

---

# CITATION DATA OF CZECH APEX COURTS

---

A PREPRINT

**Jakub Harašta**

Institute of Law and Technology  
Masaryk University  
Brno, CZ  
jakub.harasta@law.muni.cz

**Tereza Novotná**

Institute of Law and Technology  
Masaryk University  
Brno, CZ  
tereza.novotna@mail.muni.cz

**Jaromír Šavelka**

Intelligent Systems Program  
University of Pittsburgh  
Pittsburgh, USA  
jas438@pitt.edu

February 7, 2020

**ABSTRACT**

In this paper, we introduce the citation data of the Czech apex courts (Supreme Court, Supreme Administrative Court and Constitutional Court). This dataset was automatically extracted from the corpus of texts of Czech court decisions - CzCDC 1.0. We obtained the citation data by building the natural language processing pipeline for extraction of the court decision identifiers. The pipeline included the (i) document segmentation model and the (ii) reference recognition model. Furthermore, the dataset was manually processed to achieve high-quality citation data as a base for subsequent qualitative and quantitative analyses. The dataset will be made available to the general public.

**Keywords** Reference recognition · Reference extraction · Document segmentation · NLP pipeline · Citation data · Supreme Court · Supreme Administrative Court · Constitutional Court · Czech Republic

## 1 Introduction

Analysis of the way court decisions refer to each other provides us with important insights into the decision-making process at courts. This is true both for the common law courts and for their counterparts in the countries belonging to the continental legal system. Citation data can be used for both qualitative and quantitative studies, casting light in the behavior of specific judges through document analysis or allowing complex studies into changing the nature of courts in transforming countries.

That being said, it is still difficult to create sufficiently large citation datasets to allow a complex research. In the case of the Czech Republic, it was difficult to obtain a relevant dataset of the court decisions of the apex courts (Supreme Court, Supreme Administrative Court and Constitutional Court). Due to its size, it is nearly impossible to extract the references manually. One has to reach out for an automation of such task. However, study of court decisions displayed many different ways that courts use to cite even decisions of their own, not to mention the decisions of other courts. The great diversity in citations led us to the use of means of the natural language processing for the recognition and the extraction of the citation data from court decisions of the Czech apex courts.

In this paper, we describe the tool ultimately used for the extraction of the references from the court decisions, together with a subsequent way of manual processing of the raw data to achieve a higher-quality dataset. Section 2 maps the related work in the area of legal citation analysis (Section 2.1), reference recognition (Section 2.2), text segmentation (Section 2.4), and data availability (Section 2.3). Section 3 describes the method we used for the citation extraction, listing the individual models and the way we have combined these models into the NLP pipeline. Section 4 presents

results in the terms of evaluation of the performance of our pipeline, the statistics of the raw data, further manual processing and statistics of the final citation dataset. Section 5 discusses limitations of our work and outlines the possible future development. Section 6 concludes this paper.

## 2 Related work

### 2.1 Legal Citation Analysis

The legal citation analysis is an emerging phenomenon in the field of the legal theory and the legal empirical research. The legal citation analysis employs tools provided by the field of network analysis.

In spite of the long-term use of the citations in the legal domain (eg. the use of Shepard's Citations since 1873), interest in the network citation analysis increased significantly when Fowler et al. published the two pivotal works on the case law citations by the Supreme Court of the United States [2, 3]. Authors used the citation data and network analysis to test the hypotheses about the function of *stare decisis* the doctrine and other issues of legal precedents. In the continental legal system, this work was followed by Winkels and de Ruyter [1]. Authors adopted similar approach to Fowler to the court decisions of the Dutch Supreme Court. Similar methods were later used by Derlén and Lindholm [4, 5] and Panagis and Šadl [6] for the citation data of the Court of Justice of the European Union, and by Olsen and Küçüksu for the citation data of the European Court of Human Rights [8].

Additionally, a minor part in research in the legal network analysis resulted in the past in practical tools designed to help lawyers conduct the case law research. Kuppevelt and van Dijck built prototypes employing these techniques in the Netherlands [7]. Görög a Weisz introduced the new legal information retrieval system, Justeus, based on a large database of the legal sources and partly on the network analysis methods. [28]

### 2.2 Reference Recognition

The area of reference recognition already contains a large amount of work. It is concerned with recognizing text spans in documents that are referring to other documents. As such, it is a classical topic within the AI & Law literature.

The extraction of references from the Italian legislation based on regular expressions was reported by Palmirani et al. [12]. The main goal was to bring references under a set of common standards to ensure the interoperability between different legal information systems.

De Maat et al. [10] focused on an automated detection of references to legal acts in Dutch language. Their approach consisted of a grammar covering increasingly complex citation patterns.

Opijnen [9] aimed for a reference recognition and a reference standardization using regular expressions accounting for multiple the variant of the same reference and multiple vendor-specific identifiers.

The language specific work by Kríž et al. [11] focused on the detecting and classification references to other court decisions and legal acts. Authors used a statistical recognition (HMM and Perceptron algorithms) and reported F1-measure over 90% averaged over all entities. It is the state-of-art in the automatic recognition of references in the Czech court decisions. Unfortunately, it allows only for the detection of docket numbers and it is unable to recognize court-specific or vendor-specific identifiers in the court decisions.

Other language specific-work includes our previous reference recognition model presented in [13]. Prediction model is based on conditional random fields and it allows recognition of different constituents which then establish both explicit and implicit case-law and doctrinal references. Parts of this model were used in the pipeline described further within this paper in Section 3.

### 2.3 Data Availability

Large scale quantitative and qualitative studies are often hindered by the unavailability of court data. Access to court decisions is often hindered by different obstacles. In some countries, court decisions are not available at all, while in some other they are accessible only through legal information systems, often proprietary. This effectively restricts the access to court decisions in terms of the bulk data. This issue was already approached by many researchers either through making available selected data for computational linguistics studies or by making available datasets of digitized data for various purposes. Non-exhaustive list of publicly available corpora includes British Law Report Corpus [14], The Corpus of US Supreme Court Opinions [15], the HOLJ corpus [16], the Corpus of Historical English Law Reports, Corpus de Sentencias Penales [17], Juristisches Referenzkorpus [18] and many others.

Language specific work in this area is presented by the publicly available Czech Court Decisions Corpus (CzCDC 1.0) [19]. This corpus contains majority of court decisions of the Czech Supreme Court, the Supreme Administrative Court and the Constitutional Court, hence allowing a large-scale extraction of references to yield representative results. The CzCDC 1.0 was used as a dataset for extraction of the references as is described further within this paper in Section 3. Unfortunately, despite containing 237 723 court decisions issued between 1st January 1993 and 30th September 2018, it is not complete. This fact is reflected in the analysis of the results.

## 2.4 Document Segmentation

A large volume of legal information is available in unstructured form, which makes processing these data a challenging task – both for human lawyers and for computers. Schweighofer [33] called for generic tools allowing a document segmentation to ease the processing of unstructured data by giving them some structure.

Topic-based segmentation often focuses on identifying specific sentences that present borderlines of different textual segments.

The automatic segmentation is not an individual goal – it always serves as a prerequisite for further tasks requiring structured data. Segmentation is required for the text summarization [30, 20], keyword extraction [31], textual information retrieval [32], and other applications requiring input in the form of structured data.

Major part of research is focused on semantic similarity methods. The computing similarity between the parts of text presumes that a decrease of similarity means a topical border of two text segments. This approach was introduced by Hearst [20] and was used by Choi [21] and Heinonen [22] as well.

Another approach takes word frequencies and presumes a border according to different key words extracted. Reynar [23] authored graphical method based on statistics called dotplotting. Similar techniques were used by Ye [24] or Saravanan [25]. Bommarito et al. [26] introduced a Python library combining different features including pre-trained models to the use for automatic legal text segmentation. Li [27] included neural network into his method to segment Chinese legal texts.

Šavelka and Ashley [34] similarly introduced the machine learning based approach for the segmentation of US court decisions texts into seven different parts. Authors reached high success rates in recognizing especially the Introduction and Analysis parts of the decisions.

Language specific work includes the model presented by Harašta et al. [29]. This work focuses on segmentation of the Czech court decisions into pre-defined topical segments. Parts of this segmentation model were used in the pipeline described further within this paper in Section 3.

## 3 Methodology

In this paper, we present and describe the citation dataset of the Czech top-tier courts. To obtain this dataset, we have processed the court decisions contained in CzCDC 1.0 dataset by the NLP pipeline consisting of the segmentation model introduced in [29], and parts of the reference recognition model presented in [13]. The process is described in this section.

### 3.1 Dataset and models

#### 3.1.1 CzCDC 1.0 dataset

Novotná and Harašta [19] prepared a dataset of the court decisions of the Czech Supreme Court, the Supreme Administrative Court and the Constitutional Court. The dataset contains 237,723 decisions published between 1<sup>st</sup> January 1993 and the 30<sup>th</sup> September 2018. These decisions are organised into three sub-corpora. The sub-corpus of the Supreme Court contains 111,977 decisions, the sub-corpus of the Supreme Administrative Court contains 52,660 decisions and the sub-corpus of the Constitutional Court contains 73,086 decisions. Authors in [19] assessed that the CzCDC currently contains approximately 91% of all decisions of the Supreme Court, 99,5% of all decisions of the Constitutional Court, and 99,9% of all decisions of the Supreme Administrative Court. As such, it presents the best currently available dataset of the Czech top-tier court decisions.

#### 3.1.2 Reference recognition model

Harašta and Šavelka [13] introduced a reference recognition model trained specifically for the Czech top-tier courts. Moreover, authors made their training data available in the [35]. Given the lack of a single citation standard, references

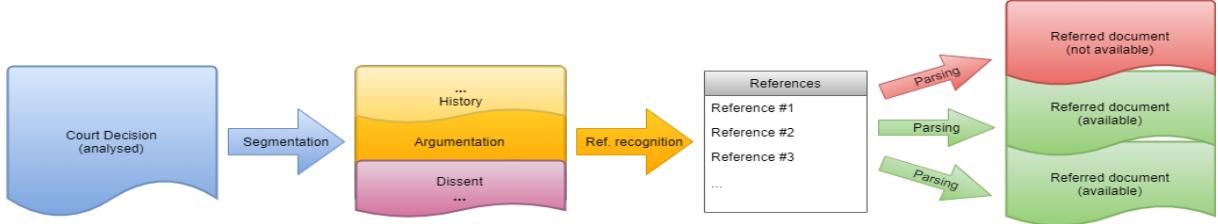


Figure 1: NLP pipeline including the text segmentation, reference recognition and parsing of references to the specific document

Table 1: Model performance

	Strict agreement			Overlap agreement		
	P	R	F1	P	R	F1
Reference recognition (court identifier) per [13]	-	-	.652	-	-	.709
Text segmentation (argumentation detection) per [29]	.885	.950	.915	-	-	-
Pipeline	.732	.716	.724	.846	.786	.815

in this work consist of smaller units, because these were identified as more uniform and therefore better suited for the automatic detection. The model was trained using conditional random fields, which is a random field model that is globally conditioned on an observation sequence  $O$ . The states of the model correspond to event labels  $E$ . Authors used a first-order conditional random fields. Model was trained for each type of the smaller unit independently.

### 3.1.3 Text segmentation model

Harašta et al. [29], authors introduced the model for the automatic segmentation of the Czech court decisions into pre-defined multi-paragraph parts. These segments include the Header (introduction of given case), History (procedural history prior the apex court proceeding), Submission/Rejoinder (petition of plaintiff and response of defendant), Argumentation (argumentation of the court hearing the case), Footer (legally required information, such as information about further proceedings), Dissent and Footnotes. The model for automatic segmentation of the text was trained using conditional random fields. The model was trained for each type independently.

## 3.2 Pipeline

In order to obtain the citation data of the Czech apex courts, it was necessary to recognize and extract the references from the CzCDC 1.0. Given that training data for both the reference recognition model [13, 35] and the text segmentation model [29] are publicly available, we were able to conduct extensive error analysis and put together a pipeline to arguably achieve the maximum efficiency in the task. The pipeline described in this part is graphically represented in Figure 1.

As the first step, every document in the CzCDC 1.0 was segmented using the text segmentation model. This allowed us to treat different parts of processed court documents differently in the further text processing. Specifically, it allowed us to subject only the specific part of a court decision, in this case the court argumentation, to further the reference recognition and extraction. A textual segment recognised as the court argumentation is then processed further.

As the second step, parts recognised by the text segmentation model as a court argumentation was processed using the reference recognition model. After carefully studying the evaluation of the model's performance in [13], we have decided to use only part of the said model. Specifically, we have employed the recognition of the court identifiers, as we consider the rest of the smaller units introduced by Harašta and Šavelka of a lesser value for our task. Also, deploying only the recognition of the court identifiers allowed us to avoid the problematic parsing of smaller textual units into the references. The text spans recognised as identifiers of court decisions are then processed further.

At this point, it is necessary to evaluate the performance of the above mentioned part of the pipeline before proceeding further. The evaluation of the performance is summarised in Table 1. It shows that organising the two models into the pipeline boosted the performance of the reference recognition model, leading to a higher F1 measure in the initial recognition of the text spans and their classification.

Table 2: References sorted by categories, unlinked

Court	Supreme Court	Supreme Adm. Court	Constitutional Court	Rest
Supreme Court cites	153 242	804	80 658	112 287
Supreme Administrative Court cites	1 342	90 217	14 756	20 709
Constitutional Court cites	8 486	2 877	137 308	45 689

Table 3: References linked with texts in CzCDC

Court	Supreme Court	Supreme Adm. Court	Constitutional Court
Supreme Court cites	140 355	658	76 003
Supreme Administrative Court cites	1 191	84 728	13 473
Constitutional Court cites	7 474	2 168	133 148

Further processing included:

1. control and repair of incompletely identified court identifiers (manual);
2. identification and sorting of identifiers as belonging to Supreme Court, Supreme Administrative Court or Constitutional Court (rule-based, manual);
3. standardisation of different types of court identifiers (rule-based, manual);
4. parsing of identifiers with court decisions available in CzCDC 1.0.

## 4 Results

Overall, through the process described in Section 3, we have retrieved three datasets of extracted references - one dataset per each of the apex courts. These datasets consist of the individual pairs containing the identification of the decision from which the reference was retrieved, and the identification of the referred documents. As we only extracted references to other judicial decisions, we obtained 471,319 references from Supreme Court decisions, 167,237 references from Supreme Administrative Court decisions and 264,463 references from Constitutional Court Decisions. These are numbers of text spans identified as references prior the further processing described in Section 3.

These references include all identifiers extracted from the court decisions contained in the CzCDC 1.0. Therefore, this number includes all other court decisions, including lower courts, the Court of Justice of the European Union, the European Court of Human Rights, decisions of other public authorities etc. Therefore, it was necessary to classify these into references referring to decisions of the Supreme Court, Supreme Administrative Court, Constitutional Court and others. These groups then underwent a standardisation - or more precisely a resolution - of different court identifiers used by the Czech courts. Numbers of the references resulting from this step are shown in Table 2.

Following this step, we linked court identifiers with court decisions contained in the CzCDC 1.0. Given that, the CzCDC 1.0 does not contain all the decisions of the respective courts, we were not able to parse all the references. Numbers of the references resulting from this step are shown in Table 3.

## 5 Discussion

This paper introduced the first dataset of citation data of the three Czech apex courts. Understandably, there are some pitfalls and limitations to our approach.

As we admitted in the evaluation in Section 3.2, the models we included in our NLP pipelines are far from perfect. Overall, we were able to achieve a reasonable recall and precision rate, which was further enhanced by several round of manual processing of the resulting data. However, it is safe to say that we did not manage to extract all the references. Similarly, because the CzCDC 1.0 dataset we used does not contain all the decisions of the respective courts, we were not able to parse all court identifiers to the documents these refer to. Therefore, the future work in this area may include further development of the resources we used. The CzCDC 1.0 would benefit from the inclusion of more documents of the Supreme Court, the reference recognition model would benefit from more refined training methods etc.

That being said, the presented dataset is currently the only available resource of its kind focusing on the Czech court decisions that is freely available to research teams. This significantly reduces the costs necessary to conduct these types of studies involving network analysis, and the similar techniques requiring a large amount of citation data.

## 6 Conclusion

In this paper, we have described the process of the creation of the first dataset of citation data of the three Czech apex courts. The dataset is publicly available for download at <https://github.com/czech-case-law-relevance/czech-court-citations-dataset>.

## Acknowledgment

J.H., and T.N. gratefully acknowledge the support from the Czech Science Foundation under grant no. GA-17-20645S. T.N. also acknowledges the institutional support of the Masaryk University. This paper was presented at CEILI Workshop on Legal Data Analysis held in conjunction with Jurix 2019 in Madrid, Spain.

## References

- [1] Radboud WINKELS, Jelle DE RUYTER. Survival of the Fittest: Network Analysis of Dutch Supreme Court Cases. Proceedings of AICOL 2011, pp. 106–115.
- [2] James H. FOWLER, Timothy R. JOHNSON, James F. SPRIGGS II, Sangick JEON, Paul J. WAHLBECK. Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis, 2007, vol. 15, no. 3, pp: 324–346.
- [3] James H. FOWLER, Sangick JEON. The Authority of Supreme Court Precedent. Social Networks, 2008, vol. 30, no. 1, pp. 16–30.
- [4] Mattias DERLÉN, Johan LINDHOLM. Select Peek-A-Boo, It's a Case Law System! Comparing the European Court of Justice and the United States Supreme Court from a Network Perspective. German Law Journal, 2017, vol. 18, no. 3, pp. 647–686.
- [5] Mattias DERLÉN, Johan LINDHOLM. Is it Good Law? Network Analysis and the CJEU's Internal Market Jurisprudence. Journal of International Economic Law, 2017, vol. 20, no. 2, pp. 257–277.
- [6] Yannis PANAGIS, Urška ŠADL, Fabien TARISSAN. Giving Every Case Its (Legal) Due – The Contribution of Citation Networks and Text Similarity Techniques to Legal Studies of European Union Law. Proceedings of Jurix 2017, pp. 59–68.
- [7] Dafne VAN KUPPEVELT, Gijs VAN DIJCK. Answering Legal Research Questions About Dutch Case law with Network Analysis and Visualization. Proceedings of JURIX 2017, pp. 95–100.
- [8] Henrik Palmer OLSEN, Aysel KÜÇÜKSU. Finding hidden patterns in ECtHR's case law: On how citation network analysis can improve our knowledge of ECtHR's Article 14 practice. International Journal of Discrimination and the Law, 2017, vol. 17, no. 1, pp. 4–22.
- [9] Marc VAN OPIJNEN. Canonicalizing Complex Case Law Citations. Proceedings of Jurix 2010, pp. 97–106.
- [10] Emile DE MAAT, Radboud WINKELS and Tom VAN ENGERS. Automated Detection of Reference Structures in Law. Proceedings of Jurix 2006, pp. 41–50.
- [11] Vincent KRÍŽ, Barbora HLADKÁ, Jan DĚDEK, Martin NEČASKÝ. Statistical Recognition of References in Czech Court Decisions. Proceedings of MICAI 2014, Part I, pp- 51–61.
- [12] Monica PALMIRANI, Raffaella BRIGHI, Matteo MASSINI. Automated Extraction of Normative References in Legal Texts. Proceedings of ICAIL 2003, pp. 105–106.
- [13] Jakub HARAŠTA, Jaromír ŠAVELKA. Toward Linking Heterogenous References in Czech Court Decisions to Content. Proceedings of Jurix 2017, pp. 177–182.
- [14] José Marín PÉREZ, Camino Rea RIZZO. Structure and design of the british law report corpus (BLRC): a legal corpus of judicial decisions from the UK. Journal of English Studies, 2012, vol. 10, pp. 131–145.
- [15] Mark DAVIES. Corpus of US Supreme Court Opinions. <https://corpus.byu.edu/scotus/>
- [16] Claire GROVER, Ben HACHEY, Ian HUGHSON. The HOLJ corpus: supporting summarisation of legal texts. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora, 2004, pp. 47–53.

- [17] Paula RODRÍGUEZ-PUENTE. Introducing the corpus of historical english law reports: structure and compilation techniques. *Revistas de Lenguas para Fines Específicos*, 2011, vol. 17, pp. 99—120.
- [18] Hanjo HAMANN, Friedemann VOGEL, Isabelle GAUER. Computer assisted legal linguistics (CAL2). Proceedings of JURIX 2016, pp. 195—198.
- [19] Tereza NOVOTNÁ, Jakub HARAŠTA. The Czech Court Decisions Corpus (CzCDC): Availability as the First Step. 2019, arXiv:1910.09513.
- [20] Marti A. HEARST. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 1997, vol. 23, no. 1, pp. 33—64.
- [21] Freddy CHOI, Peter WIEMER-HASTINGS, Johanna MOORE. Latent Semantic Analysis for Text Segmentation. Proceedings of the 6th Conference on EMNLP, 2001, pp. 109—117.
- [22] Oskari HEINONEN. Optimal Multi-Paragraph Text Segmentation by Dynamic Programming. Proceedings of the 17th COLING, 1998, pp. 1484—1486.
- [23] Jeffrey C. REYNAR. An Automatic Method of Finding Topic Boundaries. Proceedings of ACL'94 (student session), 1994, pp. 331—333.
- [24] Na YE, Jingbo ZHU, Haitao LUO, Huizhen WANG, Bin ZHANG, Improvement of the Dotplotting Method for Linear Text Segmentation. Proceedings of Natural Language Processing and Knowledge Engineering, 2005, pp. 636—641.
- [25] Murali SARAVANAN, Balaraman RAVINDRAN, Shivani RAMAN. Improving Legal Document Summarization Using Graphical Models. Proceedings of Jurix 2006, pp. 51—60.
- [26] Michael J. BOMMARITO II., Daniel Martin KATZ, Eric M. DETTERMAN. LexNLP: Natural language processing and information extraction for legal and regulatory texts, 2018, arXiv:1806.03688.
- [27] Zhenhao LI, Linxia YAO, Jidong GE, Chuanyi LI, Yuan YAO, Jin ZENG, Bin LUO and Victor CHANG. Word Segmentation for Chinese Judicial Documents. *Data Science [online]*. Singapore: Springer, 2019, pp. 466—478.
- [28] György GÖRÖG, Péter WEISZ. Legal entity recognition in an agglutinating language and document connection network for EU Legislation and EU/Hungarian Case Law. 2019, arXiv:1907.12280.
- [29] Jakub HARAŠTA, Jaromír ŠAVELKA, František KASL, Jakub MÍŠEK. Automatic Segmentation of Czech Court Decision into Multi-Paragraph Parts. *Jusletter IT*, 23. Mai 2019, pp. 1—11.
- [30] Regina BARZILAY, Michael ELHADAD. Using Lexical chains for Text Summarization. Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization, 1997, pp. 10—17.
- [31] Gonenc ERCAN, Ilyas CICEKLI. Using Lexical Chains for Keyword Extraction. *Information Processing & Management*, 2007, vol. 43, no. 6, pp. 1705—1714.
- [32] Violaine PRINCE, Alexandre LABADIÉ. Text Segmentation Based on Document Understanding for Information Retrieval. Proceedings of 12th International Conference on Applixations of Natural Language to Information Systems, 2007, pp. 295—304.
- [33] Erich SCHWEIGHOFER. The Role of AI & Law in Legal Data Science. Proceedings of Jurix, 2015, pp. 191—192.
- [34] Jaromír ŠAVELKA a Kevin D. ASHLEY. Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. *Frontiers in Artificial Intelligence and Applications*, 2018, pp. 111—120.
- [35] Jakub HARAŠTA, Jaromír ŠAVELKA, František KASL, Adéla KOTKOVÁ, Pavel LOUTOCKÝ, Jakub MÍŠEK, Daniela PROCHÁZKOVÁ, Helena PULLMANNOVÁ, Petr SEMENIŠÍN, Tamara ŠEJNOVÁ, Nikola ŠIMKOVÁ, Michal VOSINEK, Lucie ZAVADILOVÁ a Jan ZIBNER. Annotated Corpus of Czech Case Law for Reference Recognition Tasks. *Text, Speech, and Dialogue: 21st International Conference proceedings*, 2018, pp. 239—250.