A Corpus-based Analysis of Linguistic Features of IALA Publications

Li Qiang**, Ren** Zhehui

Fuzhou Aids-to-Navigation Department of Eastern Navigation Service Center, Ministry of Transport, Fuzhou, Fujian, 350004 China

Key words: IALA, corpus, communication, linguistic features, AntConc, navigation

ABSTRACT

In order to help people in Asian countries to have a better understanding of IALA publications, and to participate in information exchange more effectively, the research is mainly taken to study the linguistics features of IALA publications. The IALA standards, recommendations and guidelines of recent five years are used to establish a corpus to investigate the linguistic parameters such as the type-token ratio, keyword list, word clusters, passive voice, modal verbs and so on; and the language rules and application patterns are summarized by taking the in-depth analysis of language characteristics of IALA publications.

1. Introduction

1.1 Background

Language is an indispensable communication tool when people participate in activities of international organizations and carry out technical exchanges. Language in specific situations will affect people's communication and information exchange. IALA is an important international organization in the field of navigation, which conducts events, conferences, and release publications in English. However, most Asian are non-native speakers of English. Therefore, in order to help people in Asian countries to have a better understanding of IALA publications, and to participate in information exchange more effectively, this study carries out a scientific and in-depth analysis of the language characteristics of IALA publications by establishing a corpus, which is able to help to understand the content thoroughly and master the writing style and characteristics.

1.2 Literature Review

1.2.1 Corpora

Originally, corpora were used by linguists to represent large amounts of naturally occurring language data that could be used as a basis for language research; A corpus now refers to raw text stored on a computer, or processed text with labeled information; As a technical term, corpus refers to the electronic one that collects text data on a large scale by using computer technology according to certain linguistic principles and specific language research purposes (Liu, 2020). In a word, it refers to bringing together all the texts to be studied for processing and analysis by the software.

It has been illustrated by Qin (2021) that corpora mainly have the following characteristics: firstly, it is large in scale, and it is not rare to find a corpus with storage capacity ranging from hundreds of thousands of words to billions of words; secondly, it is representative. The selected samples can be guaranteed to represent the characteristics of the research content; thirdly, it is stored in electronic form, which can be tagged and annotated

according to the demand to facilitate automatic retrieval, query and statistics, and support empirical research.

1.2.2 Relevant research in maritime field

The output of corpus-based research in maritime field in China are mainly reflected in three aspects: One is the corpus-based study on targeted aspects of maritime English, such as textual connectives analysis, compound word analysis, context and collocations of specific synonyms, comparative analysis of verbs, etc. Secondly, corpus can be used to study the linguistic features of specific literary styles in the maritime field, such as the key words of IMO news bulletins and the linguistic features of maritime investigation reports, so as to explore the correctness and standardization of the writing of maritime investigation reports. Thirdly, the corpus is used to study the linguistic features of international maritime conventions. For example, the main vocabulary usage of the resolutions of the International Maritime Organization Conference was studied, and the stylistic features of the resolutions as legal English are analyzed through the lexical features.

2. Methodology

The IALA standards, recommendations, guidelines publicated on the IALA official website in the recent five years were collected as the profiling data, and the data that might affect the analysis results were pre-processed. After the word segmentation and part-of-speech tagging for the data, the AntConc software was used to make the processed text into a corpus, which in this study is called *IALA Corpus*. Then, through investigating the parameters of the type-token ratio, keyword list, word clusters, passive voice and modal verbs, the corresponding linguistics features of IALA documents could be studied.

3. Data Collection

The data collection process was arranged in the following steps: Firstly, download the IALA standards, recommendations, guidelines publicated on the IALA official website in the recent five years (a total of 118 documents). Secondly, the text was pre-processed to delete the data that might affect the analysis results, such as duplicate document titles and identifiers in the header and footer, address and email information of IALA headquarters, publication date and IALA website, etc. Thirdly, word segmentation is carried out to avoid lexical adhesion due to file format problems. For example, "itis" is identified as a word "itis", which will lead to analysis errors. Fourth, the part-of-speech was tagged to each word of the collected data through CLAWS part-of-speech Tagger website, so as to facilitate the in-depth analysis of the data. Finally, after the text dealt with by the above steps, the text was input into the AntConc software to get a corpus.

4. Analysis

4.1 Type-token Ratio

The number of Type and Token is obtained by AntConc, as shown in Figure 1:

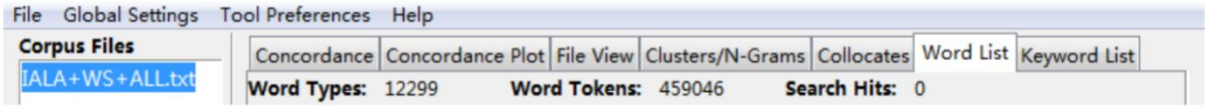


Figure 1: The number of Type and Token (Type: 12299; Token: 459046)

Calculation is made according to the following formula:

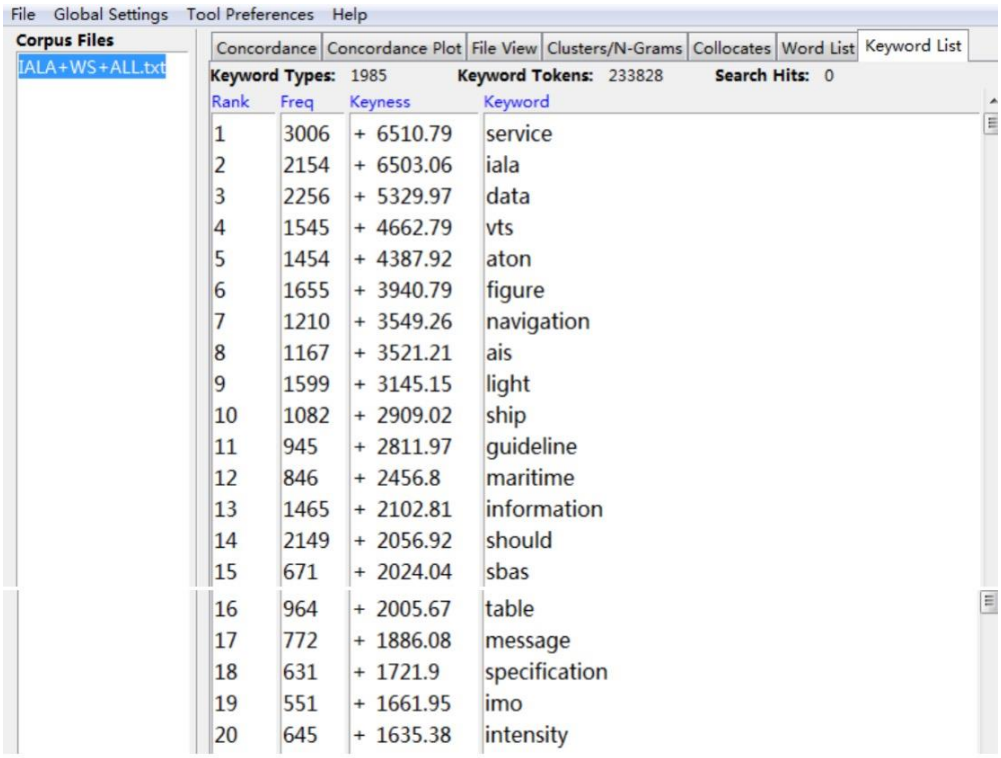
Type-token ratio (TTR) = Type/token *100%

TTR value can be easily calculated to be 2.68, and TTR value of Brown corpus (Large-scale comprehensive English corpus) is 4.21 (Zhang, 2012). In contrast, TTR value of IALA corpus is significantly lower. To a certain

extent, TTR can reflect the lexical variability of corpus (Song, 2020). As a British large-scale comprehensive corpus, Brown corpus contains rich stylistic and linguistic types, which is able to reflect the universality of language. However, the IALA corpus focuses on language related to navigation field, and its core word-base is relatively small, thus the vocabulary variation is relatively weak.

4.2 Keyword List

The keyword list is formed in terms of keyness (Table 1), and the top 20 words in the list with the highest correlation with IALA corpus are classified into three categories (see Table 2) :



Rank	Freq	Keyness	Keyword
1	3006	+ 6510.79	service
2	2154	+ 6503.06	iala
3	2256	+ 5329.97	data
4	1545	+ 4662.79	vts
5	1454	+ 4387.92	aton
6	1655	+ 3940.79	figure
7	1210	+ 3549.26	navigation
8	1167	+ 3521.21	ais
9	1599	+ 3145.15	light
10	1082	+ 2909.02	ship
11	945	+ 2811.97	guideline
12	846	+ 2456.8	maritime
13	1465	+ 2102.81	information
14	2149	+ 2056.92	should
15	671	+ 2024.04	sbas
16	964	+ 2005.67	table
17	772	+ 1886.08	message
18	631	+ 1721.9	specification
19	551	+ 1661.95	imo
20	645	+ 1635.38	intensity

Table 1: Keyword list extracted by AntConc

Theme Content	Data Presentation	Language Function
iala, vts, aton, ais, guideline, maritime, service, navigation, light, ship, sbas, imo, information, message	data, figure, table, specification, intensity	should

Table 2: Keyword categories classified by using context

The illustration of the Theme Content shows the characteristics of IALA text, which focuses on the professional navigation language, and presents the core content of IALA corpus. It could be discovered from the Data Presentation that IALA standards, recommendations and guidelines are good at using tables, figures and similarly forms to present data and related content. Additionally, IALA publications are focusing on professional navigation field, so as to the formulas, the equations, and the diagrams are used to show the results (Liang, 2016), and the words regarding which appear in the keyword list. The use of the modal verb ‘should’ also partly explains the strong non-coercive tone of the IALA publications.

4.3 Word Clusters

The software AntConc is used to retrieve the clusters of 3-5 words. The following table (Table 3) is the 50 word clusters with highest frequency in the corpus:

NO.	Clusters	NO.	Clusters
1	aids to navigation	26	the purpose of
2	of the service	27	the service provider
3	the use of	28	with respect to
4	marine aids to navigation	29	are to be
5	in order to	30	of e navigation
6	the service specification	31	service data model
7	vessel traffic services	32	in the table
8	of the light	33	can be found
9	can be used	34	the context of
10	need to be	35	be found in
11	the light source	36	prior to the
12	based on the	37	safety of life
13	to ensure that	38	the aton light
14	in accordance with	39	it should be
15	be used to	40	that can be
16	the provision of	41	life at sea
17	depending on the	42	the safety of
18	shown in figure	43	may be used
19	description of the	44	to be noted
20	as well as	45	as described in
21	should be considered	46	be able to
22	the number of	47	it may be
23	the vts area	48	should be used
24	it is important	49	a number of
25	the competent authority	50	the issue of

Table 3: Word clusters (sequenced in frequency: from most to least)

According to the classification method of the keyword list, the word cluster list is classified into three categories:

A. Word clusters expressed theme content, such as *aids to navigation, marine aids to navigation, vessel traffic services, of the service, the service specification, of the light, the light source, the provision of, the vts area, the service provider, of e navigation, Service Data Model, Safety of Life, the Aton Light, Life at Sea, the Safety of,* etc., which is to introduce the content of professional navigation field involved in IALA publications and highlight the core content that IALA concerns.

B. Word clusters expressed data presentation, such as *Shown in figure, description of the, the number of, in the table, as described in, a number of,* etc., which is to outline the main presentation manner of data in IALA publications.

C. Word clusters expressed language function, such as *the use of (can be used, be used to, maybe used, should be used), in order to, need to be, based on the, to ensure that, in accordance with, depending on the, as well as, should be considered, it is important, the purpose of, with respect to, are to be, can be found, the context of, be found in, prior to the, it should be, That can be, to be noted, be able to, it may be, the issue of,* etc., which is to summarize the commonly used functional expressions in IALA publications.

The retrieval of word clusters is an extension research of the study of keyword list. From the above word clusters, it could be seen that IALA publications focus on the passive voice, such as *Are to be, to be noted, should be used,* etc. Secondly, the use of logical phrases such as *in order to, prior to,* shows that IALA publications pay more attention to language logic and coherence. Relevant studies show that passive voice is frequently used in academic articles and scientific articles (Yue, 2019). Therefore, IALA publications can be considered to have the similar language characteristics with academic articles and scientific articles to a certain extent.

4.4 Passive Voice

As mentioned above, the use of passive voice in IALA publications is stressed, so that the regular expression of passive voice is applied to further explore the use of passive voice in IALA corpus. The frequency of passive voice regular expression (\S+_VB\w*\ S (\S+_[RX]\w+\ S)*\S+_V\wN\ S) (Liang, 2009) retrieval in AntConc software is 7692. It is estimated that 47.5% sentences of the 118 IALA documents are passive voice.

Based on the *Concordance Plot* function of AntConc, the overall distribution of passive voice in IALA *corpus* could be discovered (Figure 2). The overall distribution of the other commonly used sentence structure "it is+ adjective" in IALA *corpus* is also retrieved (Figure 3).

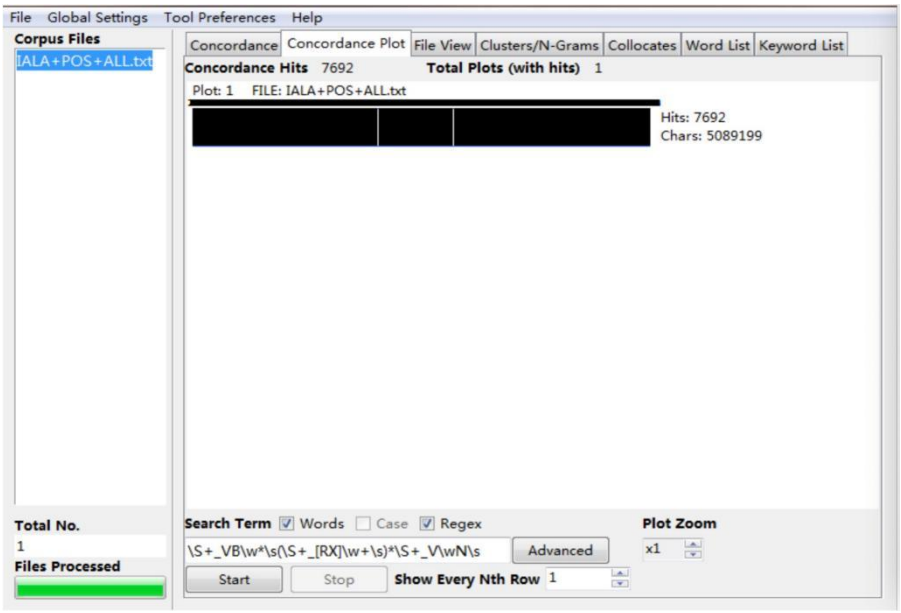


Figure 2: Overall distribution of passive voice in IALA corpus

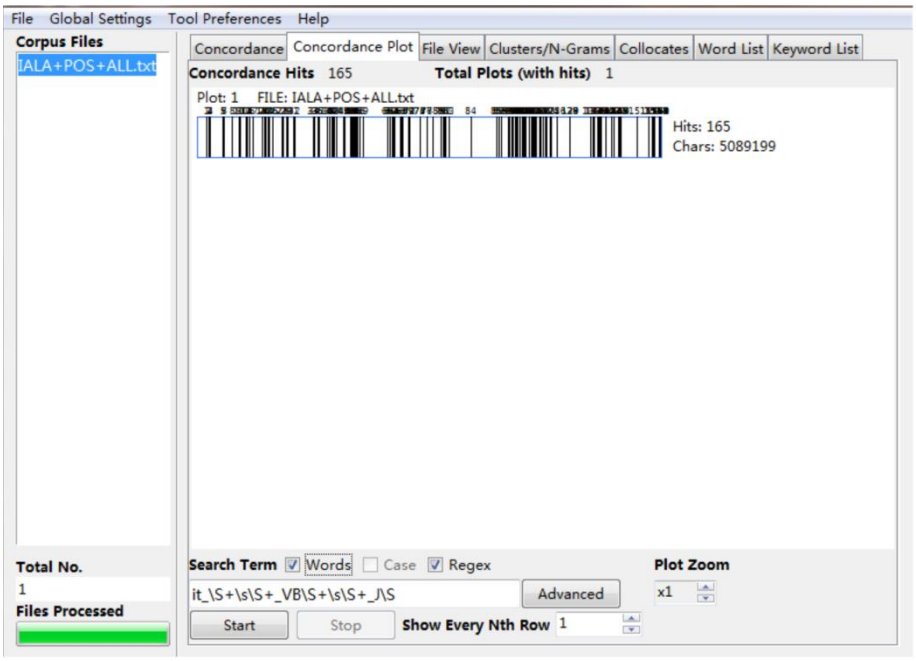


Figure 3: Overall distribution of "it is+ adjective" structure in IALA corpus

By comparing the distribution situations, the overall distribution of passive voice in IALA corpus is even and significant. Through the scientific analysis of the frequency and Concordance Plot of passive voice, the use of passive voice is equally distributed in IALA corpus, which could be judged that IALA publications attach great importance to the use of passive voice. The use of passive voice is able to help strengthen cohesion and coherence of the text, promote expression objectively and highlight the focus content (Lin, 2017).

4.5 Modal Verb

It could be found that the modal verbs such as *should, can, may* appear in the above research. The retrieval was taken to study the modal verbs, and the results are shown as follow in the order of weak to strong (Wang, 2014):

Modal Verb	Frequency
may (might)	1481
can (could)	2027
should (shall)	2566
must	372

Table 4: Frequency of commonly used modal verbs (from weak to strong)

The data shows that *can* and *should* are used in high frequency, while *must* is used in the lowest frequency, which reflects the non-mandatory characteristics of IALA standards, recommendations and guidelines. Meanwhile, IALA has a strong tone of requirements for technical specifications, which is closely related to its responsibility of developing, improving and coordinating the affairs of maritime navigation aids system.

5. Results and Discussion

In this study, IALA standards, recommendations and guidelines published in recent five years are collected as the research data, and AntConc software is used to build the corpus. Through retrieval of theme words, word clusters, passive voice and modal verbs, a series of linguistic features of IALA publications were preliminarily explored: (a) The core word-base of IALA text is relatively small, and the vocabulary variation is relatively weak; (b) What IALA concerns most is the expertise content about navigation, formulas and diagrams are commonly used

in the text to display data results; (c) IALA publications often use logical phrases to promote the logic and coherence of the text; (d) The use of passive voice is widely and evenly distributed in IALA texts; (e) *Can (could)* and *should(shall)* are the most commonly used modal verbs, while *must* and other modal verbs with a very strong tone are used cautiously.

Based on the above findings, it is suggested to pay attention to the application of the following language features in IALA-related paperwork: (a) Focusing on professional content, rather than sentence variations and flowery rhetoric; (b) Be good at using charts, table and figures to present data and research results, so as to express opinions effectively highlight key points concisely, which is able to enhance readability; (c) Pay attention to the use of logical phrases to promote coherent and cohesion of the expression; (d) Good at using passive voice to write, the use of passive voice is able to promote the expression objectively and formally; (e) It is strongly recommended to keep consistent with the style of IALA publications when using modal verbs, and to use modal verbs with strong mood, such as *must*, with caution.

References

(1) Liang, M. (2009): Ci xing fu ma yu liao ku de jian suo yu zheng ze biao da shi de bian xie [Retrieval corpus with part-of-speech tagging and regular expression editing]. *Foreign language education in China*, **2**, 65-73+81.
(2) Liang, M. (2016): Shen me shi yu liao ku yu yan xue [What is corpus linguistics]. *Shanghai foreign language education press.*
(3) Lin, H., Wu, M. and Feng, Y. (2017): Ji yu yu liao ku de guo nei ying yu xue xi zhe bei dong yu tai shi yong de dui bi fen xi [A corpus-based comparative analysis of the use of passive voice by Chinese English learners]. *Journal of Shenyang Architecture University (Social Science Edition)*, **19**, 524-529.
(4) Liu, H. (2020): Yu liao ku yu yan xue: li lun gong ju yu an li [Corpus linguistics: theory, tools and cases]. *Foreign language teaching and research press*.
(5) Qin, H. (2021): Shuang yu yu liao ku de yan zhi yu ying yong [Building and using bilingual corpora]. *Foreign language teaching and research press*.
(6) Song, Q.(2020): Ji yu yu liao ku de mo yan xiao shuo yi ben feng ge yan jiu [A corpus-based study on the translation style of Mo Yan's novels]. *Foreign language teaching and research press*.
(7) Wang, X. (2014): Ji yu yu liao ku de guo ji hai shi gong yue qing tai dong ci de shi yong yan jiu [A corpus-based study on modal verbs usage in international maritime conventions]. *Dalian Maritime University*.
(8) Yue, Y. (2019): Ji yu yu liao ku de zhong guo ying yu zhuan ye yan jiu sheng xue shu xie zuo bei dong yu tai shi yong yan jiu [A corpus-based study on the use of passive voice in academic writing of Chinese master students majored in English]. *Dalian Maritime University*.
(9) Zhang, L. (2012): Ji yu yu liao ku de ying yu ke ji lun wen wen ti te zheng fen xi: yi li nian guo ji shu xue jian mo te deng jiang lun wen wei li [A Corpus-based Stylistic Analysis of Scientific Papers in English: A case study of the grand prize papers in International Mathematical Modeling]. *Journal of Changsha Railway University (Social Science Edition)*, **13**, 179-181.

Author's Biography

Li Qiang, Bachelor of Engineering, is currently engaged in navigation management and research on international AtoN affairs in Fuzhou AtoN Department of Eastern Navigation Service Center, Ministry of Transport.
Ren Zhehui, Master of Arts, is currently engaged in navigation management and research on international AtoN affairs in Fuzhou AtoN Department of Eastern Navigation Service Center, Ministry of Transport.