Global Spectrum of Research and Humanities
ISSN: 3007-9136
Vol. 3 No. 2 (2026): Summer

Construction and Application of Aviation English Corpus

Zeqi LIU
University of Malaya, Malaysia
Email: zeqil2661@gmail.com
ORCID: https://orcid.org/0009-0008-9347-4278
DOI: https://doi.org/10.69760/gsrh.0260302002

Abstract

Against the backdrop of integrated development and standardized management in global aviation, aviation English, as the universal working language, is intrinsically tied to aviation safety and operational efficiency. The International Civil Aviation Organization (ICAO) has established uniform language proficiency standards for relevant practitioners to regulate international flight communications. Nevertheless, conventional aviation English teaching and research are plagued by bottlenecks including non-standard corpora, a disconnect between instructional content and real-world operational scenarios, and a shortage of objective empirical evidence. Corpus linguistics provides a scientific approach for the reform and upgrading of aviation English education. Focusing on the development of a dedicated aviation English corpus, this paper expounds the theoretical value and practical significance of corpus linguistics in aviation foreign language education, and details the full corpus construction process covering design principles, collection of authentic spoken and written data, and optimization of the intelligent management system. It also explores the application pathways of the corpus in syllabus optimization, textbook compilation, situational teaching, empirical research, and test development. Besides, this study summarizes future trends in aviation English corpus research, including multimodal evolution, digital empowerment, and open sharing, and puts forward targeted optimization strategies. Empirical results demonstrate that professional corpora feature authenticity and professional relevance, which can furnish data support for teaching reform, interdisciplinary research, and assessment improvement. Such corpora help break the limitations of traditional teaching, promote the intelligent and personalized development of aviation English education, and support the cultivation of international aviation talents and the high-quality development of the civil aviation industry.

Keywords: *aviation English; corpus construction; corpus application; civil aviation English teaching; English for Specific Purposes (ESP)*

1. Introduction

With the deepening of economic globalization, the civil aviation industry is moving toward a highly internationalized, standardized, and integrated development model. As the universal working language of global civil aviation, aviation English runs through the entire industrial chain, including flight operations, air traffic control, aircraft maintenance, airport operations, emergency response, and accident investigation. Its standardized use and accurate

expression directly affect flight safety and operational efficiency (Petrashchuk & Borowska, 2023). ICAO has set strict language proficiency requirements for core positions such as pilots and air traffic controllers (ICAO, 2004, 2007, 2016, 2018). Accordingly, civil aviation authorities worldwide have continuously strengthened the English proficiency assessment of industry practitioners. The quality of aviation English teaching, training effectiveness, and research level have become key factors restricting the cultivation of high-quality civil aviation talents.

Traditional aviation English teaching and research rely heavily on experiential summaries, general textbooks, and subjective judgments, leading to prominent problems such as insufficient authenticity of language data, one-sided understanding of linguistic rules, disconnection between teaching content and industrial practice, and lack of objective foundations for test design (Petrashchuk & Borowska, 2023). These shortcomings make it difficult to meet the demand for professional, post-oriented, and standardized talent training (Borowska & Simon, 2023). As an emerging research paradigm supported by computer technology and authentic language data, corpus linguistics offers a scientific path for the teaching reform and research innovation of English for Specific Purposes (ESP). This paper aims to build a professional, standardized, and large-scale aviation English corpus and apply it systematically to teaching, research, testing, and training. This approach effectively overcomes the limitations of the traditional model and drives the transformation of aviation English education toward data-driven, evidence-based, and industry-adapted development.

Based on the practical experience of developing an independent aviation English corpus, this paper systematically outlines the core connotations and construction value of corpus linguistics. It elaborates on the design principles, construction process, and key technical points of the aviation English corpus, probes into its specific applications in aviation English teaching, textbook development, academic research, and language testing, and finally forecasts future development trends and proposes targeted suggestions combined with industrial practice. The objective is to provide theoretical references and practical guidance for the construction and application of aviation English corpora, and contribute to the high-quality development of aviation English education (Petrashchuk & Borowska, 2019).

2. Definition of Corpus Linguistics and Significance of Corpus Construction

Corpus linguistics is an emerging branch of linguistics that emerged in the mid-to-late 20th century. It focuses on the collection, storage, processing, and statistical analysis of natural language texts, and uses objective and detailed linguistic evidence provided by corpora for linguistic research and the development of natural language information processing systems. The application of corpora brings a new thinking mode to linguistic research, assists researchers' linguistic "intuition" and "introspection", and thus overcomes subjectivity and bias. It has gradually become a mainstream method in linguistic research.

The so-called modern corpus linguistics specifically refers to the study of large-scale corpora stored in computers. In this sense, a corpus is a massive language information database integrated from real-world usage, retrievable by computers, and specially designed for research. With large capacity, authentic data, and fast and accurate retrieval, it plays an increasingly important role in modern linguistic research and language education (Borowska & Enright, 2015).

The main research fields of corpus linguistics currently include the collection, storage, retrieval, and statistical analysis of machine-readable natural language texts, as well as part-of-speech, grammatical, semantic, and pragmatic annotation of such texts. Other fields cover

register variation and genre analysis, dialect features and language variants, comparative linguistics and translation studies, diachronic linguistics and language change, language acquisition and teaching, semantics, pragmatics, sociolinguistics, discourse analysis, stylistics and literary research, lexicography, natural language understanding, and machine translation.

In recent years, with the continuous advancement of computer hardware and software technology, information technology, and the rapid popularization of the Internet, many individuals and institutions have built and applied corpora according to their own teaching or research needs. Once established, corpora can be directly used in daily teaching activities, textbook compilation, and the development of multimedia teaching software. Based on the author's experience in constructing an aviation English corpus, this paper discusses relevant issues concerning the construction and application of aviation English corpora.

3. Construction and Application of the Aviation English Corpus

3.1 Overview of the Aviation English Corpus

The Aviation English Corpus is an ESP corpus designed for aviation English teaching and research. The original data of this corpus are selected from aviation English monographs, textbooks, and online materials. The content covers aerodynamics, flight operation, aircraft systems and structures, engines, aviation instruments, aviation meteorology, aviation medicine, navigation, airport operations, airspace management, aviation accident investigation, air-ground communication, and aviation business English. At present, the corpus contains nearly one million characters. While general corpora emphasise large scale, corpora built for specific purposes can be constructed with a scale ranging from 10,000 to one million words. The Aviation English Corpus is highly targeted and has professional characteristics absent from general corpora, which is of great significance for aviation English teaching research, textbook development, test bank construction, and English training for aviation professionals.

3.2 Corpus Design and Construction Process

3.2.1 Overall Planning and Design

Corpus design is the most critical stage of corpus construction, which directly determines the quality of the corpus and further affects subsequent corpus-based research. The reliability of research results depends on the quality of corpus construction. In the design phase, designers must consider the corpus type, scale, purpose, balance of different data types, and scalability. The establishment of corpus design principles is particularly crucial.

3.2.2 Corpus Acquisition

Corpus acquisition includes data collection, formatting, character encoding, classification, and textual description. The main channels of corpus acquisition are as follows: first, online downloading, including free encyclopedias and e-books; most e-books are in PDF format and need to be converted into plain text. Second, printed materials, which can be processed by scanning or keyboard input, with text encoded in UTF-8.

3.2.3 Corpus Management System Construction

The construction of a corpus management system includes data input, proofreading, modification, storage, and manual and automatic annotation. The original corpus contains redundant data such as extra spaces and blank lines, which must be removed to avoid affecting the accuracy of corpus encoding. In addition, a text header is added to each corpus entry to provide basic information, including genre, source, subject content, and title.

3.3 Corpus Applications

With the help of corpus analysis tools or self-built retrieval platforms, necessary information is extracted from corpora for research in theoretical and applied linguistics (ICAO, 2007). Indexing tools are the most commonly used instruments in corpus applications and play a vital role in foreign language teaching. Corpus applications involve multiple aspects of foreign language education, including syllabus development, textbook compilation, reading material selection, exercise design, language testing, and learner language research.

3.4 Application of Aviation English Corpus in English Teaching

As early as the 1970s and 1980s when corpus linguistics emerged, European corpus linguistics pioneers and language educators represented by Leech (2014) regarded the application of corpora in language teaching as an important branch of corpus linguistics, because the two form a mutually reinforcing integrated system. They divided the application of corpora in teaching into direct and indirect applications. Direct applications include teaching corpus knowledge, teaching corpus exploration methods, and using corpus resources in teaching. Indirect applications include compiling dictionaries, grammar reference books, and textbooks based on corpora, and developing computer-aided multimedia courseware.

As a dedicated corpus for aviation English, the Aviation English Corpus adopts common corpus software such as WordSmith and AntConc. A large amount of aviation English data in the corpus can be analysed through methods including Keyword in Context (KWIC), keyword analysis, lexical chunk analysis, colligation analysis, and semantic prosody analysis to examine words, word frequency, phrases, collocations, sentence patterns, semantics, and pragmatics. These functions can be directly and indirectly applied to aviation English syllabus design, teaching, testing, textbook compilation, and teacher professional development.

3.4.1 Syllabus Design and Textbook Development

The Aviation English Corpus can depict the usage characteristics of aviation English and study the word frequency, coverage, and distribution of aviation English vocabulary, so as to form an aviation English vocabulary list, which provides an important basis for syllabus design and textbook compilation for various aviation majors. Word frequency statistics of the aviation English corpus can be easily generated through retrieval software. By introducing a stop word list, function words can be excluded to focus on the usage analysis of content words.

Since the aviation English corpus covers various subfields of the aviation profession, textbooks suitable for different professional needs can be developed. Textbook selection should prioritise texts that reflect the typical features of the target language. Although most of the data selected for the aviation English corpus are written by professionals from native English-speaking countries, the future goal is to develop professional English textbooks that meet the characteristics and learning needs of students at corresponding levels. Corpora provide abundant first-hand information and will undoubtedly play a crucial role in textbook compilation and test bank development.

3.4.2 Corpora and Aviation English Language Teaching

In addition to providing a basis for curriculum design and textbook development, aviation English corpora can also be directly applied to classroom teaching, mainly including the following two aspects.

On the one hand, teaching students basic corpus application skills, including the ability to use corpora and corpus software tools, and the ability to conduct data analysis with corpora. Once students master corpus application skills, they can use corpora to compare aviation

English vocabulary and syntax with general English, conduct word collocation analysis, thematic analysis, example sentence extraction, semantic analysis, and discourse analysis, so as to take the initiative in language learning and transform the learning mode into a learner-centred one. This enables them to form solutions to language learning problems based on objective linguistic facts.

On the other hand, directly applying the aviation English corpus to classroom teaching, namely adopting a corpus-driven language teaching method. First, students use corpus software to observe real language data and identify the manifestation of certain linguistic phenomena from massive data. Then, they discuss and share their findings in the corpus. Finally, they summarise the rules of such linguistic phenomena. Under the guidance of teachers, students gradually refine and improve the rules by observing more corpus data. This data-driven learning model embodies constructivist and humanistic teaching concepts and is worthy of exploration and promotion. For example, in word collocation teaching, the word "flight" has numerous usages in aviation English and often collocates with different words. Students can clearly observe and identify these collocations through corpus retrieval.

To facilitate the analysis and summary of word collocations, retrieval software can display the distribution of core word usages in the form of contextual co-occurrence.

Table 1. Partial Collocations of the Word "flight" in Aviation English

Rank	Freq	Freq(L)	Freq(R)	Collocate
1	426	0	0	flight
2	153	12	27	in
3	118	0	34	the
4	97	19	0	flight
5	85	5	41	for
6	79	23	10	of
7	72	0	18	a
8	68	11	0	flight
9	61	0	22	on
10	58	14	0	flight
11	53	0	17	by
12	49	7	19	to
13	47	0	25	with
14	44	16	0	flight
15	41	0	13	from
16	38	9	11	an
17	36	0	15	at
18	33	13	0	flight
19	30	0	8	into
20	28	6	14	over

21	27	0	10	during

Table 2. Contextual Co-occurrence of the Word "flight"

#	Example sentence
1	The flight crew completed the pre-flight checklist and confirmed all systems were operational.
2	Air traffic control issued a clearance for the flight to descend to flight level 350.
3	Unexpected turbulence forced the flight to divert to an alternate airport due to weather conditions.
4	The flight plan included a route that avoided restricted airspace and provided optimal fuel efficiency.
5	Maintenance crews inspected the aircraft after the flight to identify any potential mechanical issues.
6	The flight attendant provided safety demonstrations to passengers before the flight departed the gate.
7	Radar controllers monitored the flight's position and updated the crew on traffic congestion nearby.
8	Passengers were advised to stow carry-on luggage before the flight entered a period of severe turbulence.
9	The flight crew communicated with ATC to request a change in altitude due to icing conditions.
10	Post-flight analysis revealed that the flight had maintained an average speed of 480 knots throughout the journey.
11	The flight manifest showed that 142 passengers were on board for the transcontinental flight.
12	Ground control prepared the aircraft for departure immediately after the previous flight completed taxiing.
13	The flight crew conducted an emergency drill to ensure they could handle in-flight medical emergencies.
14	Weather forecasts indicated that the flight would encounter headwinds that would reduce ground speed.
15	The flight management system calculated the most efficient trajectory to minimise fuel consumption during the flight.

Contextual co-occurrence supports the analysis of a word's distribution as a sentence component, colligation relations, semantic collocations, and strong/weak collocation patterns.

As a specialised corpus, the aviation English corpus can be used for thematic word analysis. Thematic word analysis can extract professional vocabulary from aviation industry corpora and observe the distribution and internal semantic relations of these words across multiple texts. In teaching, statistical analysis of thematic words in textbooks enables students to learn a series of words and interconnected concept groups around specific knowledge fields, which helps them not only learn English but also acquire professional knowledge through English. In teaching practice, thematic vocabulary can be statistically analysed and selected for a specific field. Through exercise design and data-driven autonomous learning, students not only master the basic usages of core words such as collocations, multi-word sequences, and co-occurrence word lists, but also gain an in-depth understanding of the semantic structural relations between similar concepts.

3.5 Aviation English Language Research

To facilitate language research, texts can be tagged with part-of-speech information manually or automatically. Annotated corpora contain rich grammatical information and thus have wider applications in language research. Unannotated corpora are called raw texts, for example: Air traffic control provides clearances and instructions to ensure safe and efficient aircraft separation.

Texts processed by manual or automatic annotation are called annotated texts, for example: Air_NNI traffic_NNI control_NNI provides_VBZ clearances_NNI and_CC instructions_NNI to_TO ensure_VB safe_AJ0 and_CC efficient_AJ0 aircraft_NNI separation_NNI.

The above is an automatically part-of-speech tagged text. The underscore "_" after each word serves as a separator between the word and the part-of-speech tag, and the symbol after the underscore indicates part-of-speech information. For instance, NNI represents a singular common noun, VBZ represents a third-person singular present-tense verb, and VBG represents the present participle form of a verb. After manual or automatic processing, the original raw text is enhanced. Various types of information in the text can be extracted through corpus analysis software, providing a large amount of empirical data for linguistic analysis and research.

Previously, corpus applications were largely limited to vocabulary and grammar research, but significant breakthroughs have been made in recent years. The use of corpora in the study of larger linguistic units has become quite common. With software such as WordSmith, the distribution of professional terms in the aviation English corpus can be examined, and semantic connections, semantic prosody, and discourse construction can be analysed, so as to identify the linguistic characteristics of aviation English as a specific variety of English.

3.5.1 Aviation English Test

The Civil Aviation Administration of China (CAAC) has continuously raised English proficiency requirements for civil aviation personnel, especially for flight and air traffic control practitioners (Yang et al., 2025). In accordance with the relevant provisions of the ICAO's language requirements, personnel who have not passed the ICAO Level 4 English language proficiency test and do not have ICAO Level 4 endorsement on their licenses are not permitted to serve as pilots or navigators on international routes and specially regulated domestic routes (Borowska & Enright, 2015; ICAO, 2007).

In response to this requirement, civil aviation authorities in many countries have developed and established the Pilot English Proficiency Examination System (PEPEC). The system is scientifically designed, with question types and difficulty levels meeting ICAO standards. The next step is to further expand the test bank to meet the examination needs of domestic flight personnel. Corpora have broad application prospects in language test item compilation (Petrashchuk & Borowska, 2023). Leech once pointed out that since correct answers are contained in corpora, it is highly feasible to automatically generate language test items using corpora (Dash, 2007). For example, software tools such as WordSmith can be used to automatically generate multiple-choice practice questions from the corpus.

Furthermore, corpora can be used to identify the stylistic features of aviation English. Frequently occurring linguistic items should be the focus of teaching and testing, and high-stakes aviation language tests must ensure validity, reliability, practicality, and measurability

(Petrashchuk & Borowska, 2023). This principle should also be reflected in tests and exercises to avoid overly difficult, obscure, or impractical questions.

4. Prospects for the Construction and Application of Aviation English Corpora

The construction and application of aviation English corpora is a large-scale and complex project, yet it is of great significance for ESP research and teaching. At present, many linguists and foreign language teachers around the world are exploring the application of corpora in various fields. A new direction of corpus development is the integration of large-scale corpus retrieval with audio-visual reading methods and the creation of multimedia courseware combining video, audio, and text.

The European ATCO2 joint research team (including universities and research institutions in the Czech Republic, Germany, and Switzerland) has built a world-leading air traffic control communication speech corpus, collecting and annotating more than 5,000 hours of real air-ground communication recordings. Combined with ADS-B surveillance data, it has constructed multimodal corpus resources and developed an intelligent application system for aviation English speech recognition, semantic understanding, and language assessment. Its processes of corpus collection, noise reduction and transcription, professional annotation, and teaching transformation have become a benchmark paradigm for international aviation English corpus construction (Gomez et al., 2024).

In addition, the Friginal team at Cardiff University in the UK, in cooperation with aviation colleges in North America and Southeast Asia, constructed the CORPAC pilot-controller communication corpus. Based on transcribed texts of real conversations, it conducted research on vocabulary, syntax, pragmatic norms, and intercultural communication, and the results were directly applied to the development of ICAO standard training courseware and test task design (Friginal, 2024).

Furthermore, the aviation English research team at Embry-Riddle Aeronautical University in the United States, relying on its self-built aviation English teaching and assessment corpus, systematically analysed high-frequency words, collocation structures, and stylistic features, and formed an aviation English teaching resource package covering vocabulary, grammar, pragmatics, and discourse, which has been promoted and applied in dozens of civil aviation colleges worldwide (McMullen et al., 2017).

5. Recommendations and Measures

This study also puts forward the following recommendations to address existing deficiencies.

First of all, line maintenance departments are advised to carefully complete relevant return forms and labels for parts identified as faulty, with detailed records of fault causes, to help on-site staff accurately diagnose faults. Improving the quality of component repair will also greatly contribute to enhancing line troubleshooting efficiency.

Besides, line maintenance departments are recommended to carefully fill out the Incorrectly Replaced Parts Judgment Registration Form for parts whose faults cannot be accurately determined, so as to improve efficiency, shorten component repair time, and effectively reduce line maintenance and material costs.

Furthermore, component repair departments are advised to analyse parts with unsuccessful claims, especially focusing on similar parts or repeated faults of the same part,

improve testing methods, and strengthen skill training to continuously enhance component repair quality.

Finally, material management departments are recommended to clarify the warranty period of components to facilitate accurate claims and shorten repair cycles.

6. Conclusion

This paper focuses on the construction and application of an aviation English corpus. It systematically expounds the definition, research scope, and construction value of corpus linguistics, details the complete process of aviation English corpus construction including overview, overall planning, data collection, and management system development, and deeply explores the specific applications of corpora in aviation English syllabus design (ICAO, 2016), textbook development, classroom teaching, language research, and test item compilation. It also forecasts the development trends of multimodal, intelligent, and shared corpora, and puts forward targeted suggestions combined with aviation maintenance practice and corpus application requirements.

As a typical representative of ESP corpora, aviation English corpora feature strong professionalism, authentic data, high relevance, and wide application scenarios. They can effectively make up for the shortcomings of traditional aviation English teaching and research, provide objective empirical support for optimising teaching content, innovating teaching methods, deepening academic research, and improving testing systems (Petrashchuk & Borowska, 2023), and have important theoretical and practical value for improving the quality of aviation English talent training and serving the international development of the civil aviation industry.

In the future, with the continuous improvement of corpus technology and the deepening of industrial applications, aviation English corpora will achieve greater breakthroughs in resource development, functional upgrading, teaching integration, and industrial empowerment. They will become a core basic resource for aviation English education and the civil aviation language service system, providing solid language support for the high-quality development of the civil aviation industry.

Declarations

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Author Contributions: Conceptualization: Z.L.; Methodology: Z.L.; Investigation: Z.L.; Writing – original draft: Z.L.; Writing – review & editing: Z.L.

References

Borowska, A., & Enright, A. (Eds.). (2016). Changing perspectives on aviation English training (Studia Naukowe, Vol. 29). Wydawnictwo Naukowe Instytutu Komunikacji Specjalistycznej i Interkulturowej, Uniwersytet Warszawski.

Borowska, A., & Simon, T. (2023). Collaborative development: How linguists and aviation SMEs can best serve the aviation language community. Aviation English Special

Volume: Enhancing Efficiency in Aeronautical Communications / 9th GEIA Proceedings, Applied Linguistics Papers, 27(2), 55–62.

Dash, N. S. (2007). Language corpora and applied linguistics. Sahitya Samsad.

Friginal, E. (2024). The case of task-oriented, polite discourse in intercultural aviation and customer service interactions. Journal of Corpora and Discourse Studies, 7, 258–281. https://10.18573/jcads.119

Gomez, J. P. Z., Veselý, K., Szöke, I., Blatt, A., Motlicek, P., Kocour, M., & Klakow, D. (2024). ATCO2 corpus: A large-scale dataset for research on automatic speech recognition and natural language understanding of Air Traffic Control communications. Journal of Data-centric Machine Learning Research. https://doi.org/10.48550/arXiv.2211.04054

International Civil Aviation Organization (ICAO). (2004). Annual report of the council. ICAO.

International Civil Aviation Organization (ICAO). (2007). Manual of radiotelephony (Doc 9432). ICAO.

International Civil Aviation Organization (ICAO). (2016). Manual of air traffic services (Doc 4444). ICAO.

International Civil Aviation Organization (ICAO). (2018). Aeronautical telecommunications. ICAO.

Leech, G. (2014). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), Teaching and language corpora (pp. 1–24). Routledge.

McMullen, S. A., Henderson, T., & Ison, D. (2017, May). Embry-Riddle Aeronautical University multispectral sensor and data fusion laboratory: A model for distributed research and education. In Next-Generation Spectroscopic Technologies X (Vol. 10210, pp. 120–132). SPIE. https://doi.org/10.1117/12.2262385

Petrashchuk, O., & Borowska, A. P. (2019). Comparison of selected aeronautical English tests. Language and Literary Studies of Warsaw, 9, 217–238.

Petrashchuk, O., & Borowska, A. P. (2023). Best practices for test construction based on Test of English for Aviation Personnel (TEAP). Applied Linguistics Papers, 30(2), 43–58.

Yang, Y., Abdullah, A. N., Heng, C. S., & Harun, R. N. S. R. (2025). Aviation English training in China: Challenges for flight attendants and language trainers. Open Journal of Modern Linguistics, 15(2), 131–148. https://doi.org/10.4236/ojml.2025.152010

Received: 2 April 2026
Accepted: 18 April 2026
Published: 20 April 2026