ISSN:2790-167X DOI: 10.56028/aehssr.3.1.260

Construction of Corpus for Hydraulic Purpose: Supporting Industry Research, Education, and Development

Zhaoqin Ji [1, a], Honghui Huang [2, b], Cancan Xu[2, c], Min Xu [3, d]

[1] Technology Journal and Information Center, Nanjing Hydraulic Research Institute, Nanjing, 210029, China;

[2,3] Nanjing Hydraulic Research Institute, China.

[a] jizhaoqin@nhri.cn, [b] huanghonghui@nhri.cn, [c] ccxu@nhri.cn, [d] 122382397@qq.com

Abstract. Corpus-based approaches and tools are effectively and widely used in language studies for specific purposes. A developed corpus with a large collection of specialized textual examples can help users find language characteristics and patterns of the industry. Building a Chinese-English bilingual hydraulic industrial corpus allows practitioners and students to get access to definitions, terminology usage, relevant academic research findings, and ongoing projects, so as to support industry research, education, and development. NHRICorpus is a platform that consists of a Chinese-English bilingual corpus and a term search engine. Terms in the corpus were retrieved from internal academic reports of Nanjing Hydraulic Research Institute. Users can launch into the platform to search for hydraulic engineering terms, information, and practices. To improve user experiences, eye-tracking technologies and theories are applied in web page design.

Keywords: hydraulic industry, corpus construction, term search platform, eye-tracking

1. Introduction

Since its first presence in 1967, corpora have been widely used in various industries to support language research, regulate language teaching, standardize language use, and help industry development trend analysis [1]. Terminologies and expressions used in different industries varied. Linguists recognized and summarized such differences and created industrial corpora to clearly delineate language characteristics and patterns in relevant industries. With a large collection of textual examples, industrial corpora offer a wider context of language use. Electronic industrial corpora make it possible for users to discover the occurrences of specific structures or words, and get access to examples of the nuances of the meaning. Therefore, users can easily find out whether a structure or word is used differently according to the subject, location, register of the text, and more.

Currently, a majority of open-source corpora and corpus-related studies focus on the literature, business, law, and computer science industries. Although there are comprehensive corpora that cover all walks of life, terms in the hydraulic industry are hardly included.

The term hydraulic engineering first appeared in Master Lv's Spring and Autumn Annals, referring to water resource management for fishing. After hundreds of years of development, in 1934, the Chinese Hydraulic Engineering Society defined the content of hydraulics as flood control, drainage, irrigation, hydraulic power, waterway, water supply, sewage, and port works. A few years later, water and soil conservation, water resource protection, environmental water conservancy, environmental hydraulic engineering, and reservoir fisheries are gradually added to the content of hydraulic engineering.

According to the Global and China Hydraulic Industry Report, 2021-2026, [2] the hydraulic industry in China started late but developed at a fast pace. As of October 2021, according to the National Bureau of Statistics, China has accommodated over a thousand hydraulic parts manufacturers. Liu Weiping, Vice Minister of Water Resources, told a press conference that, during the first nine months of 2022, China started construction on 42 major water conservancy projects, absorbing investment of over CNY1.9 trillion.

China has expanded its overseas presence via larger, smarter international investments and better performance in hydraulic engineering practices. To maintain the momentum, China has attached great importance to hydraulic scientific research, academic studies, engineering projects, and higher education. Therefore, developing a hydraulic corpus is necessary for practitioners and students to gain a more extensive understanding of the characteristics and patterns of language used in the hydraulic industry, so as to help them understand professional terms, write academic papers, and analyze industry development trends.

Easy-to-use webpages with a clear structure can extensively improve user experience. According to Bergstrom and Schall [3], webpage users experience two phases, the scanning phase and the inspection phase, when visually searching for information, and both phases are impacted by webpage features, such as color, font, size, and location, of objects on that page. For instance, typing a word on a web page in bold is a good way to attract user attention. To gain authentic feedback from users, eye-tracking technology is widely applied. However, since it is even more difficult to get access to professional facilities and labs during the pandemic, I failed to implement standard eye-tracking experiments. Instead, I tried some theories in page design and invited five participants to browse the pages without eye-tracking facilities, recorded their eye motions, and interviewed them after using the term search platform. The accuracy can be impacted but some of the theories have been verified as useful.

2. Roles and Relevant Studies of Industrial Corpora

2.1 Role of Bilingual Industrial Corpora in Industry Research

Industrial corpora can help users find research and technical hotspots, screen out useful and time-sensitive information, and summarize and predict industry development trends. By matching frequently-mentioned keywords, names of locations, and technologies with years and authors, users can preliminarily, and efficiently understand current academic trends, academic leaders, and popular technologies of a specific research direction or a specific author's research focus during a time period. Compared to traditional quantitative approaches by retrieving information from open-source databases such as CNKI or Springer, industrial corpora can give more accurate results since disturbances have been screened during the data cleansing process. Also, after a corpus is completely built, information can be efficiently and accurately searched, retrieved, quantified, and measured as long as the internet is connected. In addition to all advantages and strengths of industrial corpora, the bilingual corpus includes term names, definitions, and examples in both languages, which meet practitioner and student needs for language learning. Bilingual corpora can help users, especially students, write English papers. Since the hydraulic industry develops fast in China, new expressions and terms will appear when technological and theoretical breakthroughs occur and language characteristics can also change quickly. Therefore, examples in industrial corpora should be updated on a regular basis.

2.2 Role of Bilingual Industrial Corpora in Industry Education

Corpus linguistics was applied in language teaching since the late eighties and early nineties [4]. With more people becoming computer literate, corpus-based approaches in language teaching are much easier to be introduced. Also, China has achieved higher international influence in hydraulic engineering practices. The SMART Rivers 2022 was held in Nanjing for the first time, and CAE academician HU Ya'an commented in his speech that young scholars were the new force for international hydraulic development in the future. The young generation should put their focus on exploring professional breakthroughs as well as prepare themselves with enhanced language skills for better performance on the international stage. We sorted out the key nodes of all 76 universities in China, including Nanjing Hydraulic Research Institute, Hohai University, Tsinghua University, and China Three Gorges University, which are authorized to issue master's and doctoral degrees in hydraulic engineering, hydrology and water resources, hydraulics and river dynamics, hydraulic structure engineering, water resources and hydropower engineering, and harbor, coastal and offshore engineering, to analyze the development of hydraulic education in China over the past century. By 2000, near 75% of the 76 universities established hydraulic departments or majors, while only about 23% of them started to cultivate postgraduate students. The number of master's and doctoral degree subjects soared since 2011, and reached 181 master's degree subjects and 53 doctoral degree subjects in 2022. In 2021, Li Guoying, Minister of Water Resources, emphasized the long-term policy regarding the development of the hydraulic industry in the Fourteenth Five-Year Plan of China. China will put more effort into cultivating high-level hydraulic talents in the future. Master's and doctoral subject development situations are shown in Figure 1.

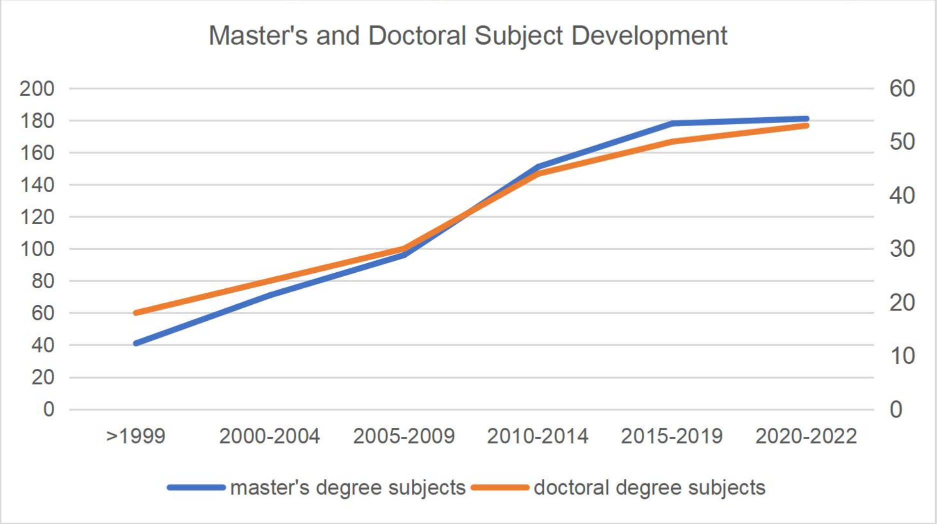


Fig. 1. Master's and doctoral subject development

2.3 Role of Bilingual Industrial Corpora in Industry Development

Industrial corpora also help study and analyze industry development trends. Significant word frequency plunges imply either outmoded technologies or focus changes. On the contrary, the increased presence of a certain technology or theory can be used to predict the upcoming trend in the industry.

2.4 Relevant Studies Regarding Industrial Corpora

We applied quantitative document approaches to analyze corpora construction and corpus study development in various industries. We took the search strategy of "discipline + corpus" (45 disciplines in total) to search for nine types of academic publications, including papers published in journals from both inside and outside China, master's and doctoral dissertations, Chinese and international conference papers, Chinese and English books, and others (research reports, standards, and technical documents). The 45 disciplines, including history, language and linguistics, performing art, visual arts, economics, and more, are listed in Wikipedia under the term academic discipline. More than 30,000 documents were found, and after preliminary data cleansing, a total of 21,469 documents were prepared for further analysis. Of those, language and linguistics, computer science, law, business, and literature rank top five among all disciplines involved. We found that corpora-related studies developed slowly before 2002 and met a drastic increase since 2008. Industrial corpora–related study is undergoing fast development. Corpus-related studies development by industry and year are shown in Figure 2.

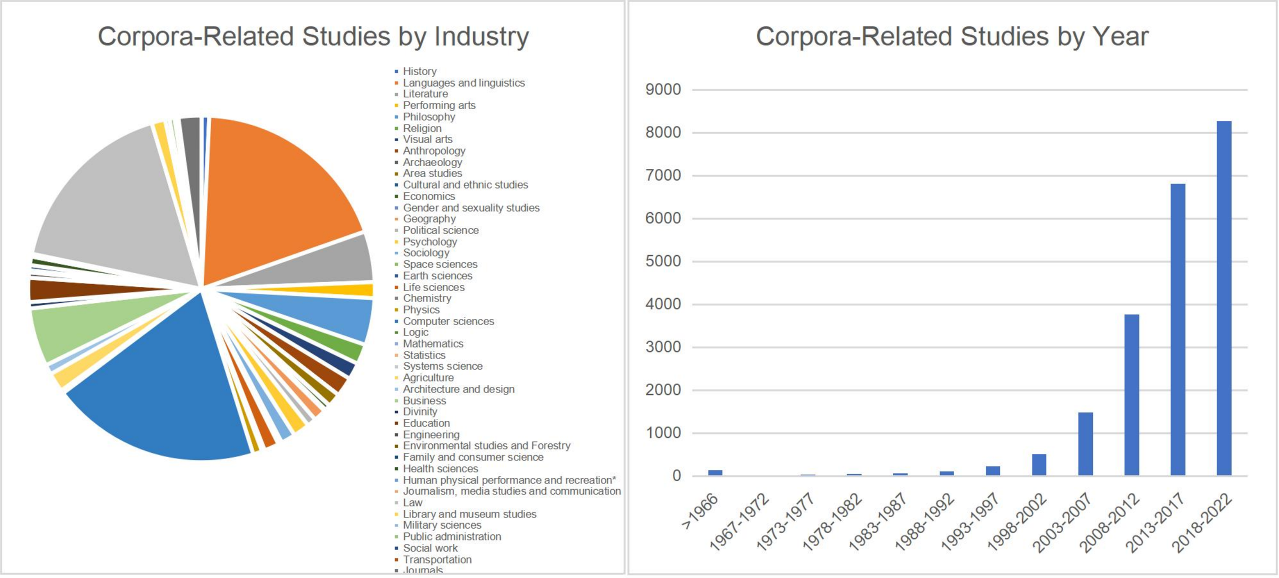


Fig. 2. Corpora-related study development by industry and year

3. Preparation for the Construction of the NHRICorpus

3.1 Corpus Construction Objectives

In 1988, when defining two variable characteristics and four absolute characteristics of English for specific purposes (ESP), Strevens explained that ESP is designed to meet specific demands of learners with content to particular disciplines, activities, and occupations [5]. According to this definition, English used in the hydraulic industry should be defined as an ESP. Academic documents in the hydraulic industry cover various topics, such as hydrology, geology, hydraulic engineering, construction, budget estimation, project contracting, environmental assessment, resettlement, dam safety, hydraulic annals, and hydraulic academic studies, featuring high professionalism and low accessibility. Therefore, it is necessary to establish a hydraulic industrial corpus to meet practitioner and learner needs in hydraulic terminology acquiring, academic writing, professional material translation, and more.

3.2 Corpus Capacity Design

It is a stereotype that only corpora with a large capacity can extensively show the language characteristics and patterns of the industry. Bowker and Pearson refuted this opinion and believed that the value of a corpus cannot be evaluated by its capacity [6]. As long as it is delicately designed, a corpus with several thousand words can be as valuable as one with hundreds of thousands of words. The value of a corpus should be assessed by its openness, diversity, and relevance between the content of the corpus and the original intention of the establishment. Based on this philosophy, NHRICorpus is designed to collect the terms that appeared in the internal academic reports produced by researchers from six departments, including hydrology and water resources, hydraulic engineering, river and harbor, geotechnical engineering, materials and structural engineering, dam safety management, eco-environmental researches, and rural electrification, of Nanjing Hydraulic Research Institute, from 2015 to 2021. Terms involved in other subjects are not included. The capacity of the corpus is approximately 150,000 to 200,000 words.

3.3 Difficulties and Challenges

Due to the limitations caused by internal reports and our lack of professional support, there are two major difficulties that occurred in corpus construction. Firstly, the internal documents are not allowed to be uploaded online due to confidential issues, we spend a large amount of time converting those manuscripts to electronic documents via keyboarding or OCR recognition, and the proofreading process was also time-consuming. Secondly, due to the lack of professional background, we are not able to extensively determine the importance and relevance of terms that appeared in the documents, so the only standard of terms retrieval was word frequency. To address this challenge, after the term retrieval process was completed, I translated those terms into English and sent my translation to scientific researchers for proofreading. Also, we took 15 documents as a set of samples, and sent them to researchers, inviting them to manually select hydraulic terms based on their knowledge. Compared their selection to my term list, some terms were falsely deleted due to less word frequency.

4. Construction of the NHRICorpus

NHRICorpus is a platform that consists of a Chinese-English bilingual, parallel corpus, and a term search engine. The corpus content is retrieved from internal academic research reports produced by Nanjing Hydraulic Research Institute. The platform supports precision search, fuzzy search, and term recommendations. Users can use the search bar on the homepage to search for specific terms or use keywords to find relevant terms. Also, they can go through term recommendations and the dynamic hot term banner to randomly check any term they are interested in. Each term page includes term names, definitions, examples in both Chinese and English, term pictures, and information about the source papers of the examples, including the title, institute, published year, and abstract. The platform enables users to efficiently get access to hydraulic terms and how they are used in academic reports. Also, the platform pages are designed based on eye-tracking theories to improve user experience.

4.1 Tools and Technologies

ROSTCM6, a software researched, developed, and coded by Professor Shen Yang from Wuhan University, was the only large-scale, free content mining and computing platform in China. This software supports various document analyses, including Weibo analysis, chat history analysis, comprehensive network analysis, webpage analysis, browse analysis, text segmentation, word frequency counting (Chinese and English), traffic analysis, and clustering analysis.

Antconc is a well-designed and easy-to-use cluster (word sequence frequency patterns) or n-grams (n word sequences within a document or corpus) finding software created by Laurence Anthony of Waseda University [7]. The software features a freeware license, small memory requirements for only about 2MB of disk space, an extensive set of text analysis tools, powerful search features, multiple-level sorting, and Unicode support.

4.2 Corpus and Term Search Platform Construction

Corpus construction is the most important part of corpus pragmatics research design [8]. The Construction of NHRICorpus technically consists of five steps, including access to data, term retrieval, alignment, descriptive fields development, and search platform development. The corpus and term search platform construction procedures are shown in Figure 3.

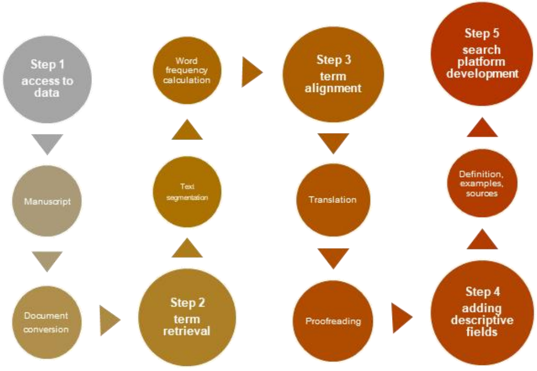


Fig. 3. Corpus and term search platform construction procedures

4.2.1 Access to data

Source papers used in this project are the internal academic research reports produced by researchers from Nanjing Hydraulic Research Institute. The printed papers can be accessed via NHRI internal library and the scanned electronic papers can be accessed via the database in the local network. Printed papers, scanned pictures, and ineditable not-editable pdf documents must be transformed into editable, electronic documents via keyboarding and offline document conversion software. A total of 127 papers are collected for document conversion. It is a time-consuming process and proofreading is required.

4.2.2 Term retrieval

This step consists of text segmentation and word frequency calculation. I applied ROSTCM6 software for text segmentation and Antconc software for word frequency calculation.

After being segmented, the texts were imported into Antconc. The View Files tool, KWIC tool, Word List tool, and Word Clusters tool are frequently used in corpus analysis and development. When a document is imported into Antconc, the regular expressions, phrases, words, or substrings can be searched by the View Files tool, which is often independently used as a powerful text search engine. When using the KWIC tool, users can define the highlight color of all resulting hits, and click buttons or use keyboard shortcuts to quickly locate a specific hit in a target file. When being asked to analyze a new file or corpus, a majority of users would choose to use the Word List tool to generate a list, which reflects the highlighted problem areas and investigation focus in a corpus. In this project, we sorted words into frequency order for term retrieval. Some words of high frequency cannot be collected into the term list since they are irrelevant to the main idea of the file. These words were eliminated during the data cleansing process. For instance, using the Word List tool, we found a word ranked first by frequency in the file. We can search the word by using the KWIC tool or the View File tool to locate it in the file. Redundant nouns, function words, adverbs of degree, prepositions, pronouns, and high-frequency words which make no contribution to the main idea of the file should be eliminated [9]. Users can generate a stop list or the reverse of a stop list, which can be specified by uploading a separate file or inputting from the keyboard, to avoid counting meaningless words. Also, the Word Clusters tool can help users acquire multi-word units and collocations, such as idioms and phrasal verbs, which are difficult for users to acquire. Although Antconc has limitations in handling annotated data, it is undoubtedly an easy-to-use, simple, lightweight corpus analysis tool that plays a significant role in corpus analysis and construction.

4.2.3 Term alignment

Since the terms are retrieved from internal academic research reports, no English version of these reports can be found. Therefore, we use hydraulics dictionaries or online references to translate a total of 30,000 words into English, and sent the translation to scientific researchers for confirmation.

4.2.4 Descriptive fields

One of the original intentions of building the NHRICorpus was to enable practitioners and students to get access to accurate terms as well as the latest research findings in the hydraulic industry. Therefore, in addition to the elements traditionally contained in corpora and term search platforms, such as term use examples, definitions, and part of speech, NHRICorpus innovatively provides users with information about source papers the examples retrieved from, such as the title, institute, and paper published year. The examples will be updated on a regular basis, so as to collect up-to-date research findings for users and support industry development trend analysis.

4.2.5 Term search engine development

Technologies applied in the term search engine are based on PHP programming language, MySQL database management system, and server environment that supports Linux, Nginx, PHP, and MySQL. The search engine is designed to realize separate management of the frontend web pages and backend content administration. The system function of the search platform is developed by different layers, so as to support second-time development, including function expansion and user interface update. The function module architecture is shown in Figure 4.

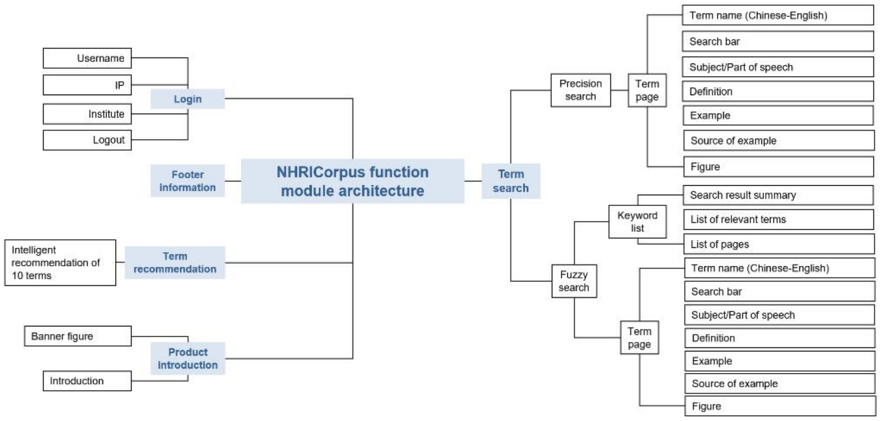


Fig. 4. NHRICorpus function module architecture

4.3 Web Page Design Based on Eye-Tracking Theories

4.3.1 Reasons for Applying Eye-Tracking Experiments

Eye-tracking technology was developed nearly a century ago, but only recently it has been applied in research regarding online dictionary/word search platform use [10]. The eye-tracking approach is used to measure users' eye motion (relative to their head) and point of gaze (where to look). Human eye motions, such as gaze behavior, are usually interpreted as a reflection of perception. Eye-tracking devices trace users' eye movements, so as to determine which features cause confusion, which features attract more attention, and which features are most easily to be neglected.

4.3.2 Experiment Procedures

During the experiment, participants received two tasks: 1) get an overview of the NHRICorpus homepage; 2) find out the headword being described on the term page.

The eye-tracking experiment included two parts corresponding to the two tasks mentioned above.

In the first part, participants were asked to 1) follow the instruction for taking a look at the NHRICorpus homepage; 2) explain their eye motions when looking at the homepage; and 3) participate in an after-experiment interview.

In the second part, participants were asked to 1) follow the instruction for browsing a term page prepared by the experiment organizer; 2) explain their eye motions when looking at the term page; and 3) participate in an after-experiment interview.

4.3.3 Participants

Five participants were invited to participate in this project, all of them aged 22–35. Due to the lack of professional facilities and labs, participants were required to tell their point of gaze (where to look) when looking at a web page, hold the point of gaze for two seconds before moving to the next feature, and participate in a brief interview after the experiment, and the whole process was recorded for further confirmation.

4.3.4 Results and Discussion

The results of task 1) are shown in Figure 5. The eye motion routes of all five participants are summarized here in the form of an eye motion route map.

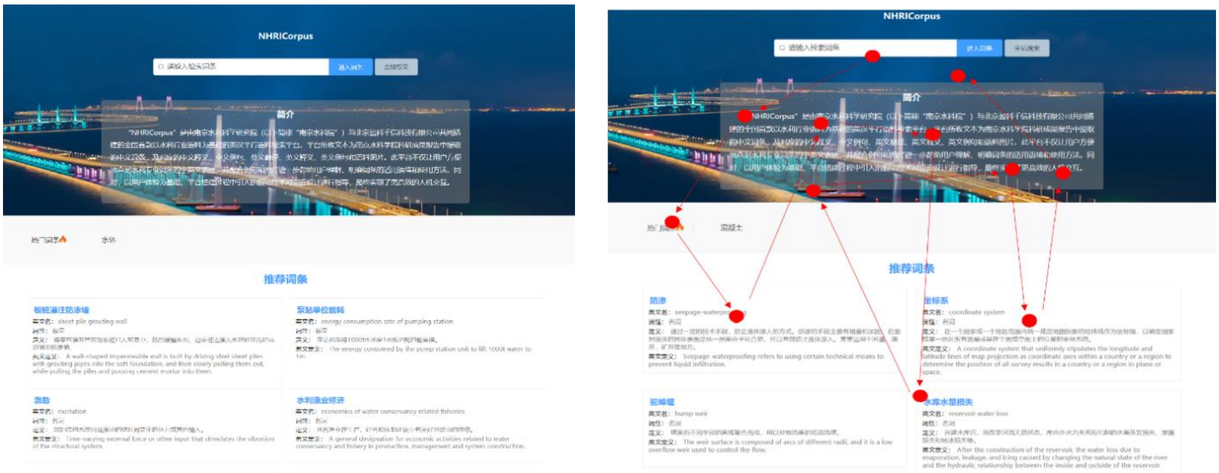


Fig. 5. Homepage design and participants' eye motion route map

It is obvious from the eye motion route map that, participants read the NHRICorpus introduction first, and their gaze point moved to the dynamic banner of "hot topics" and randomly went between the section of "term recommendation" and the introduction. All five participants scanned several times at the introduction and were able to retell the information mentioned in the introduction at the after-experiment interview.

The results of task 2) are shown in Figure 6. The original design of the term page and the eye motion routes of all participants are summarized here in the form of an eye motion route map.

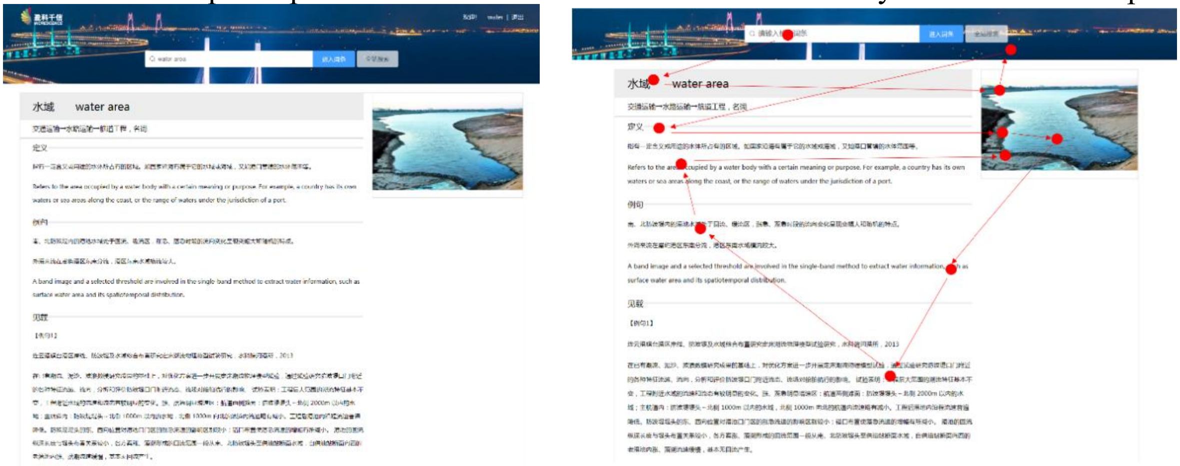
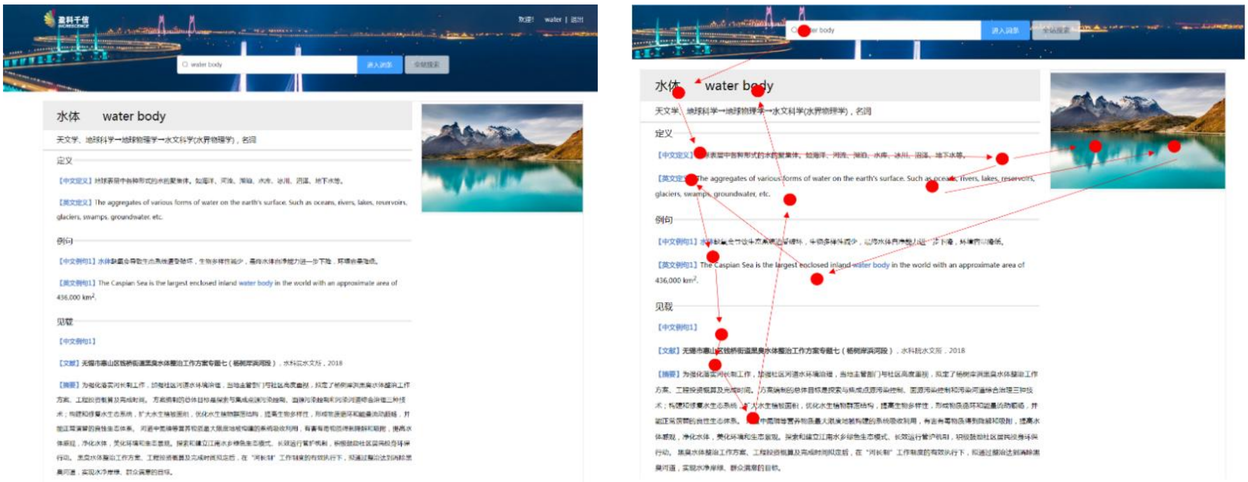


Fig. 6. Original term page design and participants' eye motion route map

It is easily noticed that, after checking the headword once on the page, participants' point of gaze stayed mostly at the picture and skimmed some of the texts on the term page. In the after-experiment interview, four participants mentioned that their attention was easily distracted by pictures on the page, and the text architecture was vague. They could not efficiently locate the information they need. Also, without informing the participants of the experiment's intention, we asked them whether the following sections were included on the webpage: 1) name of the term; 2) Chinese and English definitions; 3) Chinese and English examples; and 4) information about the papers where the examples retrieved from.

Only one participant selected all the right sections. The rest of them obviously had their views on the definition, example, and source sections, but failed to remember all sections on the page correctly. To solve this problem, we redesigned the term page into the architecture shown in Figure 7, and the results become different.



水体 water body

天文学、地球科学—地球物理学—水文科学(水界物理学)，名词

定义

【中文定义】地球表层中各种形式的水的聚集体。如海洋、河流、湖泊、水库、冰川、沼泽、地下水等。

【英文定义】The aggregates of various forms of water on the earth's surface. Such as oceans, rivers, lakes, reservoirs, glaciers, swamps, groundwater, etc.

例句

【中文例句1】水体缺氧会导致生态系统遭受破坏，生物多样性减少，导致水体自净能力进一步下降，环境容量降低。

【英文例句1】The Caspian Sea is the largest enclosed inland water body in the world with an approximate area of 436,000 km².

见载

【中文例句1】

【文献】无锡市惠山区钱桥街道黑臭水体整治工作方案专题七（[illegible]河段），水科院水文所，2018

【摘要】为强化落实河长制工作，加强社区河道水环境治理，当地主管部门与社区高度重视，拟定了[illegible]黑臭水体整治工作方案、工程投资概算及完成时间。方案编制的总体目标是探索与集成点源污染控制、面源污染控制和污染河道综合治理三种技术；构建和修复水生态系统，扩大水生植被面积，优化水生植物群落结构，提高生物多样性，形成物质循环和能量流动链路，并[illegible]

Fig. 7. Redesigned term page and participants' eye motion route map

In the redesigned term page, the information architecture was clarified by adding subtitles in front of the content in each section in a different color. Such changes helped participants clearly understand the structure of the page [11]. It is clear from the eye motion route map that participants were less distracted by the figure and went through each section easily. In the after-experiment interview, all participants mentioned that the redesigned term page was much clear and easy to use.

The NHRICorpus search platform was established to integrate and display terms and corresponding definitions, examples, and information of the source articles on the same page. Therefore, clear architecture and clear guidance are significant for creating an easy-to-use web page.

5. Summary

The rapid development in the hydraulic industry in China has attracted higher attention to hydraulic scientific research, academic studies, engineering projects, and higher education. By analyzing hydraulic subject development in 76 universities that are authorized to issue master's and doctoral degrees, there is a clear trend that China will put more effort into cultivating hydraulic talents in the future. Also, since 2008, industrial corpora–related research has entered into fast development, however, studies specialized in the hydraulic industry can hardly be found. Therefore, building a hydraulic corpus is urgently needed.

NHRICorpus is designed to not only provide language support for practitioners and students in knowledge acquiring, academic writing, and professional material translation, but also support industry development trend analysis and research hotspot prediction. NHRICorpus consists of a corpus with a capacity of 150,000-200,000 words and an easy-to-use term search engine. The examples in the corpus are planned to be updated on a yearly basis, so as to collect the latest research findings for more accurate industry development trend analysis and prediction.

For the existing deficiencies, the corpus can be further improved by firstly, importing more online, bilingual hydraulic documents for expanded openness, better coverage, and larger capacity, and secondly, inviting more scientific researchers for professional support. In this way, the NHRICorpus can be developed into a comprehensive hydraulic corpus featuring both online and local documents, higher professionalism, and user-friendly page design, so as to effectively support industrial development, empower hydraulic education, and meet practitioners' and students' language requirements.

References

[1] Yvonne A. B. Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges. Peter Lang, 2011

[2] Eggar C. Global and China Hydraulic Industry Report, 2021-2026, Giving Intelligence Teams an AI-powered advantage. ReportLinker, 2021

[3] Djamasbi, S. Hall-Phillips, A. "Visual Search", Eye Tracking in User Experience Design. A. Schall, J.Romano Bergstrom (eds.). Burlington: Morgen Kaufman, 2014, pp. 27-43

[4] John F. "Corpora in Language Teaching", The Handbook of Language Teaching. M. H. Long, C. J. Doughty (eds). Oxford: Blackwell Publishing Ltd., 2009

[5] Momtazur R. English for Specific Purposes (ESP): A Holistic Review. Universal Journal of Educational Research, 2015,3(1): 24-31

[6] Lynne B., Jennifer P. Working with Specialized Language: A Practical Guide to Using Corpora. London/New York: Routledge, 2022

[7] Laurence A. AntConc: Design and Development of a Freeware Corpus Analysis. IEEE International Professional Communication Conference Proceedings, 2005: 729-737

[8] Gisle A. "Corpus Construction", Methods in Pragmatics. A. H. Jucker, K. P. Schneider, W. Bublitz (eds). Berlin/Boston: De Gruyter Mouton, 2018: 467-494

[9] Zhen-gang Z., Yu-yuan W. Research on Text Analysis and Quantitative Evaluation of Mass Innovation Space Support Policies in Jiangsu Province. Advances in Economics, Business and Management Research, 2019(87):131-142

[10] Carolin M, Frank M., Alexander K. "Evaluation of a New Web Design for the Dictionary Portal OWID: An attempt at using eye-tracking technology", Using Online Dictionary. Berlin/Boston: De Gruyter Mouton, 2014

[11] Agnieszka B. Using Eye Tracking to Compare Web Page Designs: A Case Study. Journal of User Experience, 2006(3),1:112-120