the unpaid portion of that same invoice. Although the court also referred to the wording and

Besides, we encountered decisions criticizing lower courts for excessive formalism which contained mainly formalistic arguments (typically case law), yet their overall approach was rather non-formalistic.

Furthermore, we repeatedly observed that arguments carry different weights in judicial reasoning.[47] A decision might reference a fundamental principle only once or twice in a crucial paragraph while containing numerous formalistic arguments in less important sections addressing procedural matters. Similarly, the clarity of arguments varies significantly—some are explicitly stated while others are barely present and must be inferred based on context. This complexity is further illustrated in cases where courts reference statutory provisions that themselves incorporate non-formalistic arguments like principles, raising questions about whether such reasoning should be classified as formalistic or non-formalistic. Besides, some principles have a formalistic nature, like the so-called concentration principle, that strictly limits parties from bringing new evidence in the later stages of the proceedings. Thus, quantification does not suffice.

These observations led us to develop a complementary holistic assessment method. Building on existing research on formalistic reasoning,[48] we created a framework that considers multiple parameters both general and specific to Central and Eastern European legal contexts. When evaluating whether particular decision is formalistic, our annotating team evaluated the following aspects: types of arguments used; their frequency; clarity and explicit presence of arguments; weight of arguments; and whether the court critiqued previous courts for formalistic reasoning. While holistic assessment relies on the quantific-

some case law, it did not really put emphasis on these arguments and focused on the purpose of the interpreted provision.

[47] Similarly, Choi highlights the varying significance of arguments as a limitation in his research, which examined argument types by identifying key terms associated with them. He notes, for instance, that “legislative history might be a decisive factor in a court’s ruling, even though it is only mentioned once. Or it could be mentioned several times, even though the court ultimately decides the case on other grounds” (Choi, 2020, p. 389).

[48] Alberstein (2012) argues that formalism can be assessed using various parameters, though each parameter can often point to different direction.

ation of arguments (e.g., determining whether a decision contains more or fewer formalistic arguments), it also requires evaluating additional factors, such as to what extent the court relies on the argument or what preceded the appeal to the Supreme Court. Most importantly, the annotators had also taken into account the five core tenets of formalism and the context of the case. An example of formalistic decision is included in the Annex F.

Our holistic assessment seems to be reliable and fruitful. We achieved a Cohen's kappa coefficient of 0.65 on the overall category formalistic/non-formalistic, indicating "substantial agreement" among annotators despite the complexity of the assessment.[49] This success yields two significant findings for future empirical research. First, research assistants can be effectively trained to achieve sufficient intercoder agreement on complex holistic assessments of formalistic/non-formalistic reasoning. Second, we found a strong correlation between decisions with higher frequencies of non-formalistic arguments and holistic non-formalistic assessments, suggesting that while the dual approach provides valuable validation, the relationship between argument types and overall formalism remains strong.

2.1.4 Summary

Both approaches necessitate a clear definition of formalism, a well-structured taxonomy of arguments, detailed annotation guidelines, and a team of domain experts for the annotation process. The taxonomy and guidelines should align with current theories of legal argumentation. We provide these in the annex.

[49] A value of 0 indicates agreement by chance. According to Landis and Koch (1977), values between 0.21 and 0.40 are considered fair, 0.41 to 0.60 moderate, 0.61 to 0.80 substantial, and 0.81 to 1.00 almost perfect (Landis & Koch, 1977).

2.2 Further Methodology and Data

2.2.1. Czech Institutional Context

Before going into methodology, three key aspects need to be clarified to better understand the Czech Supreme Courts: 1) the structure of the Czech judicial system, 2) how Supreme Courts operate and 3) what is the role of Czech Supreme Courts within the system.

Firstly, Czech judiciary is dual-tiered, comprising the Constitutional Court (as separate "system") and the ordinary court system. The ordinary court system is hierarchically structured. Two courts remain at the top: the Supreme Court (Nejvyšší soud), hearing criminal and civil matters, and the Supreme Administrative Court (Nejvyšší správní soud, established in 2003), hearing administrative matters. Beneath these supreme courts, the hierarchy includes High Courts (vrchní soudy), Regional Courts (krajské soudy), and District Courts (okresní soudy).

Secondly, the Supreme Court serves as the highest court of appeal for civil and criminal matters in Czechia; similarly, Supreme Administrative Court is the highest court of appeal for administrative matters. The Supreme Court is comprised of two branches: the Civil Part and the Criminal Part.

Third, the Supreme Courts ensure uniformity of law. They mainly focus on legal questions, not determinations of facts. Both Supreme Courts hear extraordinary appeals, which are essentially appeals against rulings of lower courts. The role of the Supreme Courts is to remove and prevent inconsistencies in how lower courts interpret and apply law.

2.2.2 Dataset

Concerning sampling, the methodological literature identifies various sampling techniques for use in content analysis. Hall and Wright (2008) outline four primary techniques: (1) true random sampling, typically achieved through computer-generated random numbers; (2) systematic sampling, such as selecting every fifth case; (3) quota sampling, which involves selecting a specific number of cases per category, such as up to two hundred cases per jurisdiction per year; and (4) purposive sampling, where cases are chosen based on their relevance to the study. Krippendorff (2018) expands on these methods, detailing eight sampling techniques: (1) random sampling, (2) systematic sampling, (3) stratified sampling, (4) varying probability sampling, (5) cluster sampling, (6) snowball sampling, (7) relevance sampling, and (8) census sampling.

We annotated a dataset of 272 decisions of Supreme and Supreme Administrative Court.[50]

The dataset consists of 272 judicial decisions from the Supreme Court and Supreme Administrative Court, spanning the period from 1997 to 2024, with a focus on decisions made between 2003 and 2023. To ensure a representative sample, stratified sampling was employed, balancing cases across time periods, court agendas (civil, criminal, and administrative), and types of cases (procedural and on the merits).

The temporal dimension of the stratified sampling was structured around ten time periods spanning 1997 to 2024, with civil cases beginning in 1997, criminal in 2000 (reflecting the availability of decisions) and administrative cases in 2003 (reflecting the establishment of the Supreme Administrative Court in 2003). The stratified sampling mostly followed three-year intervals (1997, 2000, 2003, 2006, 2009, 2012, 2015, 2018, 2021, and 2023/2024), with target quotas ranging from 25 to 31 decisions per period to maintain temporal balance.

The sampling further ensured approximately equal distribution between procedural decisions (usnesení) and decisions on the merits (rozsudky), which constitute 49.45 % and 50.18 % of the dataset respectively. By design, decisions on the merits are overrepresented in the dataset (in contrast to the population), which may partially limit repre-

[50] Using online available calculators, such sample of 272 shall be representative given the population size is ca 230.000 (with the confidence level of 90 %, margin of error 5 %). See, e.g., https://www.calculator.net/sample-size-calculator.html?type=1&cl=90&ci=5&pp=50&ps=230000&x=Calculate. Of course, all this depends on the sampling process. The issue of representativeness will disappear once we manage to develop the argument mining model.

sentativeness but was intended to avoid a situation in which the dataset would be dominated by procedurally insignificant dismissals.

This sampling ensured proportional representation across the Supreme Court's civil branch (122 cases), criminal branch (58 cases), and the Supreme Administrative Court (90 cases).

Within these subpopulations, randomized sampling was used to prevent overrepresentation of specific judges or benches.

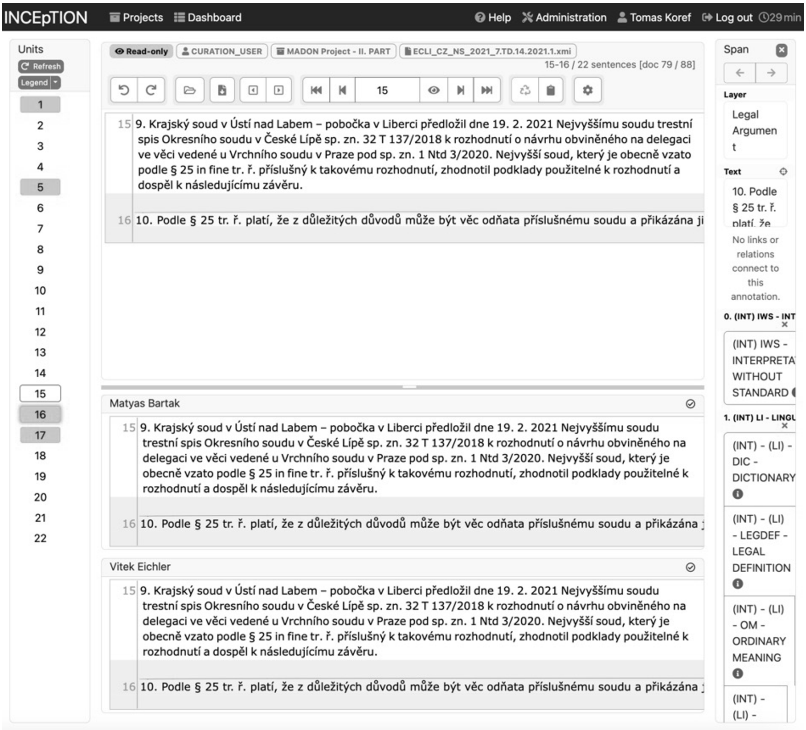
This stratification by court branch, time period, and decision type combined with randomization within each stratum, was designed to become a close representation of the relevant population of approximately 230,000 decisions of both Supreme Courts. I provide detailed overview of the dataset in Annex E.[51]

2.2.3 Annotation Process

The systematic annotation of Czech judicial decisions was conducted between July and September 2024, building on a pilot study from early 2024 that analysed 160 decisions and helped refine the methodology. We employed four law students from the Faculty of Law as research assistants. All of them had some background in legal theory and argumentation.

On the technical level, we used INCEpTION as our annotation software. Inception also calculates the intercoder agreement and enables solving the disagreement in a curation mode. This is an example of a single short decision with four annotations in paragraphs 1, 5, 16 and 17:

[51] This document was prepared by research assistants under my supervision. I pursued all the sampling.


INCEpTION
Projects
Dashboard
Help
Administration
Tomas Koref
Log out
29 min
Units
Refresh
Legend
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Read-only
CURATION_USER
MADON Project - II. PART
ECLI_CZ_NS_2021_7.TD.14.2021.1.xmi
15-16 / 22 sentences [doc 79 / 88]
15
Span
Layer
Legal Argument
Text
10. Podle § 25 tr. ř. platí, že
No links or relations connect to this annotation.
0. (INT) IWS - INT
(INT) IWS - INTERPRETA WITHOUT STANDARD
1. (INT) LI - LINGU
(INT) - (LI) - DIC - DICTIONARY
(INT) - (LI) - LEGDEF - LEGAL DEFINITION
(INT) - (LI) - OM - ORDINARY MEANING
(INT) - (LI) -
Matyas Bartak
Vitek Eichler

15 9. Krajský soud v Ústí nad Labem – pobočka v Liberci předložil dne 19. 2. 2021 Nejvyššímu soudu trestní spis Okresního soudu v České Lípě sp. zn. 32 T 137/2018 k rozhodnutí o návrhu obviněného na delegaci ve věci vedené u Vrchního soudu v Praze pod sp. zn. 1 Ntd 3/2020. Nejvyšší soud, který je obecně vzato podle § 25 in fine tr. ř. příslušný k takovému rozhodnutí, zhodnotil podklady použitelné k rozhodnutí a dospěl k následujícímu závěru.

16 10. Podle § 25 tr. ř. platí, že z důležitých důvodů může být věc odňata příslušnému soudu a přikázána ji

Matyas Bartak

15 9. Krajský soud v Ústí nad Labem – pobočka v Liberci předložil dne 19. 2. 2021 Nejvyššímu soudu trestní spis Okresního soudu v České Lípě sp. zn. 32 T 137/2018 k rozhodnutí o návrhu obviněného na delegaci ve věci vedené u Vrchního soudu v Praze pod sp. zn. 1 Ntd 3/2020. Nejvyšší soud, který je obecně vzato podle § 25 in fine tr. ř. příslušný k takovému rozhodnutí, zhodnotil podklady použitelné k rozhodnutí a dospěl k následujícímu závěru.

16 10. Podle § 25 tr. ř. platí, že z důležitých důvodů může být věc odňata příslušnému soudu a přikázána

Vitek Eichler

15 9. Krajský soud v Ústí nad Labem – pobočka v Liberci předložil dne 19. 2. 2021 Nejvyššímu soudu trestní spis Okresního soudu v České Lípě sp. zn. 32 T 137/2018 k rozhodnutí o návrhu obviněného na delegaci ve věci vedené u Vrchního soudu v Praze pod sp. zn. 1 Ntd 3/2020. Nejvyšší soud, který je obecně vzato podle § 25 in fine tr. ř. příslušný k takovému rozhodnutí, zhodnotil podklady použitelné k rozhodnutí a dospěl k následujícímu závěru.

16 10. Podle § 25 tr. ř. platí, že z důležitých důvodů může být věc odňata příslušnému soudu a přikázána

The process was divided into four phases:

1. Introduction (Week 1): Project orientation and interface familiarization
2. Training (Weeks 2–6): Training annotation of 80 decisions with reviews ca 4 times a week.
3. Coding (Weeks 7–12): Annotation of 272 main dataset decisions with reviews 2–3 times a week.
4. Finalization (Week 13): Resolution of disagreements and finalization of dataset.

Methodological Decisions and Solutions

We chose paragraph-level annotation over sentence-level. Sentence-level coding produced very low Krippendorf unitized Alpha scores. From the perspective of our research question, we were not so much interested in where a particular argument exactly ends, but more about whether a particular argument is present. Thus, this decision should not negatively influence the results.

Each argument type would be counted only once per paragraph, even if multiple instances within one paragraph appeared. However, one paragraph could include two different argument types. For example, if a paragraph cited five cases and referenced two explanatory notes, we would code it once for case law and once for historical interpretation. We hypothesize that this approach reflects how legal theorists typically analyze argumentation: focusing on argument types rather than counting individual citations within one single sentence or paragraph. When an argument type appeared repeatedly in different parts of the decisions, we annotated it as present multiple times. When courts referenced previous arguments with phrases like "with regard to above," we included these references only when the specific arguments were clearly identifiable.

We excluded rejected arguments. For instance, when a court discussed competing interpretations (e.g., court noted that historical interpretation suggests outcome X, linguistic interpretation points to outcome Y, and court ultimately decided for X rejecting linguistic interpretation), we would only annotate the accepted argument (historical interpretation in this case).[52] For more details on how we approached rejected arguments, see our guidelines.

We followed Braun's recent recommended practices for legal dataset annotation (Braun, 2024). This included maintaining detailed records of the annotation team (law students as annotators, PhD candidate in legal argumentation as arbiter, and legal practitioners as consulting experts for minority of complicated cases), ensuring at least two

[52] Similarly, Krishnakumar, 2020.

independent annotations per decision, and automatically measuring as well as documenting intercoder agreement using INCEpTION. Disagreements were resolved through "arbiter review" or a combination of arbiter review and "forced agreement" for complex cases, typically for the holistic label. Simple arbiter review meant that independent arbiter (usually me) checked the disagreement and decided it according to guidelines. The combination arbiter review and forced agreement meant that annotators were required to discuss their disagreements and propose potential solutions, accompanied by justifications for their choices. The arbiter then reviewed both the suggested solutions and their justifications. In some instances, annotators could not reach consensus even after consultation, while in a minority of cases, they agreed on solutions that appeared to deviate from the annotation guidelines. In these cases, the arbiter retained authority to override the annotators' consensus based on the established guidelines, while taking into account the annotators' reasoning.

The project included over 1,000 hours of annotations, annotating ca 350 decisions (272 after excluding training cases). The dataset of 272 decisions contains 9,183 paragraphs that include 1,913 legal arguments in total.

As mentioned above, we measured intercoder agreement to ensure reliability of our annotations. While most categories showed good agreement, Linguistic Interpretation, Systematic Interpretation and Practical Consequences categories demonstrated lower reliability. This shows the inherent complexity of legal document annotation. Annotation of legal documents is generally considered complicated and the intercoder agreement is often lower.[53] The intercoder agreement is described in the Annex D. Disagreements were solved by the process described above.

[53] See Braun (2024), who compared intercoder agreement in existing annotated legal datasets and found that average Krippendroff's alpha is 0,677, average Fleiss' kappa being 0,675.

Quantification of formalism

To measure courts' formalism, we needed certain indicators that link formalism as abstract concept to real world practices and outcomes of the courts (Ovádek et al., 2025). When developing these indicators, we relied on following hypotheses derived from the core tenets of CEE formalism:

1. Formalistic courts will use formalistic arguments more often.
2. Even when court uses some non-formalistic arguments, a court can still be formalistic if formalistic arguments disproportionately dominate
3. Formalistic courts issue more decisions that completely exclude non-formalistic arguments.
4. Formalistic court issues more decisions holistically evaluated as formalistic.

Based on these hypotheses, we implemented four key indicators to measure formalism:

1. The average number of formalistic and non-formalistic arguments per decision.
2. The proportion of formalistic to non-formalistic arguments.[54]
3. The proportion of decisions that rely exclusively on formalistic arguments or include no arguments (i.e., exclude non-formalistic arguments).
4. A proportion of decisions holistically evaluated as formalistic and non-formalistic.

According to our formalism indicators, a court is considered more formalistic when it demonstrates more formalistic arguments per de-

[54] We believe this indicator addresses the concern raised by Choi (2020): as decisions become longer, the average number of arguments may increase simply due to their length. Without considering the relative proportions of argument types, one might overestimate a court's reliance on formalistic arguments. By focusing on the proportion of formalistic to non-formalistic arguments, our approach highlights the balance between these groups rather than merely capturing an absolute increase in one type.

cision, fewer non-formalistic arguments per decision, a higher proportion of formalistic arguments overall, more decisions relying exclusively on formalistic arguments, or a higher proportion of decisions holistically evaluated as formalistic.

2.3 Limitations

This study has five main limitations:

First, while the dataset spans a broad timeframe, including decisions from 1997 to 2024 (primarily 2003–2023), it consists of 272 annotated cases. Although I employed stratified and randomized sampling to ensure representativeness across time periods, court agendas, and case types, the limited sample size means that the findings should be interpreted as tentative, especially given the study's role in developing argument mining models that will enable large scale analysis of all the decisions ever published (230k). Nonetheless, the dataset shall be representative and provides a meaningful snapshot of judicial reasoning practices.

Second, comparing the Supreme Court and Supreme Administrative Court poses challenges due to their differing agendas and appellate procedures. While the courts deal with distinct subject matter, I aimed to mitigate these issues by also focusing on relative comparisons over time. Besides, I tried to disprove the claims that already engaged in the comparison. Moreover, even if the comparative aspects were removed, the study provides valuable insights into the reasoning practices of each of the two Supreme Courts on its own.

Another limitation lies in the possibility that judicial opinions do not fully reflect the real reasons behind decisions. Courts often engage in post hoc justification, which, of course, might (and often will) differ from the actual decision-making process. However, since critiques of formalism focus on the reasoning articulated in written opinions, this study remains valid as it evaluates reasoning practices as presented.[55] Besides, the reasoning provided in decisions matters because it is the main output of judiciary for the parties to the dispute, the public and the scholarship.

Additionally, the study addresses only certain conceptualizations of formalism, as defined in existing literature. Formalism is a contested concept, and indicators of formalism vary. By narrowing the scope to a specific definition, the study ensures better reliability and validity, even though the findings might be supplemented be other operationalization of formalism in the future. The definition used reflects prominent scholarly debates, and this ensures the research contributes to the ongoing discourse on formalism in judicial reasoning.

Finally, the study encountered challenges with inter-coder reliability for certain argument types, particularly Systematic Interpretation and Practical Consequences. While this highlights the inherent complexity of annotating legal documents, substantial agreement was achieved in six out of nine categories, and disagreements were resolved through expert review. This approach aligns with best practices in legal dataset annotation, as outlined by, e.g., Braun (2024), ensuring that the findings remain robust despite these challenges.

Notwithstanding these limitations, the study provides a significant contribution to understanding formalism in judicial decision-making and offers a foundation for future research in this area.

[55] The question of whether arguments genuinely influence judicial decisions or serve as mere "window dressing" for public and professional audiences requires a different methodological approach. For instance, Abbe R. Gluck and Richard A. Posner conducted interviews with 42 federal appellate judges, exploring the role of arguments in statutory interpretation and decision-making. Their findings suggest that linguistic canons are often used as "window dressing," applied after a judge has already reached a conclusion. However, they note that determining the extent of this phenomenon remains challenging: "Linguistic canons especially, as opposed to policy canons, seem to be of this 'window dressing' variety" and "The question of how much work the canons are really doing and how much is mere 'show' (or cover for the common law tools they wish to deploy) is difficult to resolve" (Gluck & Posner, 2018, pp. 1330, 1353). Similarly, Spamann et al. used experimental methods, monitoring the computer activity of 299 judges with diverse backgrounds, to analyze the role of precedents in judicial decision-making.