ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Construction and security framework of Chinese Characteristic Cultural Tourism English Corpus in digital age

Zhuo Li[1,*]

[1] Daqing Open University, Daqing, Heilongjiang, 163311, China

SUMMARY: *This study mainly explores the scheme selection and security architecture design of a Chinese cultural tourism exclusive English corpus under the background of modern technology. It combines corpus linguistics theory with cultural tourism translation practice and contemporary data security control requirements, and systematically analyzes the key issues of corpus construction and security protection. The sampling approach is systematic based on a 10 million word corpus that covers various text types, regions, cultures, and countries. A multi-dimensional classification system organises content into text type, geographical region, cultural domain, period, and various linguistic features. The corpus uses culture-specific item annotation discriminatory id markers, which enable a systematic synthesis of translation processes that facilitate culture-specific examinations. The security model includes an all-encompassing data classification scheme with layered multi-tiered safeguarding mechanisms, authentication structure, protective encoding strategies, and compliance regulations. Mathematical models have been developed to measure the representativeness of the corpus, sensitivity of the culture, and security risk assessment. This merges practical concerns in tourism translation with theoretical developments in corpus linguistics. The created corpus offers opportunities for improving cross-cultural discourse while extending necessary safeguards for vulnerable material, developing a foundation whose principles can be tailored for analogous endeavours in specialised fields.*

KEYWORDS: *corpus linguistics, cultural tourism, translation studies, data security, Chinese cultural*

1 Introduction

In recent years, the construction of specialised linguistic corpora has become increasingly important for translation and cultural studies in the context of interdisciplinary communication in the digital era. The creation of a Chinese Cultural English Corpus is a landmark accomplishment in the linguistic and cultural integration of China with the English-speaking world. The tourism sector, as a vital gateway for culture and economic exchange, calls for precision and sensitivity in translation for conveying culture to foreign visitors and articulating culture to China and the Chinese heritage to the world [1].

The international tourism market is on the rise, as seen by the speedy growth of inbound tourism subsequent to increased international communications, with China landing on the world tourist map and receiving numerous foreign visitors every year. This is demonstrating a growing need for specialised tourism-related language materials that reflect cultural subtleties. Francesconi remarks that the tourism language is, indeed, a specialised discourse that needs

*apply22318@163.com
https://doi.org/10.65102/is20261087 Published: 10 June, 2026

corpus-based studies for contrasts and comparisons of cultures and languages to translations [2]. Enhancements in technology as part of the new age can be a double-edged sword: worries of data violation add complexity to security.

The idea of constructing a Chinese Characteristic Cultural Tourism English Corpus arises from understanding that tourism discourse is a type of specialised language with specific communicative and stylistic features. Researchers can trace discursive and stylistic patterns of translated tourism language through the analysis of translated tourist texts against texts in English and text entailment using machine learning techniques, an understanding of its effects on communicative functions and persuasive effects [3]. This method can systematically reveal the language development patterns and cultural expression characteristics of Chinese characteristic tourism text corpus, providing theoretical support for cultural tourism external publicity translation, corpus construction, and cross-cultural communication research.

Tourism is a human activity with distinct cross-cultural attributes, which has given rise to a series of special problems in the translation practice of tourism related texts. The translation of tourism related texts should not only be faithful to the literal meaning of the text, but also focus on optimizing the tourist experience as the core goal, which puts higher demands on cultural and tourism translation. In addition, cultural and tourism translation also needs to generate cultural identity and resonance among overseas tourists at a higher level. Currently, there is an urgent need to construct a cultural and tourism translation strategy that is adapted to the cultural context [4]. This relies on the creation of a specialized corpus for cultural and tourism translation, which requires both accurate language and complete contextual reference systems, as well as effective communication and interaction with foreign tourists at the cultural level.

In the digital age, the construction of a professional cultural and tourism English corpus must be accompanied by a comprehensive security protection mechanism, especially for sensitive information, effective protection strategies need to be established [5]. With the continuous improvement of governance frameworks and regulatory systems such as the Data Security Law of the People's Republic of China and the Personal Information Protection Law of the People's Republic of China, the mode of data processing and protection has been fundamentally reshaped. In the entire process of corpus construction, compliance requirements at the data governance level must be implemented to ensure the legality, standardization, and integrity of the entire lifecycle of corpus data collection, storage, annotation, use, and sharing.

This study focuses on the design and systematic research of the Chinese cultural tourism English corpus, constructing an integrated security protection model that includes method paths, practical obstacles, and application scenarios. The basic principles of corpus linguistics are deeply integrated with cultural tourism translation practice and contemporary data security compliance requirements, expanding its theoretical perspective and application dimensions. Although tourism translation has formed a relatively independent branch discipline, most research systems are still incomplete. This study is based on a standardized and structured corpus, optimizing the existing research framework and practical system [6], providing theoretical support for the standardized, intelligent, and secure development of Chinese cultural tourism translation.

2 Literature Review

In the past few decades, the blending of corpus linguistics, tourism discourse, and translation studies has become the focus of considerable academic interest. This review looks into the

primary theoretical aspects and available literature regarding Practices of Building Security Frameworks for a Corpus of English for Chinese Cultural Tourism.

2.1 Theoretical Foundations of Corpus Linguistics

Through the empirical grounding of texts, or collections of naturally occurring texts, corpus linguistics has transformed the study of language. While corpora, or textual data sets, have always existed in some fashion, advancements in computation and technology have made constructing and consulting corpora faster and easier; as a result, corpus data is now a primary resource for many linguists [7]. This methodological approach provides a unique advantage for translation and language studies.

The field has evolved from emphasising descriptive methodologies to more multi-faceted theoretical and applied approaches in translation studies. The development of these disciplines is limited to a lack of a cohesive analytical framework, which allows researchers to ignore important elements to focus on subjective factors, thus yielding diverse interpretations from shared datasets [8]. This speaks to the need for more logical corpus-based methodological frameworks in specialised disciplines, such as tourism translation.

More specifically, for corpus studies on the Chinese language, there is a significant increase in corpus-based research within the sphere of contemporary linguistics. Chinese scholars acknowledge that corpora offer a statistically substantial basis for linguistic analysis in comparison to traditional approaches [9]. This reflects a shift in the field towards data-centric methodologies across the globe, which is increasingly important for specialised domains.

2.2 Tourism English Corpus Development

As this form of interaction has its own vocabulary, cultural implications, and communicative intents, it is subsumed under the umbrella of a particular system, which is sociocultural tourism discourse. The use of corpora in teaching English for Specific Purposes (ESP) is beneficial, to some extent, despite the challenges faced by both the learners and instructors. The so-called "data-driven learning" enables students to identify patterns of language use in real-life contexts, thus becoming independent in catering to their linguistic needs. This educational focus indicates one of the numerous examples of the utilisation of tourism corpora, which transcends research purposes.

This approach acknowledges the professional and non-professional layers in tourism discourse: some researchers have already dealt with the language of travel blogs and trip reports, analysing how the language constructs tourist phenomena [10]. These tourism overviews represent the beginning of an interesting line of research.

In the case of translations of texts in Chinese and English, researchers have pointed out a range of issues that include scientific terminology, stylistic conventions of the text, and cultural elements that require specialised translations [11]. These considerations highlight the need for more corpus-based methods, which attempt to understand and explain these issues systematically through the empirical analysis of authentic texts.

2.3 Cultural Translation Theories in Tourism Contexts

Different aspects of culture tend to pose certain specific problems in the area of tourism translation due to their deep-rooted nature and possible misconceptions. Translation in tourism involves cultural mediation which is particularly important as nations begin to welcome foreign tourists and allow them access into their cultures through translation [12]. This mediation function emphasizes the need to create based resources which are capable of pinpointing culture-specific issues and treating them adequately.

The notion of culture-specific items has been widely discussed within the framework of translation studies. Studies on culture-specific item translation show the use of various strategies by translators such as neutralisation, foreignisation, and even domestication driven by the culture's markedness and the polysemous nature of the source item. Those are reasons why there is a need for developing corpus resources which address the need to articulate common strategies for elements of culture in the context of tourism [13].

Some researchers have developed specific theories aimed at providing guidance on cultural translation for the purpose of tourism. Hu Gengshen's Eco-Translatology theory, which contains elements of "natural selection and adaptation" taken from the works of Darwin as well as "focus on human" and "harmony between man and nature" derived from Chinese philosophy, provides a complete answer for the linguistic, cultural, and communicative problems of tourism translation [14]. Such an approach has permitted offering principles for the design and annotation of the corpus that will address the cultural aspects of the body.

2.4 Digital Technologies in Corpus Construction

The advancement of corpus linguistics has come as a result of the digitisation of data collection, annotation, and analysis processes. The merging of web-based resources along with data tracking and analysis systems has opened new avenues for the development and application of specialised corpora [15]. These advances in technology greatly assist in developing well-structured tourism corpora of higher efficacy.

Tools and approaches offered by computational linguistics have aided in the creation of specialised corpora. The incorporation of computational linguistics into corpus-based studies of Chinese translation has advanced the systematic study of translation phenomena, along with the development of machine translation processes [16]. Such approaches can be very beneficial in tackling the scale and complexity associated with the construction of specialised tourism corpora.

The digitisation of materials allows for the use of additional modes alongside text to be integrated into corpora. The study of linguistic landscapes in the context of tourism demonstrates how visual and textual elements work together to synthesise diverse identities and convey cultural elements [17]. Such a multimodal approach reinforces the need to create corpus resources that go beyond just textual materials prominent in tourism and include visuals as well.

2.5 Security Issues in Corpus Development

The protection of sensitive information has become important in the construction of a corpus, especially in cases where custom content that contains private cultural or proprietary information is included. China's efforts to create a legal framework for data protection began when the Personal Information Protection Law and the Data Security Law were issued in 2021, forming an integrated system that governs data circulation within China as well as influencing other countries' legislations on data protection laws [18]. These changes focus primarily on corpus development practices, which demand sophisticated security arrangements.

Developing a corpus with integrated features demands affirmative action on the protection of intellectual property rights. While corpus development tracks user participation through the corpus resources, privacy and user-generated content integration come into play [19]. The above highlights the balance of ethical and technical issues pertaining to the development of a corpus.

The development of a corpus is now microbiologically global in its analytical attempt for the application of legislation on data governance. In the European Union and China, comparative data governance law contains shared interests in data protection and distinct issues

such as cross-border data transfer [20]. These borderless considerations are crucial when it comes to corpus projects with international collaboration or data sharing.

2.6 Research Gaps in Tourism Corpus Development

While notable strides have been made in the corpus linguistics and translation studies of tourism, developing specialised Chinese cultural tourism corpora has gaps which require further research. Culturally mediated Chinese tourism translation still eludes scholarly attention, and this indicates possibilities for systematic corpus-based research to fill this gap.

Developing security frameworks alongside corpus-building techniques creates yet another persistent gap. Recent legislation, including the Data Security Law in China, introduces novel data handling requirements that have yet to be fully met by the existing corpus development practices [21]. This illustrates the lack of adequate linguistic-dominated interdisciplinary frameworks with data security that address these issues.

Though there is potential for employing corpus linguistics as authentic linguistic evidence to support systematised analysis of translation phenomena, addressing interdisciplinary tourism translation with corpus linguistics remains insufficiently developed. This marks the need for integrating more comprehensive theories alongside corpus linguistics, tourism studies, and cultural translation.

The review conducted demonstrates that there have been fundamental works in developing the history and theory linking computational linguistics, tourism translation, and information security, but as synthesised in a practical corpus, these domains have yet to be integrated. This is a gap which the current research seeks to fill through devising a systematic plan for building a corpus of English for tourism of Chinese characteristic culture with the needed securitisations.

3 Methodology

3.1 Corpus Design Principles and Parameters

The Chinese Cultural Tourism English Corpus adopts a balanced sampling framework, which can ensure comprehensive coverage and sample representativeness of tourism discourse types. The corpus design needs to strictly follow the key parameters listed in Table 1, with the aim of constructing a quantifiable and operable data collection and organization index system, providing unified standards for corpus screening, classification, annotation, and storage.

Table 1: Corpus Design Parameters for Chinese Characteristic Cultural Tourism English Corpus

Parameter	Specification	Rationale
Size	10 million words	Ensures statistical significance while maintaining manageability
Temporal range	2015-2025	Captures contemporary linguistic patterns
Text types	Website text accounts for 30%, promotional brochures account for 20%, museum explanatory texts account for 15%, heritage site explanatory texts account for 15%, travel blogs account for 10%, and social media content accounts for 10%	Balances institutional and user-generated content
Regional distribution	30% in the eastern region, 20% in the southern region, 20% in the northern region, 15% in the western region, and 15% in the central region	Reflects geographical diversity
Cultural domains	Tangible cultural heritage accounts for 40%, intangible cultural heritage accounts for 30%, natural heritage accounts for 15%, and contemporary culture accounts for 15%	Encompasses diverse cultural offerings

The overall method framework for corpus construction integrates a multi-stage construction process and a closed-loop feedback mechanism to ensure the quality and coherence of the corpus throughout the entire process, as shown in Figure 1.

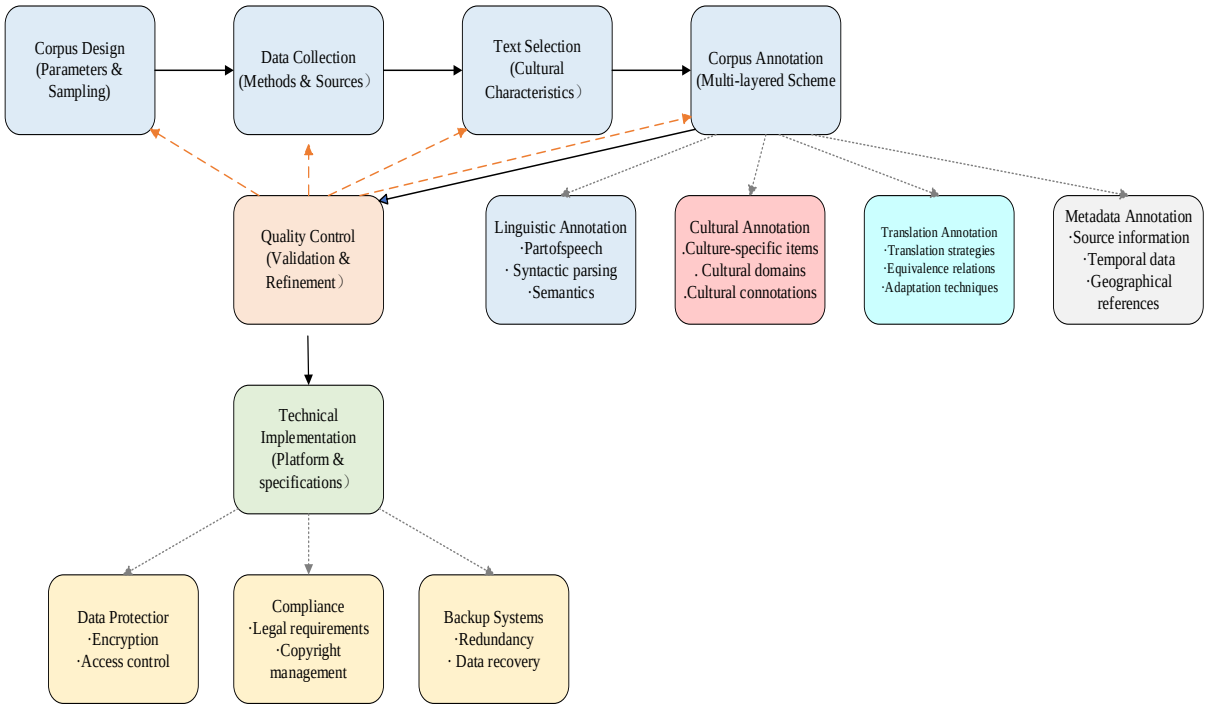

Figure 1: Methodological Framework for Chinese Characteristic Cultural Tourism English Corpus Construction

As shown in Figure 1, the construction method of the cultural tourism translation corpus proposed in this study follows an iterative optimization process. Throughout the entire development cycle of the corpus, a multi-level annotation system and data security framework components are synchronously integrated to achieve the integration of language annotation, cultural annotation, and security control, thereby ensuring the unity of language standardization and data security in the corpus.

3.2 Data Collection Methods and Sources

The data collection process of the cultural and tourism translation corpus adopts a combination of automatic web crawlers and targeted manual collection. The sources of cultural and tourism translation corpus cover official tourism websites, digital archives, museum literature, tourism publications, social media platforms, and professional cultural and tourism blogs. The data collection process of the cultural and tourism translation corpus strictly follows standardized protocols, fully recording the source, collection time, metadata information, and usage authorization status of the cultural and tourism translation corpus, ensuring that the acquisition of cultural and tourism translation corpus complies with academic ethics and copyright regulations.

3.3 Text Selection Criteria with Chinese Cultural Characteristics

The quality of cultural and tourism translation texts is verified using a multidimensional evaluation framework, which comprehensively evaluates them from three core dimensions: cultural specificity, language authenticity, and representation accuracy. The multidimensional

evaluation indicators and scoring criteria for cultural and tourism translation texts selected in this study are shown in Table 2.

Table 2: Evaluation Framework for Texts with Chinese Cultural Characteristics

Evaluation Dimension	Assessment Criteria	Scoring Range
Cultural density	Frequency of culture-specific references per 1000 words	1-5
Cultural depth	Extent of explanatory content for cultural concepts	1-5
Cultural accuracy	Degree of factual correctness in cultural representations	1-5
Translation quality	Effectiveness of rendering cultural concepts in English	1-5
Target audience awareness	Adaptation to knowledge level of international visitors	1-5

Texts achieving a minimum composite score of 3.5 qualify for inclusion, ensuring quality while maintaining representativeness.

3.4 Corpus Annotation Scheme and Procedures

This study adopts a multi-level integrated annotation framework, covering four major modules of language annotation, cultural annotation, translation annotation, and metadata annotation for cultural and tourism translation texts: (1) The language annotation module mainly includes part of speech tagging and syntactic structure analysis of translated texts; (2) The cultural annotation module mainly focuses on the cultural exclusive concepts and cultural domain identification of translated texts; (3) The translation annotation module mainly records the translation strategy and semantic correspondence of the translated text; (4) The metadata annotation module is mainly used to annotate the corpus source, genre, region, theme and other attributes of the translated text. The annotation process of cultural and tourism translation texts adopts a mixed mode of automated processing and expert manual verification. The annotation format follows the TEI text encoding standard and is specially extended based on the characteristics of cultural and tourism texts.

3.5 Quality Control and Validation Processes

The quality assurance system for the constructed cultural and tourism translation corpus adopts a multidimensional control mode combining automatic verification, expert evaluation, and user testing, and adopts two methods: quantitative verification and qualitative verification: (1) quantitative verification stage. Using statistical methods to measure the consistency and error rate of annotation in cultural and tourism translation texts; (2) Qualitative verification. Professional evaluation is conducted by experts in the fields of culture and language to ensure the cultural adaptability and language standardization of cultural and tourism translation materials. The specific content and indicators of the multi-stage quality control process selected are shown in Table 3.

Table 3: Multi-stage Quality Control Process for Corpus Development

Stage	Validation Procedure	Quality Indicators	Acceptance Threshold
Pre-collection	Source evaluation	Authority, reliability	Minimum score 3.5/5
Collection	Data integrity verification	Completeness, attribution	<2% error rate
Annotation	Inter-annotator agreement	Cohen's kappa coefficient	>0.8
User	User satisfaction	Usefulness, intuitiveness	>85% approval

3.6 Technical Specifications and Platform Selection

The Chinese cultural tourism English corpus constructed in this study adopts a three-tier architecture design: data storage layer, data processing layer, and application interface layer. Among them, the data storage adopts a hybrid SQL/NoSQL database architecture, which balances the efficiency of structured data queries and the flexibility of unstructured corpus storage. The platform security system strictly follows the ISO/IEC 27001 information security management standard and deploys role-based access control (RBAC) and multi factor authentication system to enhance user login security. On the hardware support level, a high-performance server equipped with a 64 core processor, 512GB of memory, and 500TB of storage capacity can support the storage, annotation, retrieval, and concurrent access of massive corpora. Adopting a construction model that combines customized development with standard tool integration to ensure the platform's professionalism, scalability, and practicality. The optimization indicators for dynamic allocation of corpus system resources are as follows:

$$R = \frac{C \times A \times S}{T} \tag{1}$$

where, C represents computational efficiency, R represents platform responsiveness, S represents storage performance, A represents algorithm optimization level, and T represents transmission overhead.

4 Construction of Chinese Characteristic Cultural Tourism English Corpus

4.1 Corpus Architecture and Structural Design

The Chinese Cultural Tourism English Corpus (CCCTE) is built using a modular architecture and a hierarchical structure design. Each level module operates independently and collaboratively, ensuring efficient data access while strengthening security control capabilities. The specific structure of CCCTE is shown in Figure 2.

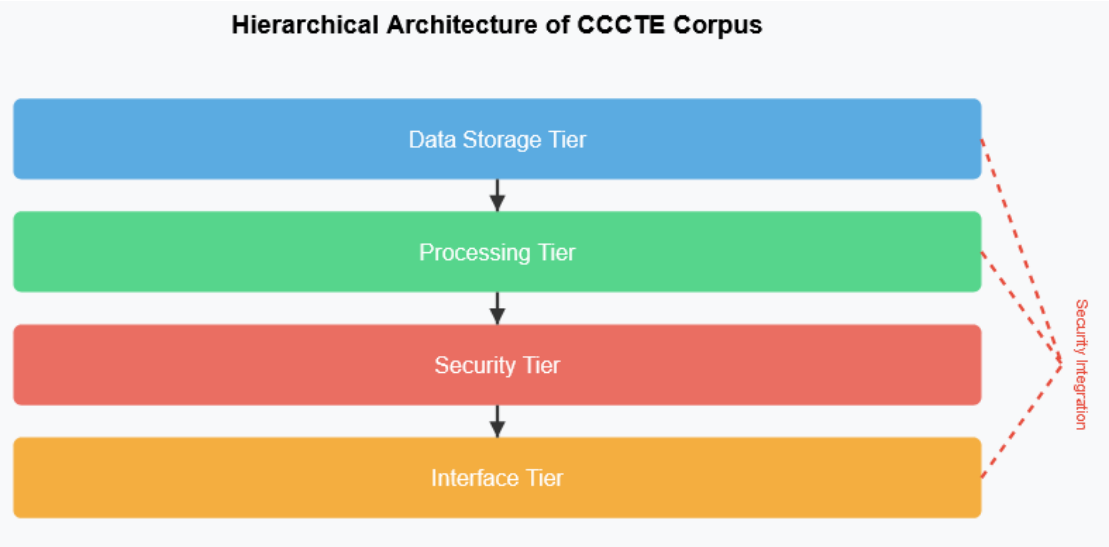

Figure 2: Hierarchical Architecture of CCCTE Corpus

As shown in Figure 2, the architecture design of CCCTE corpus consists of four main levels, adopting the design principle of "independent operation and collaborative linkage": (1) data storage layer. Adopting a distributed database system equipped with cultural classification custom indexes, balancing storage efficiency and classification accuracy; (2) Data processing layer. Integrate a natural language processing (NLP) module specifically designed for the cultural tourism field, which can complete specialized processing tasks such as automatic annotation of corpus, syntactic analysis, and recognition of cultural loaded words; (3) Defense layer. Strictly implement various security protection measures and build a comprehensive and multi-level security protection barrier; (4) Interface layer. Design differentiated access control interfaces for different user types such as researchers and translators to achieve refined management of user permissions.

The optimization objective function used is as follows:

$$A(t) = \lambda_1 \cdot P(t) + \lambda_2 \cdot S(t) + \lambda_3 \cdot E(t) \tag{2}$$

where $A(t)$ represents the architectural efficiency at time t, $P(t)$ denotes processing performance, $S(t)$ indicates security robustness, and $E(t)$ represents extensibility capacity. The coefficients λ_1, λ_2, and λ_3 are adjusted based on operational priorities and usage patterns.

4.2 Text Categorization and Classification System

The CCCTE corpus has implemented a multidimensional classification system based on multiple core parameters related to cultural tourism discourse, which can effectively organize and classify cultural tourism related texts in the corpus. Table 4 provides an overview of the main classification dimensions and their corresponding subcategories.

Table 4: CCCTE Corpus Classification Dimensions

Classification Dimension	Primary Categories	Subcategories	Coding System
Text Type	Official documents, Promotional materials, Interpretive texts, User-generated content	Website content, Brochures, Signs, Social media posts	TT-XX-YY
Geographical Region	Eastern, Southern, Western, Northern, Central	Provincial, Municipal, Site-specific	GR-XX-YY
Cultural Domain	Tangible heritage, Intangible heritage, Natural heritage, Contemporary culture	Architectural, Culinary, Performing arts, Festivals	CD-XX-YY
Temporal Period	Ancient, Imperial, Modern, Contemporary	Dynasty-specific, Republic era, Post-1949	TP-XX-YY
Linguistic Features	Translation strategies, Cultural terms, Specialized vocabulary	Domestication, Foreignization, Specialized terminology	LF-XX-YY

The classification system can effectively improve the convenience of corpus content retrieval. Each text is assigned a unique composite classification code, integrating multidimensional classification dimensions to meet the needs of complex queries and multidimensional statistical analysis. The classification process adopts a weighted algorithm to automatically assign categories, and can be accurately matched by combining text features to ensure the scientific and accurate classification results. The weighted algorithm calculation model used is as follows:

$$C(d) = \max_i \sum_{j=1}^{m} w_j \cdot s_j(d, c_i) \tag{3}$$

where $C(d)$ represents the assigned category for document d, c_i indicates category i, $s_j(d, c_i)$ denotes the similarity score between document d and category i based on feature j, and w_j represents the weight assigned to feature j.

4.3 Metadata Framework for Cultural Elements

The metadata framework of the CCCTE corpus used can effectively record the cultural components, language analysis results, and security classification information in the corpus, providing reliable support for corpus management, retrieval, and security control. This framework employs a multi-layered schema that captures fundamental cultural elements while preserving adequate security measures. Table 5 lists the primary components of the metadata framework.

Table 5: Metadata Schema for Cultural Elements

Metadata Category	Attributes	Value Range	Security Level	Example
Cultural Element Identifier	Unique ID, Element type, Cultural domain	Alphanumeric code, Type taxonomy, Domain categories	Low	CE-1024-ARC-TH
Linguistic Representation	Source term, Target term, Translation strategy	Chinese text, English text, Strategy code	Medium	Forbidden City, Cultural adaptation
Cultural Context	Historical period, Cultural significance, Associated practices	Period code, Significance rating, Practice description	Medium-High	QING-01, National significance, Imperial ceremonies
Geographical Association	Location type, Coordinates, Administrative region	Type code, Lat/Long, Region code	Low	SITE-01, 39.9163° N, 116.3972° E, BJ-01
Sensitivity Classification	Cultural sensitivity, Political sensitivity, Commercial sensitivity	Numeric scale (1-5), Yes/No flag, Restricted flag	High	CS-3, PS-Yes, CS-No

The metadata framework consists of XML schema definitions that allow for standardised tagging and still observe TEI standards in corpus linguistics. The implementation comes with automated verification procedures for checking metadata consistency and completeness.

The cultural sensitivity classification within the metadata framework utilizes the mathematical model:

$$S(e) = \alpha \cdot c_s + \beta \cdot p_s + \gamma \cdot m_s \tag{4}$$

where $S(e)$ represents the sensitivity score for cultural element e, c_s denotes cultural sensitivity, p_s indicates political sensitivity, and m_s represents commercial sensitivity. The coefficients α, β, and γ are calibrated based on domain expert evaluations.

4.4 Annotation of Culture-Specific Items and Expressions

The CCCTE Corpus uses a special annotation scheme for culture-specific items (CSIs) and expressions which allows for systematic evaluation of translation techniques and cultural representation. The identification of CSIs is partially automated, but cultural sensitivity and accuracy are ensured through expert validation.

The annotation schema for CSIs includes multiple dimensions:

CSI Type: Material artifacts, Social customs, Ecological elements, Political concepts

Translation Strategy: Preservation, Addition, Naturalization, Cultural equivalent
Functional Equivalence: Full equivalence, Partial equivalence, Non-equivalence
Cultural Loading: High density, Medium density, Low density

The annotation process applies analytical frameworks from translation studies to evaluate the effectiveness of cultural translations. This systematic approach enables quantitative analysis of translation patterns and cultural representation strategies across different text types and domains.

The annotation's equivalence assessment employs the formula:

$$E(c) = \frac{\sum_{i=1}^{n} f_i \cdot w_i}{\sum_{i=1}^{n} w_i} \tag{5}$$

where $E(c)$ represents the equivalence score for CSI c, f_i denotes the equivalence factor for dimension i, and w_i indicates the weight assigned to dimension i.

4.5 Integration of Multimodal Resources

The CCCTE corpus integrates various types of materials such as text, images, and recordings, which not only enriches the expression forms of cultural and tourism text content, but also deepens the analysis framework of cultural components under different presentation modes. At the same time, the CCCTE corpus has also specially constructed supporting control mechanisms to effectively maintain the integrity of reference information, avoid data leakage, loss, or tampering, and ensure the standardization and availability of multimodal data.

The multimodal resources are organized according to:
Modality Type: Textual, Visual, Audio, Combined
Content Function: Informational, Promotional, Interpretive, Ambient
Cultural Representation: Explicit, Implicit, Symbolic, Practical
Cross-modal Relationships: Complementary, Supplementary, Contradictory

The integration of multimodal resources is achieved through a standardized reference system, which not only effectively maintains the correlation between various resources such as text, images, and audio, but also ensures the synergy and correlation of different modal resources. In addition, the refined security control mechanism embedded in the system can strictly control the process of data access, use, and dissemination, prevent security risks such as data leakage and tampering, and ensure the standardization, security, and traceability of corpus resources.

The multimodal significance assessment utilizes the function:

$$M(r) = \sum_{i=1}^{k} \sum_{j=1}^{m} \rho_{ij} \cdot v_i \cdot w_j \tag{6}$$

where $M(r)$ represents the multimodal significance score for resource r, ρ_{ij} denotes the correlation between modality i and function j, v_i indicates the prominence value of modality i, and w_j represents the weight assigned to function j.

4.6 Corpus Size, Composition, and Representativeness

The construction of the CCCTE corpus aims to balance corpus size and organizational coherence, covering various text types from different regions and cultural backgrounds, with a total word count of approximately 10 million. Through scientific corpus screening and allocation, it achieves an organic balance between statistical accuracy and national regional representativeness. The specific allocation of the composition of the corpus is as follows: (1) In terms of regional distribution. The eastern region accounts for 30%, the southern region accounts for 20%, the northern region accounts for 20%, the western region accounts for 15%, and the central region accounts for 15%; (2) In terms of text types. Official website text accounts for 25%, tourism promotional materials account for 20%, travel guides account for 15%, and museum related text accounts for 15%; (3) In the cultural field. Tangible heritage related materials account for 40%, intangible heritage related materials account for 30%, natural heritage related materials account for 15%, and contemporary culture related materials account for 15%; (4) In terms of time coverage. Contemporary related corpus accounts for 60%, modern related corpus accounts for 20%, and historical related corpus accounts for 20%.

The representativeness assessment utilizes the statistical function:

$$R(C) = 1 - \frac{1}{k}\sum_{i=1}^{k} |p_i - q_i| \tag{7}$$

where $R(C)$ represents the representativeness score for corpus C, p_i denotes the proportion of category i in the corpus, q_i indicates the target proportion for category i, and k represents the total number of categories.

The CCCTE corpus deeply integrates security control and cultural analysis by implementing standardized security measures for sensitive text materials. At the same time, the CCCTE corpus integrates language intervention technology and security management standards to construct an implementation system that balances practicality and security, achieving the expansion and optimization of corpus application scenarios.

5 Security Framework Development

5.1 Data Classification and Risk Assessment Model

The security framework for the Chinese Characteristic Cultural Tourism English (CCCTE) Corpus implements a comprehensive data classification system that aligns with both international standards and domestic regulations. The framework establishes multi-tiered protection mechanisms based on systematic risk assessment procedures, as illustrated in Table 6.

Table 6: Data Classification Categories and Protection Requirements

Sensitivity Level	Classification Criteria	Protection Requirements	Example Content
Level 1 (Public)	General tourism information without cultural sensitivity	Standard encryption, Public access	Generic destination descriptions, Opening hours
Level 2 (Internal)	Specialized cultural content with moderate sensitivity	Role-based access control, Enhanced encryption	Detailed historical narratives, Cultural interpretations
Level 3 (Confidential)	Content with significant cultural, political or commercial sensitivity	Multi-factor authentication, Advanced encryption, Access logging	Sacred site descriptions, Politically sensitive historical content
Level 4 (Restricted)	Highly sensitive cultural material requiring special protection	Compartmentalized access, Maximum-strength encryption, Comprehensive auditing	Restricted ceremonial knowledge, Commercial trade secrets

5.2 Authentication and Access Control Architecture

The CCCTE corpus adopts a multi factor user authentication and access control architecture, which achieves standardized management of user behavior through different levels of authentication and permission authorization. While providing security protection capabilities that meet research needs, it ensures the normal access of legitimate users to relevant information for academic research purposes, achieving a balance between security and practicality. The core components of the security defense system include:

1. Identity management system: federated single sign-on identity with streamlined federated authentication for academic partners.
2. Authorisation framework: user roles mapped to permissions with level-enabled dynamic constraints.
3. Access control enforcement: multi-layer policy-based control governance.
4. Audit and monitoring system: anomalous event detection under comprehensive timestamps.

This single-draft system ensures that the needed corpus resources are available to researchers under appropriate restrictions for sensitive cultural material. The access model includes contextual boundaries where parameters such as the user’s geo-location, reason for research activity, and time constraints are considered as dynamic access factors.

5.3 Data Encryption and Protection Strategies

The CCCTE Corpus employs a comprehensive encryption framework that protects data throughout its lifecycle. The encryption strategy implements multiple protection layers addressing data-at-rest, data-in-transit, and data-in-use scenarios, as shown in Figure 3.

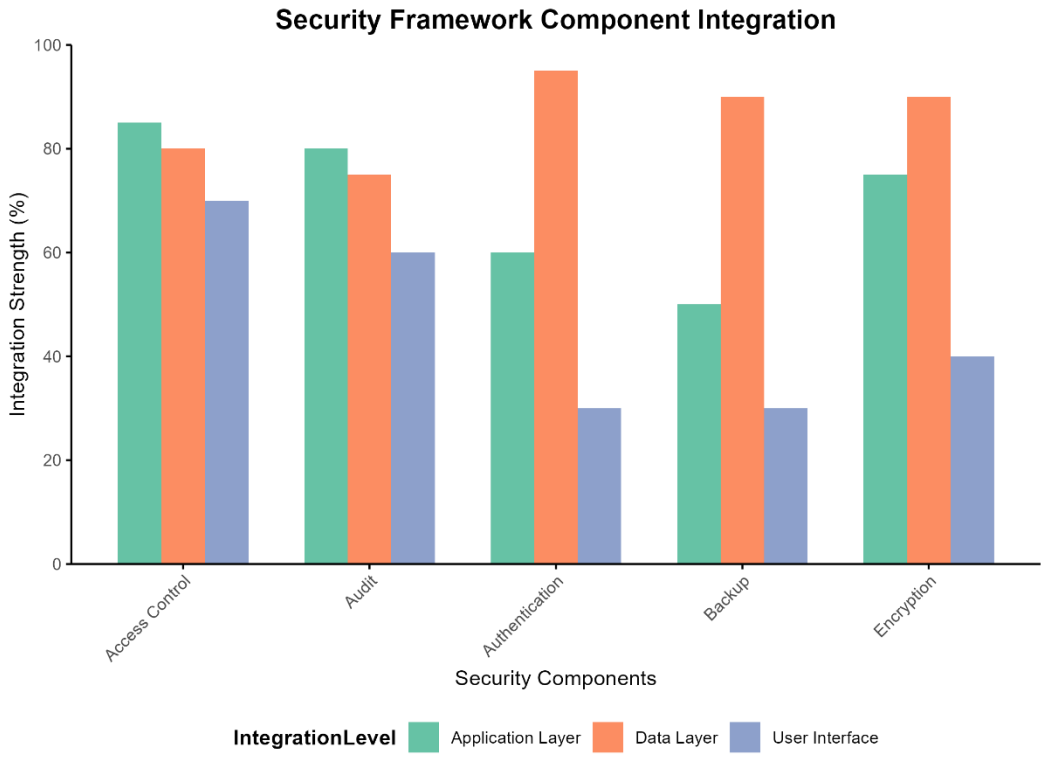

Figure 3: Multi-layered Encryption Framework for CCCTE Corpus

As shown in Figure 3, the CCCTE corpus encryption framework adopts differentiated security protection strategies at the architecture layer to achieve full process security control. Specifically, (1) in terms of storage encryption. Adopting AES-256 encryption algorithm for corpus data storage to achieve secure protection in the data storage process; (2) In terms of transmission encryption. Using TLS 1.3 protocol for data transmission, with complete network transmission forward confidentiality; (3) In terms of application-level encryption. For sensitive cultural elements, field level encryption is used to accurately protect core cultural information and sensitive data; (4) In terms of tagging processing. Tagging sensitive values and replacing sensitive information with non-sensitive equivalents ensures data security without affecting research use. This encryption framework not only achieves comprehensive security protection for corpus data, but also effectively balances system performance and data availability, providing solid support for the secure management of multimodal corpus.

5.4 Compliance with International and Domestic Data Regulations

The security framework of CCCTE corpus strictly complies with the regulatory requirements of multiple jurisdictions, with a focus on following the relevant provisions of China's Data Security Law and Personal Information Protection Law. Through systematic security control design, it ensures that the construction, operation, and use of the corpus comply with various regulatory requirements, effectively preventing data security risks and protecting personal information and cultural related data security. Table 7 shows the main regulatory considerations and the specific implementation mechanisms for each factor.

Table 7: Regulatory Compliance Implementation for CCCTE Corpus

Regulatory Framework	Key Requirements	Implementation Mechanisms	Validation Procedures
Data Security Law (PRC)	Data classification and protection	Multi-level security model	Regular compliance audits
Personal Information Protection Law (PRC)	Consent management, purpose limitation	Purpose-specific access controls	Privacy impact assessments
EU General Data Protection Regulation	Cross-border data transfer controls	Data localization options	Transfer impact assessments
ISO/IEC 27001	Information security management	Comprehensive security controls	External certification
Copyright Law (PRC)	Protection of original works	Attribution mechanisms, licensing controls	Rights management verification

The compliance framework mainly includes two core functional modules: (1) automatic supervision mechanism. Real time monitoring of the entire process of corpus processing and data usage, accurately identifying possible regulatory violations; (2) System correction mechanism. Develop clear corrective measures for detected violations or compliance gaps, quickly address deficiencies, and standardize operations. This comprehensive strategy of "monitoring+correction" ensures that the operation of the CCCTE corpus fully complies with relevant laws, regulations, and regulatory requirements, achieves legal and compliant operation, and guarantees the educational and research value of the corpus, allowing the corpus resources to be fully utilized within the scope of compliance.

5.5 Backup and Recovery Strategy

The CCCTE corpus has built a complete backup and disaster recovery framework, providing solid support for the secure storage and stable use of corpus resources. This framework strictly follows the 3-2-1 backup model, which includes keeping three copies of data, using two different types of storage media, and keeping one remote backup, combined with a cultural content exclusive protection protocol to further enhance the security and reliability of data backup. It specifically includes four core modules: (1) incremental snapshot system module. Can achieve continuous change capture, support time point recovery function, and quickly trace back corpus data from different time nodes; (2) Cold storage archive module. Build a dedicated immutable backup repository for long-term archiving and storage of corpus data; (3) Object metadata storage module. Independently backup the structural metadata of the corpus to ensure the integrity of the corpus system; (4) Geographically diverse streaming media replication module. Synchronize the corpus data stream to multiple storage sites with different geographical distributions to achieve multi regional backup. In addition, the backup and recovery framework is equipped with a predetermined backup testing process to ensure that data recovery can be quickly completed in scenarios such as sudden disasters and data anomalies, ensuring the efficient and stable operation of the corpus.

6 Conclusion and Future Directions

6.1 Summary of Key Findings and Contributions

This study has created an all-encompassing process for the creation and security procedures of a Chinese Characteristic Cultural Tourism English Corpus in the era of digital technology. The combination of disciplines such as corpus linguistics with the demands of cultural tourism translation and modern information security needs resolves a substantial gap in the current body of literature, which to some degree lays the groundwork for a systematic examination of discourse in tourism from the protected culture-sensitive content perspective.

The major accomplishments of this research include the design of a multi-security cultural tourism corpus and the construction of a conceptual model that enables the integration of control systems with the diverse textual forms of cultural tourism. The design of the classification scheme for culture-specific items facilitates a systematic examination of translation and cultural representation of the diverse culture, which was a methodological gap in preceding studies. The security framework devised addresses the corpus's data protection functions and design requirements, thus creating a flexible paradigm for similar sensitive corpus construction initiatives.

Of great importance, the study illustrates that the methodologies designed for corpus development can incorporate security features while maintaining core linguistic and research functions as value-based principles. The systematic approach taken ensures that the Chinese texts and the related materials on cultural tourism are not only comprehensive but also harmoniously dimensional and coherent for logical analysis of the discerned patterns of translation with the representation of the culture.

6.2 Practical Implications for Tourism Translation

The CCCTE Corpus is useful in the context of tourism translation practice by offering a substantial practical value claimed tourism translation strategy development and evaluation. It provides documented evidence of translation and cultural representation systems for tourism industry translators and content creators to make decisions based on facts.

Effective culture-specific item strategies can be used in training programmes as well as in quality control systems for the tourism translation service. The construction of specialised lexicographic materials for tourism translators is enhanced through the extraction of title- and domain-specific phrases, as the architecture of the corpus allows for substantiating domain-related terminology and phraseology.

Cultural constituent annotation aids transdisciplinary tourism research by permitting systematic evaluation of contextual determinants of translation efficacy, thus enabling practical development of tourism translation policy with a focus on specific context. Moreover, the security arrangement sets out sensitive cultural material translation protocols that address some ethical issues inadequately dealt with in earlier translation research.

This illustrates attempts from the integrated approach to show how errors can be made over globalisation to document on the concepts that deal with cultural inform within the document as these issues strengthen the need for Canadian cross international borders.

6.3 Theoretical Significance for Corpus Linguistics

In a theoretical sense, this research innovates corpus linguistics methodologies to include elements of culture and security which previous works have overlooked. The blend of cultural translation theories with corpus construction principles enhances the discourse on cross-cultural specialised discourse analysis by addressing gaps in previous corpus-based

translation studies.

The annotation framework for culture-specific items offers a systematic method of documenting translation phenomena that were previously dealt with only through case studies or small-scale sampling. This approach enables more rigorous qualitative analysis of cultural translation phenomena, thus strengthening the empirical base of translation studies. In setting the security classification system, theoretical determinants are established as to how research access and protective restrictions on sensitive materials can be balanced, providing an ethical framework that has largely been absent in corpus linguistics.

Furthermore, corpus representativeness sample design, classification of cultural sensitivity, and estimation of security risk level matrices all provide an enhanced quantitative foundation in corpus linguistics which other specialised corpus development projects could be designed around. Such integration of quantitative with qualitative approaches provides a stronger complex structure for meta-analysis of translation studies based on corpus.

6.4 Limitations and Future Research Directions

This research has emerging gaps that require new perspectives, even with the value it adds. For now, the corpus is concentrated on the tourist textual interplay, with little to no integration of multimodal aspects which play prominent roles in tourism communication. Building and improving the stitching methods of visual and audio elements in the future, and developing complex research paradigms that can systematically address multimodal translation phenomena in language databases, are key supports for enhancing research depth and breadth.

The current research work mainly focuses on the one-way translation scenario from English text to Chinese, which has certain limitations in this one-way research mode. In the future, it is necessary to further expand the scope of application of the corpus framework, in order to provide support for comparative cultural mediation analysis in different translation contexts and enrich the results of cross contextual translation research.

At the level of security assurance, the existing security framework needs to be dynamically adjusted and optimized in conjunction with the updates and iterations of industry regulations and the emergence of new digital threats. Future research should focus on building more adaptive dynamic security frameworks that achieve a balance between security and practicality. At the same time, the adaptive security framework should integrate advanced technology architectures such as federated learning and differential privacy, and ensure rich analysis and reasonable utilization of corpus data on the basis of continuously improving information security protection capabilities.

6.5 Potential for Cross-Cultural Communication Enhancement

The CCCTE corpus, as a core resource for enhancing cross-cultural communication in the tourism environment, helps to improve mutual understanding between Chinese tourism stakeholders and international tourists, and solve cognitive differences in cross-cultural communication. Clear presentation of conceptual differences in different cultural backgrounds for specific cultural projects, accurate identification of conceptual gaps that are prone to communication barriers in tourism scenarios, and effective resolution of communication pain points in cross-cultural communication. At the same time, by identifying mature and effective translation models, practical guidance can be provided for the compilation of tourism promotion materials, optimization of heritage interpretation texts, and the development of cultural mediation activities, promoting the high-quality development of the cultural tourism industry.

In addition, a responsible cross-cultural communication model has been established, which fully respects the sensitivity of different cultures and achieves reasonable sharing of

information, taking into account the dual needs of cultural dissemination and security control. It not only ensures the effective transmission of cultural connotations, but also avoids risks such as cultural leakage and information tampering, providing strong support for sustainable and mutually respectful tourism development practices.

The methodological framework developed in this study not only promotes theoretical innovation and practical implementation in the fields of corpus linguistics, translation research, and tourism communication, but also constructs a replicable and scalable model, providing important references for academic research and practical applications in related fields.

Funding

Research results in Daqing city philosophy and social sciences planning project (DSGB2024111).

About the Author

Zhuo Li, female, was born in 1987 and is from Daqing, Heilongjiang. P.R. China. She holds a Master's degree in English Language and Literature and is an Associate Professor at Daqing Open University. Her research interests include English teaching, Chinese-English translation, and corpus linguistics.

Reference

[1] Al-Bahrani, R. H. (2018). Research on Tourism English Translation Based on Cultural Difference. Academia.edu, 37-48.

[2] Francesconi, S. (2021). TOURISM TRANSLATION: From dictionary to corpus. ResearchGate, 350-396.

[3] Durán Muñoz, I. (2013). Translating the Language of Tourism. A Corpus Based Study on the Translational Tourism English Corpus (T-TourEC). ScienceDirect, 325-335.

[4] Liao, M., & O'Gorman, K. (2013). Reading between the lines: Multidimensional translation in tourism consumption. ScienceDirect, 157-167.

[5] Li, S., & Kit, C. (2021). Legislative discourse of digital governance: a corpus-driven comparative study of laws in the European Union and China. DeGruyter, 267-290.

[6] Li, W., Wu, J., & Holmes, K. (2021). Developing Culturally Effective Strategies for Chinese to English Geotourism Translation by Corpus-Based Interdisciplinary Translation Analysis. Geoheritage, 13(4), 616-629.

[7] Morbiato, A. (2020). Corpus-Based Research on Chinese Language and Linguistics. Academia.edu, 45-97.

[8] Wang, Y., & Dong, Y. (2024). An analytical framework for corpus-based translation studies. Humanities and Social Sciences Communications, 11(2), 250-265.

[9] Morbiato, A. (2023). Corpus-Based Research on Chinese Language and Linguistics. Academia.edu, 76-236.

[10] D'Egidio, A. (2014). The Language of Tourists in English and Italian Travel Blogs and Trip Reports: a Corpus-based Analysis. Lingue Culture Mediazioni, 1(1), 102-116.

[11] Li, W., Zhu, Y., & Holmes, K. (2024). Effective Chinese-to-English biotic interpretation in ecotourism destinations: a corpus-based interdisciplinary study. Journal of Sustainable Tourism, 32(4), 804-825.

[12] Alangari, E. (2022). Cultural Mediation in Tourism Translation: Saudi Arabia as a Case Study. Arab World English Journal for Translation & Literary Studies, 6(3), 82-98.

[13] Amenador, K. B., & Wang, Z. (2022). The Translation of Culture-Specific Items (CSIs) in Chinese-English Food Menu Corpus: A Study of Strategies and Factors. SAGE Journals, 12(2), 1-19.

[14] Hu, T. (2003). Eco-Translatology: Interdisciplinary approaches to translation from cultural and ecological perspectives. Foreign Language Research, 5(1), 62-67.

[15] Sala, M., & Maci, S. M. (2023). Corpus Linguistics and Translation Tools for Digital Humanities: Research Methods and Applications. Bloomsbury, 154-168.

[16] Wang, F., & Xu, J. (2019). THE ROUTLEDGE HANDBOOK OF CHINESE TRANSLATION. Academia.edu, 283-302.

[17] Lu, S., & Chen, P. (2024). Linguistic landscape as a way to construct multiple identities in the context of globalization: the case of an ancient town in China. Social Semiotics, 34(2), 405-429.

[18] Creemers, R. (2022). China's emerging data protection framework. Journal of Cybersecurity, 8(1), 1-12.

[19] Sala, M., & Maci, S. M. (2023). Corpus Linguistics and Translation Tools for Digital Humanities: Privacy considerations in corpus development. Bloomsbury, 167-175.

[20] Li, S., & Kit, C. (2021). Comparative analysis of data governance legislation between regions: European Union and China. DeGruyter, 278-285.

[21] Webster, G., & Creemers, R. (2021). Translation: Data Security Law of the People's Republic of China. DigiChina, Stanford University, 1-15.