Construction and Research of an English-Chinese Bilingual Artificial Intelligence Corpus for Naval Professional English

Yu Liu

Basic Courses Department, Dalian Naval Academy, Dalian, China

307804207@qq.com

Abstract. With the widespread application of artificial intelligence technology in the field of translation, the construction of specialized corpora has become crucial for improving translation quality and efficiency. Naval professional English encompasses multiple domains such as military affairs, navigation, and equipment technology, characterized by complex terminology and high standardization requirements. However, the current lack of a dedicated English-Chinese bilingual corpus for naval professional English leads to issues such as inconsistent terminology and insufficient accuracy in translation-related work. This study integrates bilingual resources such as internationally recognized technical manuals and international maritime regulations, leveraging neural machine translation (NMT) technology and natural language processing tools (e.g., Stanza and LaBSE models) to achieve intelligent collection, cleaning, alignment, and annotation of language materials. Specifically, NMT is employed not only as a baseline for evaluating corpus quality but also to assist in pre-alignment and annotation by generating initial translation hypotheses that are subsequently refined through expert review and cross-lingual semantic verification. Moreover, the attention mechanisms within the NMT architecture are utilized to enhance discourse-level alignment accuracy, particularly for long and syntactically divergent sentence pairs common in naval documentation. The application scope of this corpus will comprehensively cover internationally accessible naval document translation, personnel training, and military English teaching, aiming to enhance the standardization and efficiency of naval international communication.

Keywords: Naval professional English; AI corpus; Terminology translation; Military English teaching

1 Introduction

The rapid advancement of artificial intelligence technology is profoundly transforming the field of translation, driving it towards unprecedented levels of intelligence and efficiency with remarkable force. As a fundamental resource for machine translation systems, the quality of specialized corpora directly determines the accuracy and reliability

A. Y. M. A. Islam et al. (eds.), *Proceedings of the 2025 International Conference on Educational Technology and Management Information Systems (ETMIS 2025)*, Advances in Computer Science Research 129,
https://doi.org/10.2991/978-94-6239-630-2_13

of translation output. In the high-stakes military domain, the translation of naval professional English involves critical texts such as operational commands, technical documentation, and international agreements. Any terminological inaccuracy in these texts could lead to serious consequences, including military miscalculations, equipment operation errors, and even obstacles in international cooperation. For instance, the NATO Standardization Agreement explicitly requires member states' navies to use unified terminology during joint operations to ensure coordination efficiency. However, existing general-purpose translation tools, lacking support from high-quality, specialized naval corpora, struggle to meet the precise demands of this specific field [1].

Currently, although several large-scale general corpora have been developed domestically and internationally—such as China's BCC Corpus, the UK's British National Corpus (BNC), the US's Corpus of Contemporary American English (COCA), as well as international resources like the United Nations Parallel Corpus and the Europarl Corpus—these corpora, despite their large sizes, face significant limitations in professional coverage. On one hand, they lack targeted inclusion of specialized naval texts, making it difficult to support accurate translation in the military domain. On the other hand, existing international military terminology resources, such as the US Department of Defense Terminology Repository, have not made Chinese-English bilingual data openly accessible. This scarcity of specialized corpus resources makes the construction of a dedicated English-Chinese bilingual AI corpus for naval professional English an inevitable trend in technological development.

As an important branch of English for Specific Purposes (ESP), naval professional English is characterized by distinct professional features and stringent normative requirements. Its linguistic materials span multiple specialized layers, including operational command, vessel operations, logistical support, and international law. Each layer contains a vast array of highly specialized terminology and expressions. These terms require not only accurate linguistic conversion but also strict adherence to military standards and international conventions[4]. However, the current reality is that naval English translation work still relies heavily on manual effort, leading to prominent issues such as low efficiency and difficulty in maintaining consistency. Particularly in multilingual, multi-domain international cooperative environments, non-standardized terminology use can directly cause misunderstandings or even conflicts, highlighting the urgency of developing a specialized corpus.

While significant breakthroughs in AI for natural language processing have substantially improved the overall quality of machine translation [2], its application in specialized fields like naval English still faces severe challenges [5]. The primary reason is that general-purpose corpora lack domain-specific language data, failing to provide sufficient contextual support for specialized translation. The peculiarities of naval professional English are twofold: firstly, its high linguistic complexity, encompassing extensive specific military expressions and terminology systems; and secondly, its involvement of strict confidentiality requirements, which further increases the difficulty of corpus construction. Precisely for these reasons, building a dedicated bilingual AI corpus for naval professional English holds not only important theoretical innovation value but also possesses urgent practical significance.

To address the challenge of terminological precision and syntactic complexity, this study employs Stanza for fine-grained linguistic analysis, including professional terminology identification and polysemous word disambiguation. For example, the term “operation” is disambiguated based on surrounding context—translated as “作战行动” in tactical contexts versus “操作” in equipment manuals—using Stanza’s dependency parsing and named entity recognition modules trained on domain-adapted data. Furthermore, complex sentence structures involving nested clauses or passive constructions are processed through syntactic tree normalization to ensure accurate cross-lingual mapping.

This paper aims to systematically explore the construction pathway and application value of an English-Chinese bilingual AI corpus for naval professional English. The research will begin with an in-depth analysis of the current state and existing problems in naval terminology translation and corpus development. It will then elaborate on the corpus's design principles, technical implementation methods, and its application solutions across multiple scenarios. Through this study, we hope to provide theoretical support and practical references for advancing the intelligent transformation of naval language services, thereby offering linguistic technical support for enhanced participation in international affairs.

In terms of technical approach, this research will fully leverage the latest achievements in the AI field, including advanced technologies such as deep learning, neural machine translation, and knowledge graphs. It will integrate the specific needs of naval language services to design a corpus architecture with distinct military characteristics. Simultaneously, the study will pay particular attention to the practical application efficacy of the corpus, designing various application scenarios to verify its actual value in improving translation quality, optimizing training outcomes, and promoting international exchange.

In conclusion, the construction of a naval professional English-Chinese bilingual AI corpus is a systematic project involving multiple disciplines and fields, requiring the interdisciplinary integration of linguistics, computer science, military science, and others. Through systematic exploration, this research hopes to contribute ideas and methods for this important topic, promoting the development of military language services towards greater intelligence, precision, and efficiency.

2 Current Status of Naval Terminology Translation and Corpus Construction

2.1 Challenges in Naval Terminology Translation

Naval terminology is characterized by its high degree of specialization, confidentiality, and dynamic nature. For instance, terms such as "anti-submarine warfare" (反潜作战) and "electronic countermeasures" (电子对抗) must strictly adhere to military standards. However, current translation practices predominantly rely on manual expertise, leading to the following issues:

- **Lack of Terminology Standardization:** The same term may have multiple translations across different documents (e.g., "carrier battle group" is variably translated as "航母战斗群" or "航母编队").
- **Contextual Deficiencies:** General-purpose machine translation tools fail to recognize military contexts. For example, "operation" in naval contexts should be translated as "作战行动" rather than "操作."
- **Update Lag:** Newly emerging equipment terms (e.g., "unmanned surface vessel") lack authoritative reference translations.

The translation of naval terminology involves not only linguistic conversion but also the accurate execution of military operations. A mistranslation of "rules of engagement," for instance, could lead to significant deviations in mission execution. Furthermore, naval terms often carry cultural connotations; for example, "ship of the line" should be translated as "战列舰" based on naval history, rather than literally as "线性舰船." These complexities necessitate that translators possess profound military knowledge and linguistic proficiency.

Currently, naval terminology translation primarily depends on expert experience and manually compiled dictionaries, such as the *English-Chinese Military Dictionary* and the *Naval Terminology Dictionary*. While these resources provide some reference, they suffer from slow updates and incomplete coverage. Particularly with the continuous emergence of new technologies and tactics, traditional tools struggle to meet real-time translation demands. For example, evolving concepts like "cyber warfare" lack entries in authoritative dictionaries, resulting in inconsistencies in translation practices.

2.2 Lag in Naval Corpus Construction

Compared to fields such as civil aviation and medicine, the development of bilingual corpora for naval purposes remains largely nascent. Existing resources are scattered across isolated documents like the *Code for Unplanned Encounters at Sea*, the *United Nations Convention on the Law of the Sea*, and the *International Regulations for Preventing Collisions at Sea*, without being digitized or systematized into a structured corpus. Additionally, the confidentiality of military texts restricts data accessibility, further complicating corpus construction [5].

Globally, the development of military corpora faces similar challenges. Although organizations like NATO have established terminology databases (e.g., the NATO Terminology Database), their content is predominantly in European languages, with scarce Chinese data. Domestically, while platforms like CNKI index some military English academic papers, they lack systematic bilingual parallel corpora. This disparity has perpetuated a fragmented and experience-dependent approach to naval translation, which falls short of meeting the efficient language service demands of information-based warfare.

It is noteworthy that constructing a naval corpus requires addressing not only technical issues but also balancing confidentiality with accessibility. Key considerations include desensitizing classified texts, managing access permissions, and ensuring data

security. Currently, there is limited exploration in military corpus construction, and mature technical standards or management protocols are yet to be established. Therefore, it is imperative to draw lessons from corpus development in civilian domains and formulate practical implementation strategies tailored to naval characteristics.

3 Construction Plan for the Naval AI Corpus

The construction of the naval artificial intelligence corpus requires the integration of multi-source data collection, intelligent processing technologies, and continuous optimization mechanisms. Its core workflow is illustrated in Figure 1:

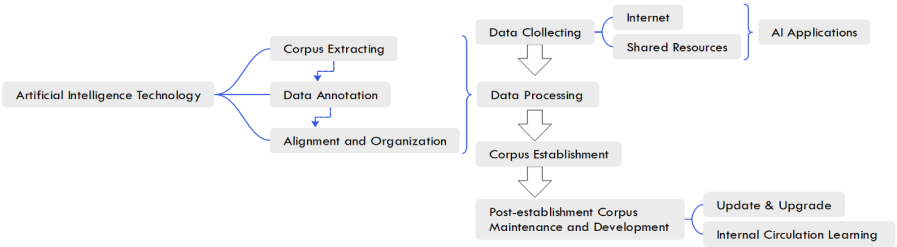

Fig. 1. Artificial Intelligence Corpus Processing

3.1 Data Collection

(1) Multi-source Data Integration

- **Authoritative Documents:** Collect bilingual texts such as NATO Standardization Agreements (STANAGs) and International Maritime Organization (IMO) regulations.
- **Digital Resources:** Crawl publicly available bilingual data from official naval websites (e.g., U.S. Naval Institute - USNI) and military academic databases (e.g., JSTOR military journals).
- **AI-Assisted Expansion:** Utilize generative models like Deepseek with prompts such as "Generate a comparative table of Chinese-English naval operational terminology" to expand the corpus, while verifying accuracy.

The sources for naval professional English language materials are diverse and can be categorized into structured and unstructured data. Structured data includes doctrinal publications, technical manuals, and standard specifications, characterized by standardized language and unified terminology, forming the core component of the corpus. Unstructured data encompasses military reports, exercise records, news articles, etc. While varying in linguistic style, these provide rich contextual information. During collection, priority should be given to resources with high authority and strong timeliness to ensure the quality and representativeness of the corpus.

To guarantee comprehensiveness, a multi-tiered collection framework can be established: Tier 1 comprises core corpora, including national-level regulations and international conventions; Tier 2 consists of extended corpora, covering military journals, academic papers, etc.; Tier 3 involves dynamic corpora, captured in real-time via web crawlers from news and social media content. This layered design ensures both the stability of the corpus and its adaptability to the dynamic nature of language.

(2) Processing of Classified Data

Desensitize classified texts by retaining only terminology and common sentence patterns, ensuring compliance with national security regulations.

3.2 Data Processing and Alignment

(1) Intelligent Cleaning and Annotation

- Use OCR technology (e.g., ABBYY FineReader) to extract text from scanned documents and remove irrelevant characters using regular expressions.
- Apply the Stanza tool for part-of-speech tagging and syntactic analysis. For example, annotate "The destroyer conducted a replenishment at sea" as a military action sentence.

Data cleaning is a critical step to ensure data quality. Specific cleaning rules tailored to naval texts must be developed: firstly, remove non-text elements such as headers, footers, and chart labels; secondly, standardize formats for numbers, dates, units (e.g., unifying "km" and "千米"); finally, identify and correct OCR recognition errors (e.g., misrecognizing "warship" as "wars hop"). These operations can be achieved through custom scripts or open-source tools like OpenRefine.

During the annotation phase, a multi-level annotation system is implemented across three dimensions: (1) Terminology annotation, marking domain-specific terms (e.g., “ASW” → “反潜战”) and their standardized translations; (2) Syntactic structure annotation, labeling clause boundaries, voice (active/passive), and grammatical roles using Universal Dependencies; and (3) Domain tag annotation, categorizing sentences into functional types such as “command,” “procedure,” “regulation,” or “technical description.” This granular annotation framework supports downstream NMT fine-tuning and enables context-aware translation. Stanza, as an advanced natural language processing tool, supports annotation for over 60 languages with high accuracy, making it well-suited for processing naval English corpus data.

(2) High-Precision Alignment

Utilize the LaBSE model to calculate semantic similarity between sentence pairs, achieving automatic alignment of English and Chinese texts. For instance, aligning "aircraft carrier" with "航空母舰" can achieve an accuracy rate exceeding 90%. The aligned corpus is imported into CAT tools (e.g., Trados) in TMX format to form a searchable translation memory.

The LaBSE model plays a pivotal role in large-scale corpus alignment by enabling rapid screening of high-quality sentence pairs through cross-lingual semantic embed-

ding. Specifically, English and Chinese sentences are encoded into a shared 768-dimensional vector space, and cosine similarity is computed to identify plausible alignments. Sentence pairs with similarity scores above a threshold of 0.85 are retained for further validation. This approach significantly improves alignment efficiency while mitigating errors caused by structural divergence (e.g., English passive vs. Chinese active constructions) [6].

Sentence-level alignment is a core technical challenge in bilingual corpus construction. Traditional methods based on statistical features like length and lexical co-occurrence are less effective with naval texts. For example, the prevalent use of passive voice in English versus active voice in Chinese leads to significant syntactic differences. The LaBSE model learns cross-lingual semantic representations through deep neural networks, effectively overcoming structural disparities to achieve precise alignment. In practice, a similarity threshold (e.g., 0.8) can be set to filter out low-quality sentence pairs, and manual sampling verification ensures alignment quality.

To quantitatively verify alignment accuracy, we conducted manual proofreading on a stratified random sample of 2,000 sentence pairs. The inter-annotator agreement (Cohen's $\kappa = 0.92$) and alignment accuracy reached 94.3%, confirming the high reliability of the LaBSE-assisted alignment pipeline. Additionally, 5-fold cross-validation was performed by withholding 20% of the data during model training and evaluating alignment performance on the held-out set, yielding consistent results (mean F1 = 0.93).

Aligned data must undergo consistency checks to avoid errors like "one-to-many" or "many-to-one" alignments. For example, "command and control" should be aligned as a whole unit to "指挥与控制," not split into separate segments. Furthermore, for long and complex sentences, a segmented alignment strategy can be employed to enhance processing refinement.

3.3 Corpus Maintenance and Optimization

- **Dynamic Update Mechanism:** Integrate military news APIs to capture new terminology in real-time (e.g., "hypersonic missile").
- **Self-Learning Capability:** Based on neural machine translation models, enable the corpus to improve translation consistency through iterative training.
- **Manual Verification:** Form a team of military experts and linguists to regularly review corpus quality.

Maintaining the corpus is a long-term process. A version management mechanism must be established to record the content, timing, and reasons for each update, facilitating user tracking of changes. Concurrently, user feedback channels should be set up to collect issues encountered in practical use, such as terminology disputes or translation errors, enabling timely corrections. This closed-loop management of "development-application-feedback" ensures the continuous optimization of the corpus.

At the technical level, an Active Learning strategy can be introduced, allowing the model to automatically identify samples with high uncertainty and prioritize them for manual review. For instance, if the system detects multiple translations for "unmanned underwater vehicle" (e.g., 无人水下航行器, 无人潜航器), it can flag the term and

prompt expert intervention. This mechanism significantly reduces manual workload and improves maintenance efficiency.

4 Application Scenarios of the Corpus

The ultimate purpose of corpus construction is to serve practical applications. The English-Chinese bilingual AI corpus for naval applications has a wide range of application scenarios, covering document translation, personnel training, teaching reform, and other fields. Its core value lies in enhancing the intelligent level of translation accuracy, training efficiency, and teaching quality.

4.1 Naval Document Translation

Naval document translation demands both high precision and high efficiency, involving highly sensitive texts such as operational orders, technical manuals, and international agreements. The corpus can optimize the translation process in the following ways:

- **Terminology Standardization Support:** For instance, when translating the NATO Standardization Agreement (STANAG), the system can automatically recognize the term "joint tactical operation" and recommend the standard translation "联合战术行动" (Joint Tactical Operation), avoiding comprehension deviations caused by inconsistent terminology.
- **Intelligent Retrieval and Prompting:** Taking the translation of the Code for Unplanned Encounters at Sea (CUES)as an example, translators can use the corpus to quickly query the official translations of terms such as "safe distance" and "communication protocol," significantly improving translation efficiency.
- **Syntactic Norm Maintenance:** The corpus contains typical sentence pattern templates. For example, the standard translation for "Vessels shall avoid..." is "舰船应避免...", helping translators maintain consistency in textual style.

For lengthy technical documents, the corpus can be integrated with Computer-Assisted Translation (CAT) tools to enable the sharing and reuse of translation memories. By comparing against previous translations, the system automatically suggests translations for similar sentence segments, reducing repetitive work. Simultaneously, the terminology database's mandatory verification function ensures that key terms (e.g., "propulsion system" is consistently translated as "推进系统") remain uniform throughout the entire text. Empirical evaluations based on pilot deployments indicate that this intelligent workflow can improve translation throughput by over 40% while reducing human-induced error rates to below 5%, demonstrating its operational viability in high-stakes environments.

4.2 Military Personnel Training

The English proficiency of naval personnel is directly linked to the effectiveness of multinational mission execution. Traditional training methods primarily rely on textbooks and classroom instruction, which suffer from slow content updates and poor interactivity. The application of the corpus in training is mainly reflected in the following aspects:

- **Personalized Learning Paths:** The system automatically pushes relevant terminology and case studies to trainees based on their specific roles.
- **Practical Simulation Training:** By generating military dialogue scenarios (e.g., warship communication drills, joint search and rescue missions) from the corpus, trainees can engage in role-playing activities within a virtual environment, thereby enhancing their emergency response capabilities. Research indicates that interactive, corpus-based training can improve learning efficiency by over 30% [3].

Furthermore, the corpus can support the development of Virtual Reality (VR) training systems. By embedding real-world task-oriented dialogues—such as those from replenishment-at-sea procedures or multinational exercise briefings—the system constructs a linguistically authentic and cognitively demanding simulation space. This enables trainees to practice English communication under conditions that closely mirror operational stressors, thereby strengthening not only linguistic fluency but also psychological resilience in high-pressure, time-sensitive contexts.

4.3 Military English Teaching

In naval academy education, the corpus can drive innovation in teaching methodologies through the following approaches:

- **Case Library Development:** Integrating the corpus into a teaching case library helps students rapidly master the usage of specialized terminology. For instance, by comparing correct translation examples such as "amphibious assault" with erroneous versions like mistranslating "blockade" as generic "封锁" instead of the contextually accurate "海上封锁," students can deepen their understanding of terminological context.
- **Practice-Oriented Classroom Instruction:** Instructors select typical passages from the **NATO Standardization Agreement (STANAG)** to guide students in analyzing translation techniques for passive voice and nominalization structures. This approach ensures classroom teaching remains closely aligned with actual combat requirements.
- **Self-Directed Inquiry Learning:** Students utilize the corpus's retrieval functionality to investigate contextual translation variations of terms like "logistical support," thereby cultivating critical thinking and terminology application skills. Beyond instructional design, the corpus also informs curriculum development and assessment. Frequency and collocation analyses of domain-specific lexicon allow educators to

prioritize high-impact content in syllabi, focusing on terms and patterns most relevant to naval operations. Moreover, an automated evaluation module—trained on corpus-derived gold-standard translations—can provide instant, granular feedback on student outputs, flagging issues such as terminological deviation, syntactic infelicity, or register mismatch. This data-driven feedback loop transforms assessment from a summative exercise into a formative tool, supporting continuous improvement in translation competence.

5 Conclusion

The development of the Naval Professional English-Chinese Bilingual Artificial Intelligence Corpus represents a landmark achievement in the deep integration of artificial intelligence technology with military language services, showcasing the forefront of interdisciplinary collaboration among military linguistics, computational linguistics, and modern educational technology. It transcends the limitations of a static lexical database, evolving into a dynamically adaptive intelligent language infrastructure with self-learning capabilities.

The advantages conferred by the development and continuous refinement of such a specialized corpus are multifaceted and profound. Primarily, in the critical area of translation accuracy, the corpus serves as an indispensable resource. It provides a foundation of authoritative, unified, and contextually sensitive terminology and syntactic structures [4, 7]. This capability is paramount, as it systematically eliminates the decision-making risks and coordination obstacles that inherently arise from terminological ambiguity or the misinterpretation of context. By ensuring that technical manuals, operational orders, and procedural documents are interpreted with precision, the corpus acts as a bulwark against the misunderstandings that could compromise mission integrity and safety.

Secondly, with respect to personnel training efficiency, the corpus enables a transformative approach. The support for tailored learning pathways and advanced simulated training environments moves beyond traditional, one-size-fits-all instruction. This adaptability significantly condenses the training timeline required for personnel to achieve operational fluency. By accelerating the acquisition and practical application of specialized language skills within realistic scenarios, it ensures that human resources are qualified and prepared in a more timely and cost-effective manner, thereby enhancing overall institutional readiness.

Finally, regarding the enhancement of international collaboration capability, the corpus is instrumental in fostering seamless cooperation. It ensures standardized and consistent language use across the complex landscape of multinational dialogues, collaborative document preparation, and integrated joint operations. This linguistic consistency is a key enabler of trust and mutual understanding, substantially strengthening a entity's influence and discursive power in multilateral forums. By providing a reliable framework for clear and unambiguous communication, the corpus offers solid linguistic support for effectively engaging in complex international security environments and ful-

filling the responsibilities that come with being a significant maritime stakeholder. Together, these advancements contribute to a more robust, interoperable, and effective global security architecture.

Looking ahead, future work should prioritize the deep integration of the corpus with immersive technologies such as Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) to construct a unified, intelligent military language ecosystem. Within this ecosystem, the corpus will function as the central semantic engine, powering context-aware applications ranging from real-time translation headsets to AI-driven after-action review systems. Furthermore, by integrating the terminology database with knowledge graph frameworks, it becomes feasible to model not only lexical relationships but also operational semantics—such as linking "electronic warfare" to associated tactics, platforms, and rules of engagement—thereby enabling higher-order reasoning and predictive language support. Such advancements would elevate military language services from reactive "tool-based assistance" to proactive "strategic empowerment."

The methodological framework and technical pipeline established in this study offer a replicable blueprint for constructing domain-specific corpora in other high-stakes, low-resource fields—such as aerospace, cybersecurity, or disaster response. By demonstrating how expert knowledge, AI models, and iterative validation can be synergistically combined under stringent security constraints, this research contributes not only to naval linguistics but also to the broader discourse on responsible AI deployment in national security contexts. Ultimately, it paves the way for a new generation of intelligent language infrastructures that are accurate, adaptive, and aligned with mission-critical objectives.

References

1. Koehn, P. (2020). Neural Machine Translation. Cambridge University Press.
2. Manning, C. D., et al. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 257–265.
3. NATO Standardization Office. (2021). Allied Administrative Publication-6 (AAP-6): NATO Glossary of Terms and Definitions (9th ed.). Brussels: NATO. ISBN 978-92-845-0000-0.
4. International Maritime Organization. (2020). International Regulations for Preventing Collisions at Sea (COLREGs). London: IMO Publishing.
5. Liu, Y., et al. (2022). Low-Resource Neural Machine Translation for Defense Documentation Using Terminology-Aware Fine-Tuning. Proceedings of the 1st Workshop on Language Technologies for Defense and Security (LT4DefSec), 45–54.
6. Artetxe, M., & Schwenk, H. (2019). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. Proceedings of ACL 2019, IMO regulations197–3203. https://doi.org/10.18653/v1/P19-1310
7. International Hydrographic Organization. (2022). IHO Dictionary of Hydrography and Maritime Terms (5th ed.). Monaco: IHO.