2025 Volume 3, Issue 5
Educational Innovation Research
ISSN: 3029-1844(Print)
ISSN: 3029-1852(Online)

# Research on the Pathways and Challenges of Constructing Legal Corpus

**Chao Meng, Qinglin Ma**
School of Foreign Languages, Northwest University of Political Science and Law, Xi'an 710122, Shaanxi, China

**Abstract**

Legal corpus serves as the core data foundation for Legal AI, playing an increasingly important role in fields such as natural language processing, legal reasoning systems, intelligent legal question-answering platforms, and legal policy analysis. However, constructing high-quality, secure, and compliant legal corpus still faces numerous practical pathways and challenges. This paper systematically explores the multidimensional pathways for constructing legal corpus, including data source selection, the collaboration between manual and machine annotation, the standardized management of legal terminology, the intelligent processing framework for legal corpus, and the data integration mechanism for multiple institution collaboration. At the same time, this paper analyzes the main challenges faced during the construction of legal corpus, such as data quality and standardization, the identification and handling of legally sensitive content, and the ongoing adaptability to legal policies. The research suggests that a combined approach of technological empowerment and institutional guarantees can effectively enhance data quality, ensure compliance and security, and achieve intelligent management in the construction of legal corpus. Finally, the paper proposes future research directions and practical recommendations, aiming to provide theoretical guidance and practical support for the construction and application of legal corpus.

**Keywords**

Legal Corpus Construction
Multi-source Data Fusion
Intelligent Annotation
Legal Semantic Model

***Online publication:*** June 26, 2025

## 1. Introduction

### 1.1. Research background and significance

With the rapid development of artificial intelligence and big data technologies, legal corpus has become an important infrastructure for promoting legal intelligence. Legal corpus is not only used for NLP tasks such as text classification, entity recognition, and sentiment analysis, but also plays a key role in training legal reasoning models, intelligent legal question-answering, and legal policy prediction. To ensure the effective operation of these application scenarios, constructing a high-quality, secure, and compliant legal corpus has become a core issue in legal technology research [1].

### 1.2. Research issues and challenges

The construction of legal corpus faces numerous practical challenges. First, the diversity and complexity of legal texts pose a major dilemma. Legal documents typically include various types such as legal provisions, judicial documents, administrative files, policies and regulations, and legal consultation records, each with significant differences in data collection, annotation, and management methods. How to achieve unified processing and storage of cross-type legal texts has become one of the key issues in constructing the legal corpus [2].

Second, the annotation process of legal corpus has long relied on manual efforts, leading to low annotation efficiency and the potential for subjective bias. Although existing research has explored automated annotation methods based on natural language processing (NLP), these technologies still have certain limitations when faced with the semantic complexity of legal texts and the frequent changes in legal terminology. Therefore, establishing a collaborative mechanism between manual and machine annotation, as well as constructing a high-quality legal corpus annotation system, is an important task in the current construction of legal corpus.

### 1.3. Research objectives and innovations

The research objectives of this paper are mainly threefold:

(1) Systematically outline the construction path of legal corpus: by analyzing the core components of legal corpus (such as data collection, annotation management, privacy protection, and compliance governance), clearly define their technical implementation and management processes to provide scientific guidance for the construction of legal corpus.

(2) Reveal the main challenges in legal corpus construction: by addressing challenges related to data standardization, privacy protection, annotation accuracy, and legislative dynamics, this paper proposes corresponding categories of issues and analyzes their impact on the construction of legal corpus.

(3) Explore solutions and technical support: by combining cutting-edge technologies such as natural language processing, machine learning, big data analysis, and blockchain, this paper proposes strategies for constructing legal corpus and discusses their application prospects in legal intelligent systems[3].

The innovations of this research lie in the following aspects:

(1) Proposing multi-source data fusion pathways: unlike traditional single-source data collection methods, this paper emphasizes the cross-institutional, multi-type, and multilingual integration of legal corpus to address the uneven sources of legal texts.

(2) Constructing an intelligent annotation system: based on the combination of natural language processing and deep learning, this paper proposes an annotation system that balances human intervention and automated processing, effectively improving the annotation accuracy and efficiency of legal corpus.

(3) Designing a privacy computing and compliance governance framework: by introducing privacy computing technologies such as differential privacy and federated learning, along with governance mechanisms driven by legal policies, this paper aims to achieve intelligent, secure, and compliant management of legal corpus.

## 2. Path to Constructing Legal Corpus

### 2.1. Data collection and multi-source integration

The construction of legal corpus primarily relies on a stable and diverse data collection system. Currently, the sources of legal corpus mainly include legal documents, judicial rulings, legal consultation records, legal journals, and legal literature. Legal texts from different sources have varying linguistic expressions and text structures, making unified management and integration a key aspect of constructing a legal corpus.

To achieve the integration of multi-source legal corpus, this paper proposes a cross-source data collection and preprocessing mechanism. First, the data collection system needs to interface with multiple legal databases, such as the China Judgments Online, legal consulting platforms, government policy disclosure systems, and legal literature from academic publishers. Second, due to the lack of uniformity in legal text formats, the collected data needs to undergo standardization processes, including the removal of irrelevant information, unification of terminology definitions, and adjustment of text structures to facilitate subsequent annotation and semantic analysis.

## 2.2. Annotation system and human participation mechanism

Annotation is an indispensable part of the construction process of legal corpus; it not only determines the quality of the corpus but also affects the training effectiveness and reasoning capabilities of legal AI models. The annotation content of legal corpus typically includes aspects such as recognition of legal entity, reasoning of legal relationship, classification of legal events, and segmentation of legal texts. However, the complexity of legal text content makes annotation challenging, and traditional annotation methods often rely on manual annotation, which is inefficient and carries the risk of subjective bias[4].

To improve the efficiency and accuracy of legal corpus annotation, this paper proposes a hybrid annotation system that combines manual and automatic annotation. First, by introducing pre-trained language models (such as BERT, RoBERTa, etc.), the system can automatically perform preliminary annotations on legal texts, providing a reference for subsequent manual review and optimization.

## 2.3. Data standardization and terminology unification

Data standardization and legal terminology unification are another important aspect of constructing a legal corpus. Due to the diverse sources of legal texts, legal terminology may vary across different institutions and regions, affecting the consistency and scalability of the corpus. Therefore, the construction of the legal corpus must ensure the uniformity of data structure and terminology definitions to support the training and reasoning of legal NLP models.

This paper proposes a method for unifying legal terminology, combining legal knowledge graphs and natural language processing techniques to achieve automatic recognition and unified mapping of legal terminology. First, based on the analysis of legal texts, the system can identify common legal terminology and construct a legal terminology graph.

## 2.4. Intelligent analysis and legal semantic model construction

The construction of a legal corpus not only requires high-quality data and accurate annotations but also necessitates the development of intelligent legal semantic models to support semantic understanding and legal analysis capabilities of legal texts. The application of legal semantic models can significantly enhance the intelligence level of the legal corpus in scenarios such as natural language processing, legal reasoning, and legal queries.

This paper proposes a framework for constructing legal semantic models based on deep learning and natural language processing. First, through the automatic preprocessing and feature extraction of legal texts, the system can provide accurate input data for the legal semantic model. Second, by utilizing deep learning models (such as Transformer, BERT, RoBERTa, etc.) to perform semantic analysis and semantic modeling of legal texts, it can acquire the ability for legal knowledge reasoning and understanding of legal texts.

## 2.5. Multiple institution collaboration mechanism and legal data sharing

The construction of a legal corpus often requires the collaborative participation of multiple legal institutions, including courts, procuratorates, law firms, legal research institutions, and government regulatory departments. There are differences in data structures, collection standards, and annotation methods among different institutions. Achieving data integration from multiple institutions while ensuring data quality is one of the significant challenges in constructing a legal corpus.

To this end, this paper proposes a multiple institution collaboration mechanism based on legal data sharing standards. First, by formulating a unified legal corpus sharing agreement, it ensures that the data access methods of different institutions are consistent and meet data protection and legal compliance requirements. Second, by adopting a decentralized data sharing model, such as Federated Learning and Privacy Computing technologies, it achieves cross-agency data sharing while ensuring data privacy and security.

# 3. Major Challenges in Constructing a Legal Corpus

## 3.1. Data quality issues

The data quality of legal corpus directly affects the training effectiveness and reasoning ability of legal AI models. However, the current construction of legal corpus faces issues such as low data quality, semantic inconsistency, and inaccurate annotations. First, during the data collection process, there may be problems such as data loss, data duplication, and data noise, which directly impact the comprehensiveness and accuracy of the legal corpus [5].

## 3.2. Complexity of legal terminology and annotation difficulty

The diversity and complexity of legal terminology make the annotation work of legal corpus particularly challenging. On one hand, legal terminology often has specific legal definitions and application scenarios; certain terminology may belong to a particular legal domain or a specific type of legal content, and these differences make standardization and unified annotation a challenge. On the other hand, there are significant differences in the semantic structures and expressions of different legal texts, leading to difficulties in maintaining consistency during the annotation process.

Therefore, accurately identifying legal terminology during the construction of legal corpus and establishing an annotation system that can adapt to changes in legal terminology is one of the core challenges faced in the construction of legal corpus.

## 3.3. Personalized data governance and privacy protection issues

The personalized governance capability of legal corpus determines its balancing strategy between public data and sensitive data. On one hand, legal corpus need to support various data governance models to meet the requirements of different legal systems and policies. For example, China's Personal Information Protection Law imposes strict management requirements on privacy information in legal corpus, while the European Union's General Data Protection Regulation (GDPR) requires stricter controls on query and usage permissions for legal data. How to achieve personalized data governance and privacy protection in the construction of multi-jurisdiction legal corpus has become an important topic in the design of legal corpus governance mechanisms.

These privacy and compliance issues not only affect the scope of use of legal corpus but also relate to the traceability and auditability of legal data.

## 3.4. Training needs of annotation consistency and legal AI model

The consistency of annotations in legal corpus directly affects the training effectiveness and reasoning ability of legal AI models. However, there is still significant inconsistency in the annotation work of current legal corpus, mainly reflected in the following aspects:

(1) Large differences in the professional backgrounds of annotators: different annotators may have varying interpretations of legal terminology, leading to inconsistent annotation of the same legal term across different corpora.

(2) Ambiguity and context dependence of legal terminology: certain terminology in legal texts may have multiple layers of meaning and are closely related to context, resulting in different outcomes for annotation tasks in different contexts.

(3) Lack of unified annotation standards: due to the diversity of legal corpus, annotation rules may vary depending on the source of the text, leading to inconsistencies in the training data for legal AI models, which affects the model's generalization ability and reasoning accuracy.

## 3.5. Classification of legal texts and adaptation to legal AI scenarios

The classification of legal texts is a key step in the construction of legal corpus, and its classification results directly affect the application effectiveness of legal AI models. Legal texts typically include civil judgments, criminal judgments, administrative cases, intellectual property texts, and legal consultations. Different types of legal texts exhibit significant differences in semantic expression, structural characteristics, and application scenarios, which poses certain challenges for the classification task[6].

# 4. Solutions and Technical Support

## 4.1. Multi-source data fusion and standardization processing path

The construction of legal corpus requires the integration of legal texts from different institutions and of different types, which places high demands on data fusion and standardization. However, there is still a widespread phenomenon of data silos in current legal corpus, making it difficult to meet the demand for high-quality legal texts across data sources for legal AI systems.

To address this issue, this paper proposes a multi-source legal data fusion path, which mainly consists of the following steps:

(1) Cross-institution data access mechanism: first, design a cross-institution data access system to efficiently aggregate legal texts from different sources such as courts, law firms, legal research institutions, and government agencies.

(2) Unified processing of legal texts: second, by means of cleaning, deduplication, and format standardization of legal texts, unify the representation of legal texts to enhance the generality and scalability of the legal corpus.

(3) Unified mapping mechanism for legal terminology: furthermore, construct a legal terminology graph to achieve unified mapping and definitions of legal terminology, making the use of terminology in legal texts more standardized and avoiding inconsistencies in the corpus content due to terminology differences.

## 4.2. Intelligent annotation and manual review mechanism

The annotation of the legal corpus is a key step in constructing high-quality legal text data, and its effectiveness directly impacts the training accuracy and reasoning ability of legal AI models. However, the annotation of legal corpus currently faces many challenges, such as low annotation efficiency, poor consistency, and high costs [7].

## 4.3. Privacy protection and data compliance framework

In response to the privacy data issues in the legal corpus, this paper proposes a data management framework centered on privacy computing and data compliance. This framework aims to ensure the security of legal data during the processes of collection, storage, usage, and sharing through privacy protection technologies.

## 4.4. Legal semantic modeling and adaptation to legal AI application scenarios

The construction of the legal corpus must not only consider data quality and annotation accuracy but also the adaptability of legal semantic modeling to application scenarios. Currently, the application of legal semantic models in scenarios such as legal reasoning, legal prediction, and legal consulting services is significantly influenced by the quality of legal texts.

To address this, this paper proposes a path for adapting legal semantic modeling to legal AI applications. First, by utilizing natural language processing (NLP) and deep learning technologies, this paper builds a legal semantic model to accurately understand and reason about legal texts. Second, based on different legal AI application scenarios (such as legal question-answering systems, legal analysis tools, and legal prediction platforms), the structure and training methods of the legal semantic model are adjusted to enhance its adaptability.

## 4.5. Multiple institution collaboration construction and legal data sharing model

Multiple institution collaboration construction is one of the important paths for constructing the legal corpus, but how to achieve cross-institutional data sharing and collaborative governance remains a practical challenge. On one hand, legal corpus needs to involve multiple legal institutions for data collection and annotation to improve data comprehensiveness and accuracy; On the other hand, there are differences in data collection standards and annotation methods among different legal institutions, leading to issues such as inconsistent data formats and difficulties in data integration between legal corpora.

# 5. Conclusion and Outlook

## 5.1. Research conclusions

This paper systematically explores the construction paths and main challenges of legal corpus, proposing a comprehensive solution suitable for the construction of legal corpus. The construction of a legal corpus involves multiple aspects such as data collection, annotation management, terminology standardization, privacy protection, data compliance, and multiple institution collaboration, each of which is crucial for the quality and secure use of the legal corpus.

Research findings indicate that the core pathways for constructing a legal corpus include: multi-source data collection and standardized processing, a combination of intelligent annotation and manual review, unified management of legal terminology, establishment of legal privacy computing and compliance governance mechanisms, as well as multiple institution collaboration construction and the traceability of legal data flows. These pathways can effectively enhance the quality, security, and intelligent application value of the legal corpus, providing solid data support for the construction of legal AI systems[7].

## 5.2. Practical application value

The construction and optimization of the legal corpus hold significant practical implications for the development of legal technology. As legal AI technology continues to mature, the legal corpus will become an important foundation for legal intelligent systems. However, the current legal corpus still has many deficiencies, such as low data quality, untimely updates, and poor annotation consistency, which severely affect the training effectiveness and reasoning capabilities of legal AI models.

The pathways and solutions proposed in this paper for constructing a legal corpus can effectively address the main issues currently faced by legal corpus. For example, by introducing a multi-source data fusion mechanism, the legal corpus can achieve more comprehensive data collection; by constructing a legal terminology graph and a unified annotation rule scheme, the legal corpus can adapt to changes in legal terminology and improve annotation consistency; through privacy computing and blockchain technology, the legal corpus can achieve secure data management; through a multiple institution collaboration construction mechanism, the legal corpus can achieve cross-institutional data sharing and the integration of legal liability elements.

**Funding**

This paper is a phased outcome of the National Social Science Fund Project “Translation and Construction of a Chinese-English Parallel Corpus of the Regulations and Rules of the Communist Party of China and the Regulations of Major Political Parties in the World” (Project No.: 19XYY014).

**Disclosure statement**

The author declares no conflict of interest.

## References

[1] Jiang H, 2025, Intelligent Auxiliary Judgment of Legal Corpus Technology and the Literal Meaning of Criminal Law. Journal of Jiaotong University Law, (03): 137-150. DOI: 10.19375/j.cnki.31-2075/d.2025.03.004.

[2] Song L, 2023, Linguistic Data Foundation, Methods, and Applications of Digital Jurisprudence: Taking the Birth and Development of Legal Corpus Linguistics as an Example. Eastern Law, (06): 118-129. DOI: 10.19404/j.cnki.dffx.20231116.004.

[3] Yuan Y, Cui Y, Sun J, et al., 2023, How to Build a Legal Specialized Corpus for Research on Factual Expression? Contemporary Rhetoric, (02): 16-28. DOI: 10.16027/j.cnki.cn31-2043/h.2023.02.009.

[4] Wu S, Li J, 2025, A Critical Cognitive Analysis of Judges’ Reported Speech in Judicial Opinions Based on Corpus

Linguistics. Foreign Language Teaching, 46(04): 25-32. DOI: 10.16362/j.cnki.cn61-1023/h.2025.04.003.

[5] Brian G. Slocum, Stephen TH. Grace, Gu R, 2023, Evaluating Corpus Linguistics in Legal Contexts. Legal Method, 44(03): 95-108.

[6] Tang Y, Yang Y, 2017, A Study on the Quality of English Translation of Chinese Legal Texts from the Perspective of Lexical Chunk Theory-Based on a Bilingual Legal Corpus. Chinese Science and Technology Translation, 30(03): 41-44. DOI: 10.16024/j.cnki.issn1002-0489.2017.03.012.

[7] Xu J, Wang Q, 2017, Analysis of the Current Situation of Legal Translation Research Based on Corpora: Problems and Countermeasures. Foreign Language Research, (01):73-79. DOI:10.16263/j.cnki.23-1071/h.2017.01.013.

**Publisher' s note**

*Whioce Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.*