5-ŞÖBA · OLIY TA'LIM VA KASB-HUNAR TA'LIMI SOHASIDA XALQARO HAMKORLIK

RESEARCH ON THE CONSTRUCTION OF A CHINESE–ENGLISH–UZBEK AGRICULTURAL TERMINOLOGY DATABASE

Yang Wenge

Master's degree, Professor, College of Chinese Language, Literature and Art, Xinjiang Agricultural University, Urumqi

Zhang Chunmei

Master's degree, Associate Professor, College of Chinese Language, Literature and Art, Xinjiang Agricultural University, Urumqi

Submission section: Section 5, International Cooperation in Higher Education and Vocational Education

DOI 10.66073/ReFocus-Conf/1311155

NASHR SANASI 13.11.2026 · Nukus

SAHIFALAR 718–727

NASHR HAMKORI Research Focus
ISSN 2181-3833

IQTIBOS UÇUN Yang W., Zhang Ch. Research on the construction of a Chinese–english–uzbek agricultural terminology database // Veterinariya meditsinasi va çorvaçilikdagi dolzarb muammolar va ularning yeçimlari: xalqaro ilmiy va ilmiy-texnik anjuman materiallari. – Nukus, 2026. – B. 718–727. – DOI: 10.66073/ReFocus-Conf/1311155

ABSTRACT The widespread terminological asymmetry in China–Uzbekistan agricultural cooperation raises the costs of qualification verification, technical training, and contract negotiation. This paper explains the construction pathway for a Chinese–English–Uzbek agricultural terminology database: using a Chinese–English agricultural dictionary as the base text, 15,158 trilingual entries were empirically annotated, and 275 newly added terms were supplemented, forming a dual-source corpus of "stock resources plus incremental resources." The terminology database uses entry numbers as the primary key and contains eight fields: Chinese headword, Pinyin annotation, English equivalent, Uzbek translation, subject classification, original page number, status marker, and cross-reference entry. It supports three-way retrieval in Chinese, English, and Uzbek. The database construction process is divided into six stages: layout parsing, column-by-column transcription, main-table alignment, discrepancy checking, translation-name generation, and output release. Artificial intelligence significantly reduces manual costs in transcription, batch comparison, and preliminary translation drafting. Quality control relies on a four-stage process of "machine drafting—rule constraints—expert review—version consolidation," with emphasis on preventing four types of risks: misrecognition of visually similar characters, coexistence of multiple translations, semantic drift, and instability in the translation of policy terminology. This paper summarizes four roles and six practices of artificial intelligence in terminology construction and compares the applicable boundaries of general large language models, execution-oriented agents, and terminology-specific systems.

KEYWORDS agricultural terminology database; Chinese–English–Uzbek; artificial intelligence; China–Uzbekistan cooperation

I. INTRODUCTION

In cross-border cooperation in agriculture between China and Uzbekistan, terminological asymmetry is a frequently occurring yet long-neglected communication barrier. Agreements signed by both sides, official documents issued, and technical training conducted in areas such as cotton cultivation, grain production, livestock breeding,

veterinary health, agricultural machinery, and plant quarantine all require cross-language conversion among Chinese, English, and Uzbek. At the practical level, three typical problems are widespread. First, some concepts have corresponding terms in both Chinese and English, but Uzbek lacks a standardized equivalent expression. Second, the same professional concept has multiple translations in different texts, and there is no unified standard. Third, some translated names conform to the linguistic conventions of the target language in literal terms but deviate from the core meaning of the original term. These three types of problems are intertwined and superimposed. They not only significantly increase transaction costs for both sides but also negatively affect cross-language mutual recognition of phytosanitary certificates, agricultural technical standards, and professional textbooks [1]. Therefore, integrating agricultural terms scattered in paper dictionaries, policy documents, and industry materials into a trilingual professional terminology database that is searchable, maintainable, and shareable is not an auxiliary add-on construction task but a core infrastructure for deepening China–Uzbekistan agricultural cooperation. This paper systematically reviews the entire construction process of this trilingual agricultural terminology database. The research covers corpus sources, database structure design, construction process organization, quality control system, and application methods of the terminology database in teaching scenarios and practical work.

II. CORPUS BASIS: STOCK RESOURCES AND INCREMENTAL RESOURCES

(1) Sources of Stock Corpus

The stock corpus of this terminology database comes from a Chinese–English agricultural dictionary arranged in Chinese phonetic order. It covers disciplines such as crop breeding, tillage and cultivation, soil and fertilizer, plant protection, animal husbandry and veterinary medicine, agricultural machinery, agricultural product processing, and agricultural economics, with approximately 15,000 entries. The dictionary uses a two-column layout, and the scanned version has no native searchable text layer. The entry format is "Chinese headword + English equivalent," and Uzbek translations are completely absent: the Chinese and English content constitutes the existing corpus of the dictionary, while Uzbek translations must be separately supplemented and constructed. This corpus structure determines that the core workload of database construction is concentrated in the generation of Uzbek translations. The dictionary is divided into letter sections in Pinyin order; 23 of these sections contain entries and were therefore covered by the present standardized sorting, while a few letters have no entries in the base dictionary. A total of 15,158 entries were verified and included in the trilingual database. Among them, section S contains the largest number of entries, reaching 1,717, while section O contains the smallest number, only 19.

(2) Collection of Incremental Corpus

Because the base dictionary was published relatively early, most policy concepts and digital technology terms newly added after 2012 are not included. Terms such as "藏粮于地、藏粮于技" (English: storing grain in land and technology; Uzbek: doni yer hamda texnologiyaga tayanib ta'minlash), "智慧农业" (English: smart agriculture; Uzbek: aqlli qishloq xoʻjaligi), "植物工厂" (English: plant factory; Uzbek: oʻsimlik fabriki), as well as veterinary terms such as "非洲猪瘟" (English: African swine fever; Uzbek: Afrika choʻchqa

vabosi), “无规定动物疫病区” (English: specified animal disease-free zone; Uzbek: belgilangan hayvoniy kasalliklardan xoli hudud), and “兽药残留限量” (English: maximum residue limits of veterinary drugs; Uzbek: veterinar dori qoldiqlarining ruxsat etilgan chegarasi), do not appear in the paper dictionary but occur frequently in China–Uzbekistan agricultural cooperation documents and project application materials. On this basis, this study adopts “newly proposed after 2012 or given new connotations” as the selection criterion, uses China’s successive “No. 1 Central Documents” and industry plans and annual reports issued by the Ministry of Agriculture and Rural Affairs as corpus sources, and completes letter-section classification according to the same Pinyin alphabetical order as the stock base text. Ultimately, 275 incremental entries were included and distributed across 23 letter sections; the sections that have no entries in the base dictionary (A, C and V, among others) are populated by incremental entries alone. With stock and incremental resources combined, the database covers 23 letter sections and comprises approximately 15,430 entries.

(3) Processing Standards for Veterinary and Animal Husbandry Corpus

Such entries mostly involve animal disease names, veterinary drug names, feed additive names, and international generic names. For translation-name processing, it is appropriate to adopt the standardized method of “generic-name transliteration plus Latin scientific name in parentheses,” such as 灭蝇胺 (English: cyromazine; Uzbek: siromazin) and 灭线磷 (English: ethoprophos; Uzbek: etoprofos). This method is consistent with the term names officially registered by the Uzbek side and can also support traceability verification of translated names. For animal disease terms, it is appropriate to use the English names published by the World Organisation for Animal Health as an intermediary for translation, thereby avoiding semantic deviation caused by word-for-word literal translation.

III. DATABASE STRUCTURE DESIGN

(1) Field and Primary Key Design

The terminology database uses entry numbers as the database primary key. The coding rule is “letter-section number plus serial number,” thereby ensuring that any terminology entry can be traced to the specific page and column of the source dictionary. Each terminology record has eight structured fields, namely Chinese headword, Hanyu Pinyin annotation, English equivalent term, Uzbek translation, subject classification, source document page number, review status marker, and cross-reference related entry. The review status marker is divided into three levels: “approved,” “pending verification,” and “questionable.” This mechanism allows entries that have not yet completed final review to be temporarily included in the terminology database, preventing them from being mixed into the final approved terminology set.

(2) Subject Classification Rules

Based on the original outline system of the source dictionary and current agricultural discipline classification standards, nine first-level discipline categories are divided: crop science, horticulture, soil and fertilizer science, plant protection, animal husbandry, veterinary medicine, agricultural engineering (agricultural machinery direction), agricultural product processing, and agricultural economics. In addition to serving a navigational retrieval function, the subject classification field also constitutes the applicable boundary of

https://doi.org/10.66073/ReFocus-Conf/1311155

the rule "one term corresponds to one translation." The same term may correspond to differentiated translations in different disciplinary contexts, and subject classification can provide contextual constraints for semantic determination of terms.

(3) Cross-Reference and Deduplication Mechanism

For different word-order expressions of the same concept, cross-reference guides are established according to the fact that the first character's Pinyin belongs to different letter sections. For example, "文明乡风" is placed in section W, while "乡风文明" is placed in section X. The two expressions are separately established as entries and cross-reference markers are added, which both prevents duplicate inclusion of redundant entries and ensures full coverage of multi-path retrieval. For terminology entries with the same form but different meanings, separate entries are established independently, and semantic distinction is achieved through the subject classification field.

(4) Retrieval Function Design

The terminology database has three-way retrieval capabilities in Chinese, English, and Uzbek. It is equipped with fuzzy matching and rapid positioning by Pinyin initial letters, and also supports combined filtering by section and discipline. In practical application scenarios, the most frequently used function is "reverse verification," that is, using an Uzbek translation to search back for the corresponding Chinese headword, thereby determining whether a translated name has already been included and whether it conflicts with existing entries. This function is a digital advantage that traditional paper dictionaries cannot achieve.

IV. DATABASE CONSTRUCTION PROCESS AND TECHNICAL PATHWAY

(1) Layout Parsing and Column Segmentation Process

For scanned pages with a two-column layout, if whole-page recognition is carried out, entries in the left and right columns will cross and become misaligned in reading order, thereby generating incomplete entries. This study adopts the following processing pathway: after rendering each scanned page as an image, the page is segmented along the center gutter of the type area, so that the left and right columns each form an independent image unit; then text concatenation is completed according to the original book layout logic of "left column first, right column second." The recognition files for each column follow the naming rule of "letter + page number + column position," ensuring that any terminology entry can be traced to the physical location in the original book.

(2) Column-by-Column Transcription and Immediate Saving Mechanism

The recognition model transcribes the content of a single column and generates a standardized headword list. During project implementation, serious losses were encountered due to imperfect data storage logic: preliminary transcription results were retained only in the dialogue context, and all content was lost after the process restarted. Based on this lesson, the project team established the operating principle of "immediate saving after each column," that is, after completing the transcription of each column, the text is immediately written into a plain text file and marked with the corresponding column identifier. Such files are both standardized records of the project process and core evidence for breakpoint resumption.

(3) Main Table Construction and Entry Alignment Method

The transcribed corpus is integrated into a "main table" corresponding one-to-one with the entries in the original book, and all subsequent procedures use this main table as the baseline coordinate. In the initial stage of the project, a sequential pointer advancement alignment method was adopted, which was prone to entry misalignment when encountering headwords with the same name. After optimization, a "headword pool" strategy was adopted, in which candidate entries for the same headword are placed into corresponding pool units and processed by sequential popping. This both fully preserves duplicate entries and effectively avoids alignment misalignment problems. Discrepancies found in the alignment stage can be divided into two types: expressive differences and substantive errors. The former, such as differences between abbreviated forms and full forms, can be summarized into matching rules for batch processing; the latter must be returned to the original book page and column for verification and correction.

(4) Translation Generation and Review Standards

The construction of Uzbek translations follows three constraint rules. At the orthographic level, the Latin alphabet system is uniformly adopted, strictly following the writing norms of o', g', sh, and ch. At the word-formation level, productive affixes are preferentially selected, such as -chilik to indicate "industry," -shunoslik to indicate "discipline," and parvarish to indicate the meaning of "cultivation or raising." The principle of one term, one translation is strictly observed. For terms lacking cultural context, the processing method of "loanword plus qualifier" is adopted. After the model drafts translations, they must be compared one by one with the corresponding English senses, and candidate translations with semantic drift are eliminated.

(5) Tool Selection and Task Division

During database construction, five types of tools were successively introduced: WorkBuddy, DeepSeek, Doubao, ChatGPT, and Tongyi Qianwen, and task division was configured according to task attributes. WorkBuddy undertakes long-process operations, and the table-generation procedures for each section are completed relying on this tool. DeepSeek is used for sense discrimination and inference optimization of Uzbek translations. Doubao and Tongyi Qianwen are mainly used for definition rewriting and grammatical compliance verification. ChatGPT is used for accuracy review of English equivalents. Among them, WorkBuddy and DeepSeek have the most prominent application effectiveness—the former's core advantage is that it can implement long-process tasks, and the latter's core advantage is clear semantic discrimination logic. A detailed comparison of relevant platforms is provided in Part V.

(6) Structured Processing Pathway for Incremental Entries

Newly supplemented entries lack corresponding base-text support, so the structural constraint stage needs to be strengthened. First, a large language model extracts candidate terms from policy texts and generates draft definitions. Then manual work completes three verification tasks: whether they conform to the policy orientation of the new era, whether they belong to the corresponding section, and whether the English expression adopts the official translation. After verification is completed, such entries adopt a unified field structure with base-text entries and can be incorporated into the same retrieval system, achieving integrated retrieval and verification of stock corpus and incremental corpus.

https://doi.org/10.66073/ReFocus-Conf/1311155

V. PRACTICAL PATHWAYS FOR AI-EMPOWERED TERMINOLOGY DATABASE CONSTRUCTION

(1) Overall Research Framework

The core process of terminology system construction can be summarized into four core links: "term selection, standardized naming, multi-level verification, and dynamic maintenance." Under the traditional terminology construction model, the above links all depend on manual completion and, constrained by labor costs, can only achieve sampling-based processing. After the introduction of artificial intelligence technology, each link can generate initial results in batches, and the human role accordingly shifts to both ends of the process, namely pre-establishing normative rules and post-completing result adjudication. Based on this, artificial intelligence assumes four functional roles in the terminology database construction process. First, it is a recognition subject, realizing the transformation of scanned and image-type corpus into searchable structured text. Second, it is an extraction subject, extracting candidate terms from policy documents, industry standards, and professional technical materials. Third, it is a creation subject, generating candidate sequences of Uzbek translations by combining English senses and word-formation norms, while simultaneously explaining word-formation rationale. Fourth, it is a verification subject, locating terminology inconsistency problems through reverse back-translation and cross-section comparison. The above four functions correspond to different task attributes, and targeted selection of adapted models is required; the entire process cannot be handled through a single round of conversation.

(2) Six Implementable Pathways

The first is batch corpus preparation. A complete policy compilation is imported into a model with context-processing capability, and candidate terms are extracted using the dual selection criteria of "whether it refers to an independent concept" and "whether a cross-language correspondence needs to be established." Hundreds of candidate corpus items can be obtained in a single pass, and manual work only needs to complete final screening. The second is terminology table alignment. The newly transcribed section terminology table and the existing translation database are simultaneously imported into the model, which completes item-by-item comparison and generates a discrepancy list. Discrepancy types are divided into four categories: missing terms, spelling deviations, variant expressions, and inconsistency between simplified and traditional Chinese. The third is sense disambiguation. For polysemous terms such as "稻" and "大米" (English: paddy and rice; Uzbek: sholi and guruch), and "苗" and "秧" (English: seedling and rice seedling; Uzbek: nihol and sholi koʻchati), the model is required to first sort out the English senses and then match the corresponding Uzbek expressions one by one, avoiding mixed translation of polysemous terms. The fourth is translation generation. Combining orthographic rules and word-formation norms, the model is required to output three candidate translations at a time and annotate the word-formation rationale, supporting subsequent selection of the best. The fifth is consistency inspection. A complete section terminology table is input into the model, which identifies variant translations of the same concept and entries with the same form but different meanings. The sixth is database operation and maintenance. New terminology entries automatically reuse the unified field template, generate cross-reference markers and subject classification identifiers, and reduce manual copying costs. Based on accounting of

the project implementation cycle, a single section of terminology requires only 1–3 hours from page rendering to table delivery. If the same scale of work were completed manually, it is estimated that several weeks would be required.

(3) Classification and Introduction of Commonly Used Technical Platforms

The artificial intelligence platforms applied in the construction of the Chinese–English–Uzbek agricultural terminology database can be divided into three categories according to functional attributes. The first category is general large language dialogue models, including ChatGPT, Claude, Doubao, DeepSeek, Tongyi Qianwen, Kimi, and others. Each platform has differentiated advantages: ChatGPT is suitable for English sense verification and academic normative expression generation; Claude has stable long-document processing performance; Doubao is equipped with a multimodal recognition module and can directly read scanned page content; DeepSeek has higher semantic understanding accuracy for Chinese expressions and policy vocabulary, with strong long-text processing stability; Tongyi Qianwen is suitable for definition rewriting and structured table organization; Kimi has the ability to read super-long documents in a single pass. Such tools have a low operating threshold and adopt a conversation-based billing model, making them suitable for semantic discrimination and terminology generation. The second category is execution-oriented agents, represented by WorkBuddy, which have local file reading and writing, batch transcription, and continuous multi-step process execution capabilities, and can directly export processing results as Word documents or tables. They are suitable for "whole-section process automation" scenarios. Their core advantage is not single-round response quality but end-to-end process implementation capability. The third category is terminology management and translation-specific systems: the terminology modules of SDL MultiTerm, memoQ, and Trados have entry management, review workflow configuration, and TBX format interaction functions; the domestic "Termonline" platform is led by the China National Committee for Terms in Sciences and Technologies and can query approved standardized translations. Such systems are suitable for long-term hosting of finished terminology databases and cross-team collaboration scenarios.

(4) Technical Selection Principles and Application Boundaries

At the technical selection level, the selection of multi-model tools for the terminology recognition link follows the principle of functional adaptation: long-context language models are selected for term extraction and translation generation to ensure context coverage; execution-oriented agents are selected for batch process processing to improve batch processing efficiency; specialized terminology management systems are selected for finished terminology database hosting to ensure system stability and maintainability. The functional boundaries of the four types of tools are clear, and there is no functional substitutability. General language models focus on semantic generation and optimization of terminology matching, while specialized terminology management systems focus on system stability and long-term maintainability construction. Based on practical application data from this project, WorkBuddy and DeepSeek have the most significant application effectiveness. Among them, WorkBuddy's core advantage is strong process implementation and execution capability, while DeepSeek's core advantage is high semantic discrimination accuracy. In addition, data security principles must be strictly observed: unpublished policy texts, crop germplasm resources, and sensitive quarantine data are strictly prohibited from

https://doi.org/10.66073/ReFocus-Conf/1311155

being uploaded to public cloud large models. Local deployment modes or desensitization processing must be adopted before application. After terminology translation verification is completed, it must be synchronously written back to the terminology database and the word-formation rationale retained, eliminating the problem of inconsistent translation expressions for the same concept in different batch processing procedures.

VI. QUALITY CONTROL AND LONG-TERM MAINTENANCE MECHANISM

(1) Representation of Four Typical Risks

The first is recognition deviation of visually similar characters, such as “田埂” and “田垄” (corresponding English: field bund and field ridge; Uzbek: dala toʻsigʻi and or), and “拳师犬” and “拳狮犬” (corresponding English: Boxer; Uzbek: boksyor). Subtle differences at the character level will lead to completely different terminology objects being referred to. The second is the problem of multiple coexisting translations. The same terminology concept is constructed into different Uzbek expressions in different sections, and long-term persistence will undermine the systematic consistency of the terminology system. The third is semantic drift, that is, the literal expression of the translation conforms to target-language expression norms but deviates from the sense of the source term. Such problems are the most difficult to identify and have the most significant negative impact. The fourth is insufficient stability of translations for policy-type terms. Such concepts usually first form Chinese expressions and later promulgate official external translations. Generation models based on literal construction are prone to producing inconsistent translations before and after. Authoritative foreign-language release versions should be used as the basis for naming. When there is no official translation, the original Chinese form should preferentially be retained with explanatory notes [2].

(2) Four-Stage Quality Control Process

The four stages of “machine drafting—rule constraints—expert review—version consolidation” have a progressive logical relationship rather than a mutually substitutive one. The machine drafting stage solves the problem of terminology coverage; the rule constraint stage solves the problem of terminology stylistic standardization; the expert review stage solves the problem of terminology sense accuracy; and the version consolidation stage realizes standardized retention of revision results, avoiding repeated occurrence of similar errors at the process level. Any operating model that directly takes generative model output as the final draft will lead to quality loss of control at the expert review stage.

(3) Version Management and Dynamic Update Mechanism

The terminology database releases incremental versions on a quarterly cycle. Each release is accompanied by a revision log recording details of newly added entries, revised entries, and withdrawn entries. Controversial terminology entries are separately marked and archived, and are regularly submitted to a review group composed of Chinese and Uzbek experts for deliberation. After deliberation is completed, the final naming is written back to the database, while historical translations are retained to meet traceability needs. The database adopts a general data exchange format and can achieve compatible docking with existing information systems of universities and enterprises [3].

VII. APPLICATION SCENARIOS AND DEVELOPMENT PROSPECTS

In the teaching application dimension, the Chinese–English–Uzbek trilingual terminology database can be directly used as practical training material for terminology and translation courses: given a Chinese headword and English equivalent, trainees are asked to draft an Uzbek translation and then compare it with the approved entries in the database. The differences between the two can serve as core entry points for classroom discussion. In the industry practice dimension, the terminology database can support three typical application scenarios: first, terminology translation verification of foreign trade documents and cooperation contracts; second, bilingual comparative editing of agricultural technical training handouts; third, injection of domain terminology word lists into machine translation systems to reduce terminology mistranslation rates when general large models process professional texts [4].

Subsequent development plans cover three directions. First, expand the coverage of the veterinary and animal husbandry sub-database, include terminology entries for animal disease classification, veterinary drug varieties, and feed additives, and align with the English naming systems of international organizations. Second, build a joint review mechanism with Uzbek universities and agricultural research institutions, embed the terminology translation review link into bilateral cooperation project processes, and realize the transformation of translations from "unilateral construction" to "bilateral consensus." Third, rely on the terminology database to carry out teaching and training applications, build a closed-loop iterative system of "use—feedback—revision," and promote continuous improvement of the database in dynamic application [5].

VIII. CONCLUSION

The practical experience of this project can be summarized into a core conclusion: the application value of a terminology database does not depend on the scale of included entries but on the core attributes of being "traceable, searchable, and iteratively revisable." Artificial intelligence technology has not changed the essential nature of terminology work; it has only reconstructed its organizational implementation model—full terminology comparison and full preliminary translation work that could not be completed due to manpower constraints can now complete the first round of processing within several hours, and the role of terminology workers has shifted from "entry inputters" to "final-draft adjudicators." For China–Uzbekistan agricultural cooperation, a Chinese–English–Uzbek agricultural terminology database with long-term maintenance capability has strategic value no less than a formal bilateral cooperation agreement. It is the foundational support for both sides to conduct in-depth dialogue under a unified conceptual system.

REFERENCES

1. Feng Zhiwei. Introduction to Modern Terminology (Revised and Enlarged Edition) [M]. Beijing: Commercial Press, 2011.
2. GB/T 10112—2019 Terminology Work—Principles and Methods [S]. Beijing: Standards Press of China, 2019.
3. ISO 704:2009. Terminology Work—Principles and Methods [S]. Geneva: ISO, 2009.
4. Rondeau G. Introduction to Terminology [M]. Translated by Liu Gang and Liu Jian. Beijing: Science Press, 1985.

5. Wang Huashu. Translation Technology Tutorial [M]. Beijing: Commercial Press, 2017.
6. China National Committee for Terms in Sciences and Technologies. Termonline [EB/OL]. https://www.termonline.cn.
7. A Chinese–English Classified Agricultural Dictionary [Z]. Beijing: China Agriculture Press, 2017.