

Corpus-aided Lexical Retrieval and Analysis of a Telecommunication Corpus

Yahui Jin^{1*}, Lihua Xu¹

¹ School of Humanities and Foreign Languages, Xi'an University of Posts and Telecommunications, Xi'an, China

Email Address

1679270348@qq.com (Yahui Jin), xulihua@xupt.edu.cn (Lihua Xu)

*Correspondence: 1679270348@qq.com

Received: 10 March 2024; Accepted: 12 April 2024; Published: 30 April 2024

Abstract:

Based on *Artificial Intelligent Techniques for Wireless Communication and Networking*, the article builds its own wireless communication corpus, and selects other English texts in non-communication fields as reference corpus, and uses the corpus software AntConc to retrieve and count the specialized vocabulary and subject words of communication texts, taking the retrieval results of communication engineering English as the basis of analysis. The search results of communication engineering English are taken as the basis of analysis, and the characteristics of specialized vocabulary of communication texts are derived by combining the method of qualitative analysis. The results of the study show that: (1) the vocabulary density and length of communication texts are higher than the vocabulary of general-purpose texts, which is manifested in the high density of terminology and the large number of professional real words; (2) the vocabulary of communication texts needs to be combined with the professional context to reflect the terminological specification of the terminology of the text within a word with multiple meanings in the profession; (3) in the frequency comparison, AI(578), logistic(165), algorithm (150), the frequency counts of the words reflect the connection between communication engineering and computer aspects, thus reflecting the importance of technological empowerment of communication technology. This study explains the characteristics of specialized vocabulary in the field of communication on the basis of data support provided by corpus construction, which facilitates English learners in the field of communication engineering to master the expression of the terminology and the summarization of the core vocabulary in the field, which is of theoretical significance to the study of vocabulary characteristics in the specialized field of communication engineering and has application value in the construction of the communication-related corpus and the abstraction of the terminology and specialized vocabulary.

Keywords:

Corpus, Communication Engineering English, Subject Words, Lexical Retrieval

1. Introduction

The quantitative extraction and qualitative analysis of vocabulary in communication writings using corpus search [1] combines quantitative corpus analysis and qualitative

discourse analysis in a “balanced” way in the analysis process [2]. Under the perspective of corpus linguistics, the article adopts the corpus research method to build its own corpus with the book *Artificial Intelligent Techniques for Wireless Communication and Networking* as the research object, and utilizes the corpus retrieval software AntConc4.0.4.0 to analyze the word frequency (word) and word frequency (word) of the book, and then analyzes it with the help of the corpus search software AntConc4.0.4.0. The word frequency (word) function of the corpus retrieval software AntConc4.0.4.0 is used to carry out character counting to analyze the lexical features of the communication text; then the word frequency distribution and collocation are used to analyze how the communication text embodies the vocabulary’s specialization and accuracy; and then the search for the communication-related subject words leads to the combination of vocabulary in the field of communication with other aspects as well as the importance of the technological empowerment in the field of communication to enrich the lexical analysis and research of this field.

British linguist Baker (1993) proposed that “the tools and methods provided by the corpus are complementary to traditional translation research rather than replacing it” [3], advocating the combination of corpus linguistics and translation theory [4,5], this research method provides great convenience for domestic and foreign language research [6,7], and can provide authentic and accurate information for the analysis of communication vocabulary. It can provide real linguistic data for the analysis of communication vocabulary, and thus the characteristics of communication-based vocabulary derived from the analysis of real samples are innovative. The corpus-based or corpus-driven research paradigm [8] can reduce the subjective reliance on single-text analysis and make the academic value of language research more convincing.

With the rapid development of the contemporary electronic communication field, the communication technology represented by artificial intelligence, big data, Internet of Things, etc. effectively promotes the intelligent development of society and the efficiency of production and life. The study of English terminology in this field is crucial for learners of communication engineering, especially for the reading of overseas communication engineering books can more directly help learners understand the development of the professional frontier. In addition, IT professionals, researchers and academics working on 5G communication and network technologies are linked to technical assessments and comparative analyses of existing technologies, and then experimental case studies related to real-time scenarios are illustrated. The extraction of the corpus based on specific communication engineering books makes the study more relevant and enables to refine the learners' knowledge and mastery of the English terminology of artificial intelligence and wireless communication technologies under the field of communication engineering.

2. Research Methodology and Procedures

2.1. Corpus Selection and Data Collection

This study is based on a self-constructed English monolingual corpus consisting of the book *Artificial Intelligent Techniques for Wireless Communication and Networking*, and utilizes the corpus retrieval tool AntConc4.0.4.0 to extract the corpus and analyze the data, and the number of form characters of the corpus is 102811, and the number of class characters is 11948. The number of form characters is 102811,

and the number of class characters is 11948. In addition, the American Brown corpus (Brown) is selected as the reference corpus, which is an earlier and more stable reference corpus. As the first standard corpus, the standards adopted by the Brown Corpus are one of the norms followed in corpus linguistics research. As the authoritative research data of corpus linguistics, it has classic research value. English texts in the field of communication, as part of scientific and technological texts, are characterized by the rigor and accuracy of scientific and technological English. Therefore, the data derived from the text analysis of this reference corpus are more detailed and reliable, and after comparative analysis, the differences between English in the field of communication engineering and general texts as well as the lexical usage characteristics can be found.

2.2. Tool Selection

The article chooses AntConc 4.0.4.0 as the research tool of the article, plays the function of N-Gram to retrieve the target corpus and the reference corpus word clusters, uses the collocate function to retrieve the functional units of the target vocabulary contextual collocations, and at the same time, with the help of the keyword function to retrieve and sort the subject words of the target corpus in order. The keyword function is also used to search and sort the subject words of the target corpus.

2.3. Research Issues

(1) What features of the English vocabulary in the field are reflected in the frequency distribution of specialized vocabulary in communication texts?

(2) Which lexical collocations in the target corpus reflect the importance of technological enablement in the field of communication?

3. Word Frequency Retrieval and Analysis based on AntConc

The word frequency (word) function of AntConc was utilized to summarize the word frequency statistics in the self-constructed corpus in Table 1 after removing the dummy words in the top 50 ranked items.

Table 1. Word Frequency Data Sheet.

Number	Word	Rank	Frequency
1	data	12	618
2	AI	14	578
3	network	15	554
4	wireless	17	508
5	communication	27	367
6	system	33	294
7	techniques	34	290
8	model	40	259
9	supply	40	259
10	artificial	44	250
11	routing	47	242

As can be seen from Sheet 1, excluding dummy words, it is found that the number of nouns in the top 50 high-frequency words shown in Table 1 totals 10, which are data, network, wireless, AI, communication, system, techniques, model, artificial, and supply. In order of frequency, from the largest to the smallest In descending order of frequency, they are data (618 times), AI (578 times), network (554 times), wireless

(508 times), communication (367 times), system (294 times), techniques (290 times), supply (259 times), model (259 times), artificial (250 times). artificial(250 times) Thus, it can be seen that communication discourse, as a branch of scientific and technological discourse, tends to use conceptual vocabulary with practical significance to introduce a specific domain.

The word communication is usually translated as “communication” in the field of communication, while it has the meaning of “communication; exchange; dissemination; news” in the general text. In addition, among the words appearing in the search results, artificial intelligence and other words are obviously specialized nouns, which basically appear in communication texts and are less common in other fields. In view of this phenomenon, the vocabulary of English texts in the field of communication engineering can be divided into two categories, the first of which is a word with multiple meanings with different meanings from those in the common texts, which has a prompting effect on learners related to the field of communication, i.e., the words that appear as common words but not common meanings should be summarized in the texts to form an inherent cognition. The second category is specialized terms that appear more frequently in the vocabulary of the field of communication and less frequently in other generic texts.

3.1. Thematic Word Analysis

Utilizing the function of contextual co-occurrence (KWIC) in AntConc 4.0.4.0, the retrieval yielded the first 15 items of Table 2 related to wireless communication before and after the 5 word count.

Table 2. Total results for “wireless communication”.

Left Context	Hit	Right Context
Learning Era for Future 6G	Wireless communication	Theory, Applications
Way to a new prosperous	Wireless communication	Networking technology
Has propelled the industry of	Wireless communication	Forward the challenges
Structure deals with 6G	Wireless communication	Enabled deep learning

As shown in Table 2, when the word “wireless communication” has a vocabulary value of 5 in the target corpus, the words appearing on the right side are mostly related to theory, application, technology, virtual reality, etc., which means that the word “wireless communication” appears in this corpus in the form of theory popularization and explanation, and it is combined with the explanation of the theory and the use of technical software in this field. This indicates that the term “wireless communication” appears in this corpus in the form of theory popularization and explanation, and it is combined with the explanation of the theory of the field and the use of technical software. In addition, the term wireless communication appears together with future technologies, future 6G, and deep learning enabled 6G, which reflects the current technological development in the field of communication and highlights the importance of technological empowerment.

The keyword list in Table 3 reveals that the field of communication engineering has been combined and developed in the field of logistics, artificial intelligence technology, and mobile communications, which shows that the field covers a wide range of other technological dimensions and highlights the importance of communication engineering.

Table3. Comparison of the Top 100 Keywords in Communication with the Frequency of the Reference Corpus.

Words	Rank	Frequency of Occurrence in the Target Corpus	Frequency of Occurrence in the Reference Corpus	resemblance	Effect
data	1	618	3	460.237	0.012
AI	2	578	1	447.159	0.011
network	3	554	1	428.116	0.011
wireless	4	508	0	403.874	0.01
communication	5	367	3	263.428	0.007
system	6	294	3	206.683	0.006
techniques	7	290	3	203.584	0.006
routing	8	242	0	192.191	0.005
artificial	9	250	1	187.733	0.005
supply	10	259	2	186.694	0.005
model	11	259	3	179.61	0.005
intelligence	12	227	2	161.782	0.004
logistics	13	165	1	121.011	0.003
IEEE	14	147	0	116.699	0.003
algorithm	15	150	1	109.285	0.003

Table 3 reflects the combination of the communication field with logistics, artificial intelligence technology, and mobile communication aspects that have revolutionized the modern logistics industry, such as the use of wireless sensor networks and mobile communication technology to achieve real-time monitoring of the location of goods and accurate logistics management and tracking. While providing reliable data transmission and information exchange, communication technology also provides solid technical support for the intelligent and automated development of the logistics industry, and plays an important role in enhancing the competitiveness and efficiency of the logistics industry. Mobile communication technology, on the other hand, realizes the timely transmission and communication of logistics information and improves the efficiency of collaborative work.

In summary, the search of subject terms plays an important role in grasping the subjectivity of the text and its relevance to other fields. According to the frequency of the occurrence of the subject words can be found with the text related to the above core vocabulary, so as to guide the reader to the text of the core vocabulary, to help the learners in this field to grasp the core content of the target text, and to get a certain degree of cognitive enhancement of the professional knowledge of the field of communications, from the depth to the breadth of the knowledge.

3.2. Lexical Collocation

For the real words with high word frequency occurrences in the previous operation, the article carries out further retrieval of them, i.e., retrieving the collocations of high-frequency real words using collocate and N-Gram functions. In this article, the collocation of high-frequency real words is retrieved using the collocate and N-Gram functions.

Table 4. Search for “algorithm”.

Collocate	Rank	Likelihood	Effect
algorithm	5	53.304	3.192
encryption	6	52.836	5.173
detection	17	30.526	2.798

For example, the retrieval of collocation words with algorithm as the center word yields the retrieval results in Table 4, indicating that when introducing algorithmic strategies and technologies, they are inseparable from encryption and detection technologies. Meanwhile, the search of collocation words with algorithm as the center word reveals that the word algorithm has more collocations with words such as encryption, detection, etc., which indicates that when introducing algorithmic strategies and technologies, they are inseparable from encryption and detection technologies, and the search of collocation frequency with intelligence as the center word reveals that the top three words are encryption and detection technologies, respectively. The top three terms are artificial, machine and automation. Among them, intelligence and artificial form a phrase with the meaning of “artificial intelligence”, which is common in the field of communication, while the second and third search results show that The second and third search results illustrate the connection between artificial intelligence technology in the field of automation and machine building. Communication technology improves productivity by enabling real-time data transfer and facilitating collaboration between devices. In the automation process, communication networks can be used to monitor and control equipment, enabling remote operation and data collection. Meanwhile, artificial intelligence technology enables systems with intelligent decision-making and adaptive capabilities by means of machine learning and deep learning, realizing intelligent production processes and adaptive manufacturing systems, and improving the flexibility and efficiency of industrial automation.

The retrieval of the term logistic using the N-Gram function illustrates that communication technology provides the basis for technical support of the logistics model, for example, the use of GPS positioning technology and mobile communication networks to achieve real-time tracking of vehicles and goods, and to improve the efficiency and safety of transportation.

4. Conclusions

The article uses AntConc software to search the vocabulary aspects of the representative English monographs in the field of communication, to get the information about the use of communication vocabulary in the texts, and analyzes them based on the statistical data, and finds that the characteristics of English vocabulary in the field of communication are mainly as follows: first, from the point of view of vocabulary content, the English texts in the field of communication, as the books of specialized fields, involve vocabulary with strong specialization, and have proprietary meanings in the field. field has proprietary meanings, so professional learners in this field should pay attention to the memorization of proprietary vocabulary meanings when reading English texts in the field of communication in order to avoid misunderstandings; secondly, there are significant differences between texts in the field of communication and other general texts in terms of vocabulary density, degree of nominalization, and central words, and compared with the general texts, the vocabulary of the field of communication as a whole has the characteristics of high vocabulary density, high degree of nominalization, more common phrases formed by two words, and more center words. Thirdly, the above features indicate that English vocabulary in the field of communication engineering is more abstract and specific in meaning. According to the frequency statistics and collocational word co-occurrence found that the communication industry, as one of the important infrastructures in China, is deeply integrated with the development and widely used in

the field of computer artificial intelligence and in the aspect of logistics field, which indicates that the development of the communication business is closely related to the advancement of communication technology. Limited to space, the article can not represent all the core vocabulary in the field of communication engineering, but it facilitates the learners in this field to summarize the vocabulary related to artificial intelligence technology and wireless communication, improves the cognition of the characteristics of English vocabulary of communication majors, promotes the international academic exchanges in the communication technology on this basis, and enriches the research of the English aspect of this field.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by Shaanxi Provincial Philosophy and Social Science Research Special Project (2023HZ0967) “Construction and Talent Cultivation of First-class Specialties in Business English”, Project Leader: Xu Lihua; also was supported by Shaanxi Provincial Social Science Fund Project (2023K014) “Study on the Influence Mechanisms of International Students’ Bicultural Identity” and External Communication in Shaanxi Province Research on “Bicultural Identity and Influence Mechanism of Foreign Communication of International Students Coming to Shaanxi”, Project Leader: Xu Lihua.

References

- [1] McEnery, T.; Brezina, V.; Gablasova, D.; Banerjee, J. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics*, 2019, 39, 74-92.
- [2] Paquot, M.; Gries, S.T. (Eds.). *A Practical Handbook of Corpus Linguistics*. Springer Nature. 2021.
- [3] Wallis, S. *Statistics in Corpus Linguistics Research: A New Approach*. Routledge. 2020.
- [4] Fazal, M.A.M.H.Z.; Moavia, H. Formulaic Language in Social Sciences: A Functional Analysis of Lexical Bundles in Native and Non-Native Academic Discourse. *Pakistan Social Sciences Review*, 2019, 3(1), 234-249.
- [5] Shirazizadeh, M.; Amirfazlian, R. Lexical Bundles in theses, Articles and Textbooks of Applied Linguistics: Investigating Interdisciplinary Uniformity and Variation. *Journal of English for Academic Purposes*, 2021, 49, 100946.
- [6] Huang, L.P. Reflections on corpus translator style research. *Foreign Language Teaching*, 2018, 01, 77-81.
- [7] Wang, Y.Q.; Liu, H.T. Progress and foresight in data-driven corpora research. *Modern Foreign Languages*, 2023, 05, 700-710.
- [8] Jin, B.X.; Wei, N.X. Corpus paths in discourse research: methods, challenges and prospects. *Foreign Languages and Foreign Language Teaching*, 2023, 01, 1-11+144.



© 2024 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)