Design Principles and Functionality of Chinese Interlanguage Corpora: A Case Study of the HSK Dynamic Composition Corpus 2.0

Baolin Zhang

Abstract Since the beginning of the twenty-first century, great progress has been made in terms of the construction of the Chinese Interlanguage Corpora and the essential role that these corpora has played in the study of Chinese second language teaching and research. However, there still exist some technical issues in terms of corpora design and functionality, such as the simplicity of the search function, difficulty searching for a certain interlingual phenomenon, and inconvenience caused by its not-so-user-friendly interface design. Furthermore, especially in current times, network safety has become an increasingly prominent issue and has resulted in a lack of operational corpora that satisfy the needs of the academic community. Under these circumstances, and in order to aim for Generation 2.0 which requires more delicacy and abundance, it has become necessary to adjust the design concepts and motivations behind the corpora. To ensure the uninterrupted operation and accessibility of the corpora, attentiveness and improvements in terms of system security are crucial. Meanwhile, optimizations and improvements of corpora features, especially the search function, are also essential for comprehensively meeting the needs of all users.

Keywords Chinese interlanguage corpus · Design concepts · System security · Search function

This research was funded by the Beijing Advanced Innovation Center for Language Resources (Code: KYD17004), and the Ministry of Education of the People's Republic of China's Key Philosophy and Social Science Research Project (Code: 12JZD018) and Key Beijing Social Science Foundation Project (Code: 15WYA017).

B. Zhang (✉)
Research Institute of International Education of Chinese Language, Beijing Language and Culture University, Beijing, China
e-mail: zhangbl@blcu.edu.cn; baolin08@126.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

H. H.-J. Chen et al. (eds.), *Learner Corpora: Construction and Explorations in Chinese and Related Languages*, Chinese Language Learning Sciences,
https://doi.org/10.1007/978-981-19-5731-4_2

1 Introduction

1.1 The Development of Corpus Construction and Its Applied Research

Considered as the first Chinese interlanguage corpus in academia, the "Chinese Interlanguage Corpus System" was developed at the Beijing Language Institute in 1995. Zhang (2019: 86) notes that "despite there still being some problems with the corpus, such as the small scale of the corpus size, the limited breadth and depth in processing, and the lack of corpus retrieving speed (see Chen, 1996), the corpus system still holds a good reputation and practical value as a pioneering sharer in academia".

Several corpora have been built during the first decade of the twenty-first century, among which the most influential are the HSK Dynamic Composition Corpus (Beijing Language and Culture University), the Chinese Interlanguage Corpus for International Students (Jinan University College of Chinese Language and Culture, including written and spoken corpora), the Interlanguage Corpus for International Students (Sun Yat-sen University), and the Corpus of Chinese Interlanguage Error Analysis for Foreign Students (Nanjing Normal University).

During the second decade of the twenty-first century, more corpora were built as more Chinese teachers, experts, and scholars have devoted themselves to corpus construction. Some of these include the Chinese Interlanguage Corpus for Korean International Students (Ludong University), the Chinese Written Language Corpus for International Students (Beijing Chinese Language and Culture College), the Chinese Acquisition Corpus for Foreigners (Shanghai Jiaotong University), the Language Acquisition Corpus for Spoken Chinese (LAC/SC, The Chinese University of Hong Kong), the small-scale Foreign Student Oral Interlanguage Corpus (Suzhou University), the Corpus Based on Oral Telephone Examinations (Peking University), the Errors in Continuity of Chinese Characters Interlanguage Corpus(Sun Yat-sen University), the Global Chinese Interlanguage Corpus (led by BLCU and co-constructed by academia), the TOCFL Learner Corpus (Taiwan Normal University), and the Guangwai-Lancaster Chinese Learner Corpus (CLC, Guangdong University of Foreign Studies and Lancaster University, UK).

The construction and development of the Chinese Interlanguage Corpus have promoted the emergence of corpus-based research and achieved numerous important research results. Representative publications include Zhao et al. (2008), Zhang et al. (2008), Xiao et al. (2009), and Zhang et al. (2014). Taking the HSK Dynamic Composition Corpus as an example, 3858 research papers of various types involving this corpus were found in the China Knowledge Network (CNKI) database as of May 26, 2019 (Fig. 1).

These research papers mainly consisted of two categories: Master's thesis with a total of 2,929 papers and journal publications with a total of 732 papers (Fig. 2).

More importantly, the rapid development of corpus construction and corpus-based applied research has propelled a shift for Chinese interlanguage and Chinese

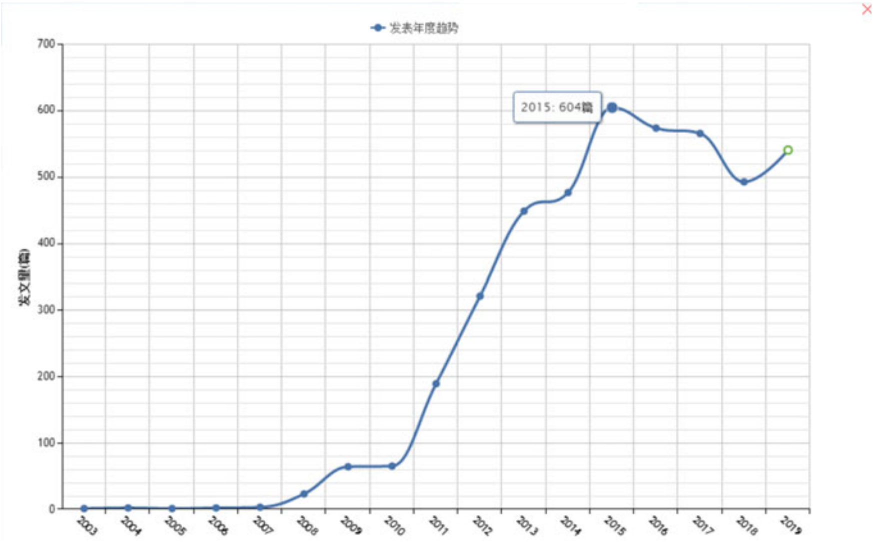


Fig. 1 Annual number of research papers based on HSK

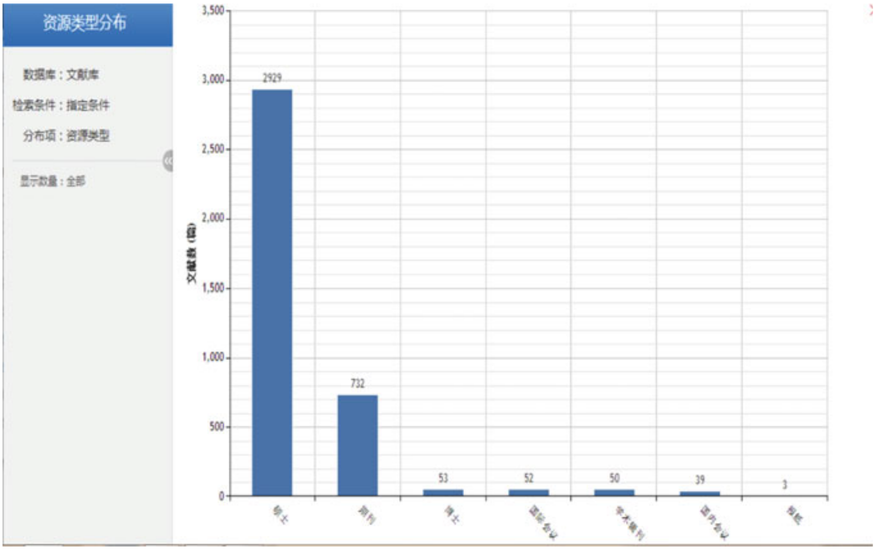


Fig. 2 Distribution of research resource types based on HSK

second language acquisition research. The field has seen a move away from small-scale, empirical, and speculative research toward large-scale, real data-based research that combines both quantitative and qualitative analyses. This shift has resulted in researchers being able to make conclusions with greater objectivity, universality, and stability.

1.2 *Ontological Study of Chinese Interlanguage Corpus Construction*

The so-called ontology research refers to the research on the related theoretical issues of corpus construction. The reason for this naming is for the tendency of emphasizing the practice of corpus construction and despising theoretical discussion in the construction of corpus and emphasizing that theoretical research is an integral part of the construction of the corpus. Mainly it includes the following.

The overall design of a corpus occurs after the specific objectives of the corpus have been clarified and the necessity for its construction has been addressed. Addressing the feasibility of constructing the corpus involves (1) researching how to construct the corpus so that it meets the desired objectives; (2) clarifying its features; (3) determining its scale, materials, and structure; and (4) deciding annotation methods, principles of construction, and methods of application (Zhang & Xiliang, 2015). Representative research papers on overall corpus design include Chu and Xiaohe (1993) "The Basic Idea of Establishing a 'Chinese Interlanguage Corpus System'" and Cui and Zhang's (2011) "The Construction Plan of 'Global Chinese Learner Corpus'".

In addition to the overall design, labeling conventions are also an important component of corpus construction. The specification of corpus annotation schemes makes it easier to centralize what type of content is annotated, i.e. labeled, and how. The existence of a centralized annotation system is very important, especially in the case of corpus-based research paradigms. However, what to label can vary in corpus-building practices depending on the professionalism, knowledge, and hands-on experience of the project leaders, resulting in different functions and different use values of the corpus. Despite the importance of labeling conventions, not much research has been done in this area with the exception of several articles about the comprehensiveness of labeling (see Xiao et al., 2014; Zhang and Xiliang, 2018). Therefore, more extensive academic discussions or debates are needed to clarify the reasoning behind, and unify the understanding of, labeling conventions so as to form a labeling scheme that can be generally accepted by academia, thus promoting further development of corpus construction. Zhang et al. (2019)'s publication "Study on Standardization of Chinese Interlanguage Corpus Annotation" is one of the important related research projects in the field in recent years.

Construction standards are another aspect of corpus construction that requires attention. Research in this area is a response to the lack of unified standards in the construction of Chinese interlanguage corpora and the great arbitrariness in the practice of database construction. Construction standards are a summary of the experience in Chinese interlanguage corpus construction. These standards draw on various lessons, consolidate academic theories of corpus construction, designate the levels of corpus construction, and have important guiding significance for corpus construction (Zhang & Xiliang, 2015). Not much research has been carried out in this area yet, and Zhang and Xiliang's (2015) "On Construction Standards of Chinese Interlanguage Corpus" is a relatively comprehensive and systematic discussion of this issue.

A final noteworthy component of corpus construction is that of software systems. In the past, the construction of Chinese interlanguage corpora focused more on the size, composition, and labeling conventions and lacked attention with regard to the development of management and retrieval/search systems. In fact, the development of software systems, including management and retrieval systems, plays a very important role in enhancing the practical functions of a corpus and improving the level of corpus construction.

1.3 Existing Problems

The development and progress of the Chinese interlanguage corpus are undoubtedly huge, and it has been widely recognized by the academic circles. However, the problem is also obvious, which not only determines the construction level of the corpus, but also affects the application research based on the corpus. Mainly it is manifested in the following aspects:

(1) Annotated content is not comprehensive and cannot meet the needs of teaching and research in many aspects. For example, the "HSK Dynamic Composition Corpus" only has entries from intermediate and advanced learners and can only be used for static horizontal research, not for vertical research of the acquisition process. In addition, only character, word, sentence, and writing composition errors are annotated, making the corpus suitable for error analysis but not performance analysis.

(2) The search and retrieval function is too simple, resulting in the inability to search for some important language phenomena and limitations with regard to the functions of the corpus. For example, it is not possible to search for sentences using the "是……的" *Shi……de* construction, the "连……也/都……" *Lian……ye/dou……* construction, semi-fixed collocations, nor sentences where "离" *Li* is used as a separable word.

(3) The functional design is not user-friendly and is inconvenient. For example, you cannot automatically download the corpus results generated by a specific query; wrong recordings, mislabeling, or omissions found while browsing the corpus cannot be fixed or amended by users; thoughts, comments, and suggestions from users cannot be relayed to the creator.

(4) The network security is not up to industry standards and cannot be accessed outside specific school networks, which seriously affects the use of the corpus. This is largely attributed to the fact that the corpus was developed ahead of its time, and the programming language and technology used in its original construction have now become outdated. This has resulted in security loopholes in the system and the failure of the corpus to meet the open requirements. Consequently, the corpus could not continue to be accessed by domestic and foreign users, diminishing its application in Chinese language teaching and related research.

Faced with the four major issues outlined above, we have made various adjustments to the corpus to improve its functionality and applicability for Chinese language teaching and related research. We regularly patched the system which resolved some of the issues preventing the corpus from being available on the campus network and to those on campus. However, the patched system does not meet the needs of domestic and foreign academia. The entire HSK corpus was copied to the BCC corpus so that everyone could at least browse the HSK corpus. However, BCC is a native language corpus with a different retrieval method from the HSK corpus. It is still difficult to meet the needs of academia due to the inconvenience and inability to search in a mislabeled corpus.

1.4 Solutions

Faced with Academia's urgent need for corpora despite their shrinking number, we decided to redevelop the HSK corpus' software system by using the current mainstream programming language. We did this in order to continue and better serve Chinese teachers, scholars, researchers, graduate students, and Chinese learners domestically and globally.

The task of developing the new system was approved and commenced on January 5, 2018. On February 11 of the same year, the new system was complete and deployed to the server. On March 28, the system was officially opened to the public following trial operations and debugging. The new version of the system is known as the "HSK Dynamic Composition Corpus (Version 2.0)", and can be accessed at hsk.blcu.edu.cn. The new system ranks high in safety performance, which has allowed it to remain open to the public. As a result, we have achieved the goal of rebuilding the corpus system and the HSK corpus can continue to serve Chinese language teaching and research around the world.

2 Design Principles

In view of the various problems in the construction of the Chinese interlanguage corpus, and in accordance with the purpose of building the corpus, we have formulated the basic principles for the redevelopment of the corpus system.

2.1 Aim

Our aim in the construction of Chinese interlanguage corpora has always been to serve the teaching and research of Chinese as a foreign language. At the Third International Symposium on Construction and Application of Chinese Interlanguage Corpus held in the Summer of 2014, our focus became to proactively and wholeheartedly serve Chinese language teaching and research all over the world. It is precisely under the guidance of this understanding and aims that the HSK corpus, whether it be version 1.0 which was completed and launched at the end of 2006, version 1.1 which was upgraded in August 2008, or version 2.0 which has been recently re-developed, was made available free of charge and without delay to users all over the world.

2.2 Principles

There are three core principles with regard to the re-designing of the HSK corpus: (1) ensuring reliable and secure operation; (2) ensuring that the functions meet user demands; and (3) ensuring a fast, simple, and user-friendly experience.

The reconstruction of the corpus software system was mostly due to network security issues. Therefore, the first requirement for the reconstruction of the corpus is that there should be no or minimum security risks. It must be ensured that the corpus can operate normally and continue to serve the academic community uninterruptedly. Specifically, first of all, the new corpus system must not have any high-risk and medium-risk vulnerabilities, and low-risk vulnerabilities should be kept to a minimum as much as possible so that it can successfully pass the security inspections implemented by relevant departments and units. Secondly, when there are high-risk and medium-risk vulnerabilities, it can respond quickly and solve the problem in time, so as to ensure that the corpus is normally opened and not closed. This is a new problem brought about by the rapid development of information technology in the Internet era, and corpus builders must pay close attention to this problem.

The second consideration that came into play when redesigning the HSK corpus was the need to ensure that the functions of the corpus were computationally powerful enough to meet user demands. The HSK corpus version 1.0 and version 1.1 are products of the 1.0 generation, which were built during the initial period of Chinese interlanguage corpora construction (cf. Zhang, 2019). These two versions embody defining characteristics of that time in that they are simple and large scale, not fully functional, and have difficulty satisfying users' demands in many aspects. For example, in these earlier versions, it is possible to search for the usage of the separable word "合" *He*, but not for the usage of "离" *Li*. One can search for some sentence constructions with marker words such as the "把" *Ba* construction and the "比" *Bi* construction, but not for sentences with the "是……的" *Shi……De* construction or the "连……也/都……" *Lian……Ye/Dou……* construction, because these sentence constructions require two search terms.

These deficiencies in functionality may lead to incomplete research conclusions since the generated results, limited by the corpus' lack of functionality, do not comprehensively or accurately reflect second language speaker use. That is, in cases where the corpus does not have the ability to process the queried entry, as is the case with separable words like "Li", the phenomenon cannot be fully analyzed. Furthermore, the value of the corpus cannot be realized for relative corpus research under such conditions because certain sentence patterns cannot be retrieved. The new corpus system solves these problems and facilitates users to explore various language phenomena so as to better serve Chinese language teaching and research.

The third aspect of attention was to ensure the new design was a fast, simple, and user-friendly interface. The imperfect design and inconvenience of use with regard to version 1.0 and version 1.1 of the HSK corpus were also sources of problems. For example, the queried results could be downloaded automatically which resulted in negative user feedback. Users noted that the huge amount of queried results could only be downloaded manually page by page, resulting in sore wrists. Furthermore, users could not adjust the quantity of output results when browsing; they could not communicate with the administrator and give feedback in a timely manner when they encountered problems; they could not make corrections for errors they found in the corpus entry and annotation, so the errors continued to exist and cause problems for other users. The new system also solves these problems and is more user-friendly, allowing users to use the corpus more conveniently and to correct any errors they may find.

3 Functional Design

In order to solve the various existing problems in the construction of the corpus and make up for its shortcomings, the functions of the corpus should be improved in terms of corpus retrieval, presentation, data statistics, maintenance, message feedback, automatic download, etc., so that it can better serve Chinese teaching and research.

3.1 Search

The basic way for users to use the corpus is corpus retrieval. From a user's point of view, the value of the corpus lies in the retrieval, presentation, and accessibility of the corpus. What they care about is whether the search function can retrieve the results that they need, and whether it can provide the convenience of collecting and retrieving data for their own teaching and research work.

The search function of a corpus should include the retrieval of specific characters, words, phrases, and sentences; the retrieval of annotated content; the retrieval of special sentence constructions, fixed and semi-fixed structures, compound sentences, and the usage of separated words such as "Li"; collocation searches; and the ability to retrieval data based on parts of speech.[1] The search parameters of a corpus should be constructed based on the characteristics of a language user's nationality, gender, age, composition topics or oral topics, and scores; the search function should help users gain access to the error corpus, the correct corpus, and all the entirety of the corpus.

The search function should be simple, convenient, and easy to use.

3.1.1 General Search of Strings

This is the basic retrieval function of the corpus which allows one to search for specific characters, words, phrases, and sentences in the corpus. Generally speaking, this function is available in any corpus. As far as the HSK corpus is concerned, search parameters can be set based on factors such as the candidate's nationality, composition topic, certificate level, test time, and test score.

It should be noted that there are two "composition scores" in the retrieval conditions, which can indicate the selection range of the two scores before and after. For example, the first score is set at 60 and the latter one is set at 80, which means the retrieved corpus results originate from compositions from a score range of 60 to 80.

Below are examples of entries for specific characters, words, phrases, and sentences.

Take "帮" *bāng* as an example for word query (Fig. 3).

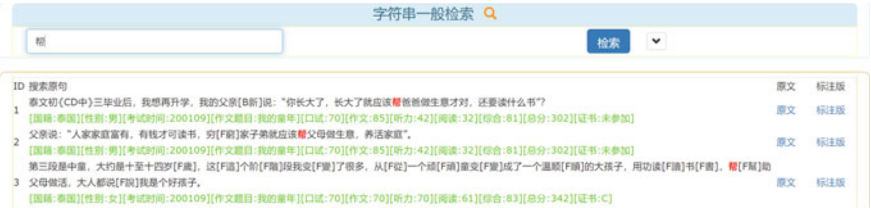
The word query takes "帮助" *bāngzhù* "help" as an example (Fig. 4).

Phrase query taking "帮助别人" *bāngzhù biérén* "helping others" as an example (Fig. 5).

Sentence query taking "我们应该帮助别人" *wǒmen yīnggāi bāngzhù biéren* "we should help others" as an example (Fig. 6).

The usage query of the separable word "离" takes "帮忙" and "见面" as examples. A space must be added between the two components of the separable word (e,g, 帮[space]忙, Fig. 7; 见[space]面, Fig. 8) in order to generate relevant corpus results.

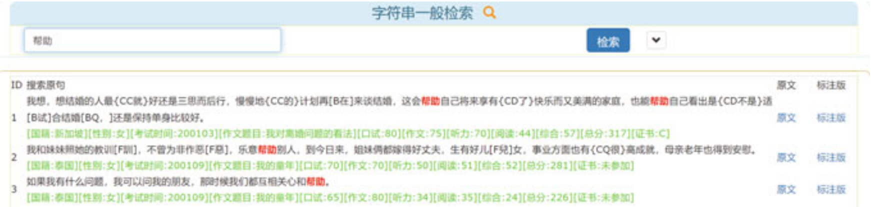
[1] The role of part-of-speech retrieval is of great significance to the construction of the corpus and the use of the corpus by users. However, it is a pity that the HSK corpus does not realize this function. The "Global Chinese Interlanguage Corpus" fulfills this function.



帮

ID	搜索原句	原文	标注版
1	泰文初{CD中}三毕业后，我想再升学，我的父亲[B新]说："你长大了，长大了就应该帮爸爸做生意才对，还要读什么书"？ [国籍:泰国][性别:男][考试时间:200109][作文题目:我的童年][口试:70][作文:85][听力:42][阅读:32][综合:81][总分:302][证书:未参加]	原文	标注版
2	父亲说："人家家庭富有，有钱才可读书，穷[F窮]家子弟就应该帮父母做生意，养活家庭"。 [国籍:泰国][性别:男][考试时间:200109][作文题目:我的童年][口试:70][作文:85][听力:42][阅读:32][综合:81][总分:302][证书:未参加]	原文	标注版
3	第三段是中童，大约是十至十四岁[F歲]，这[F這]个阶[F階]段我变[F變]了很多，从[F從]一个顽[F頑]童变[F變]成了一个温顺[F順]的大孩子，用功读[F讀]书[F書]，帮[F幫]助父母做活，大人都说[F說]我是个好孩子。 [国籍:泰国][性别:女][考试时间:200109][作文题目:我的童年][口试:70][作文:70][听力:70][阅读:61][综合:83][总分:342][证书:C]	原文	标注版

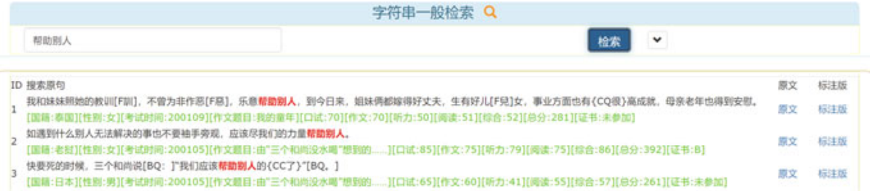
Fig. 3 Results produced searching for the character "帮" *bāng*



帮助

ID	搜索原句	原文	标注版
1	我想，想结婚的人最{CC就}好还是三思而后行，慢慢地{CC的}计划再[B在]来谈结婚，这会帮助自己将来享有{CD了}快乐而又美满的家庭，也能帮助自己看出是{CD不是}适[B试]合结婚[BQ，]还是保持单身比较好。 [国籍:新加坡][性别:女][考试时间:200103][作文题目:我对离婚问题的看法][口试:80][作文:75][听力:70][阅读:44][综合:57][总分:317][证书:C]	原文	标注版
2	我和妹妹照她的教训[F訓]，不曾为非作恶[F惡]，乐意帮助别人，到今日来，姐妹俩都嫁得好丈夫，生有好儿[F兒]女，事业方面也有{CQ很}高成就，母亲老年也得到安慰。 [国籍:泰国][性别:女][考试时间:200109][作文题目:我的童年][口试:70][作文:70][听力:50][阅读:51][综合:52][总分:281][证书:未参加]	原文	标注版
3	如果我有什么问题，我可以问我的朋友，那时候我们都互相关心和帮助。 [国籍:泰国][性别:女][考试时间:200109][作文题目:我的童年][口试:65][作文:80][听力:34][阅读:35][综合:24][总分:226][证书:未参加]	原文	标注版

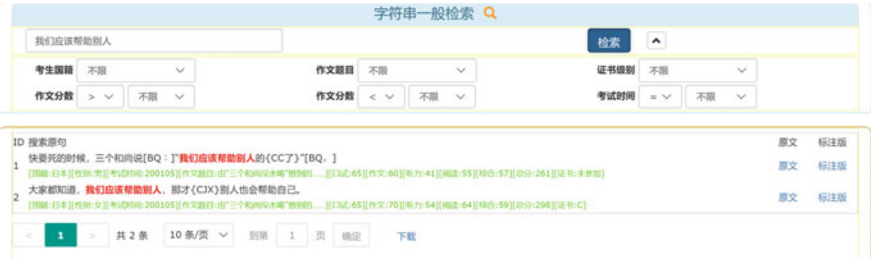
Fig. 4 Results produced searching for the character "帮助" *bāngzhù* "help"



帮助别人

ID	搜索原句	原文	标注版
1	我和妹妹照她的教训[F訓]，不曾为非作恶[F惡]，乐意帮助别人，到今日来，姐妹俩都嫁得好丈夫，生有好儿[F兒]女，事业方面也有{CQ很}高成就，母亲老年也得到安慰。 [国籍:泰国][性别:女][考试时间:200109][作文题目:我的童年][口试:70][作文:70][听力:50][阅读:51][综合:52][总分:281][证书:未参加]	原文	标注版
2	如遇到什么别人无法解决的事也不要袖手旁观，应该尽我们的力量帮助别人。 [国籍:老挝][性别:女][考试时间:200105][作文题目:由"三个和尚没水喝"想到的……][口试:85][作文:75][听力:79][阅读:75][综合:86][总分:392][证书:B]	原文	标注版
3	快要死的时候，三个和尚说[BQ：]"我们应该帮助别人的{CC了}"[BQ，] [国籍:日本][性别:男][考试时间:200105][作文题目:由"三个和尚没水喝"想到的……][口试:65][作文:60][听力:41][阅读:55][综合:57][总分:261][证书:未参加]	原文	标注版

Fig. 5 Results produced searching for the phrase "帮助别人" *bāngzhù bié ren* "help others"



我们应该帮助别人

考生国籍 不限　作文题目 不限　证书级别 不限
作文分数 > 不限　作文分数 < 不限　考试时间 = 不限

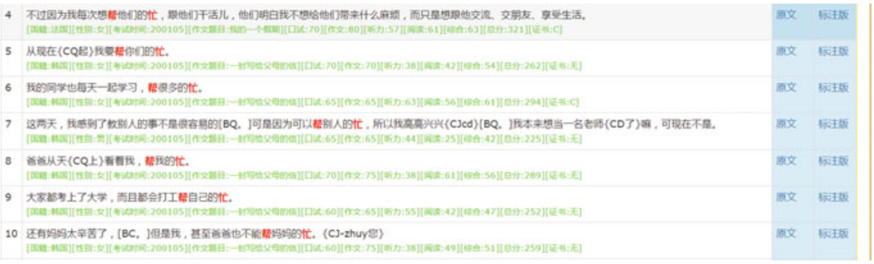
ID	搜索原句	原文	标注版
1	快要死的时候，三个和尚说[BQ：]"我们应该帮助别人的{CC了}"[BQ，] [国籍:日本][性别:男][考试时间:200105][作文题目:由"三个和尚没水喝"想到的……][口试:65][作文:60][听力:41][阅读:55][综合:57][总分:261][证书:未参加]	原文	标注版
2	大家都知道，我们应该帮助别人，那才{CJX}别人也会帮助自己。 [国籍:日本][性别:女][考试时间:200105][作文题目:由"三个和尚没水喝"想到的……][口试:65][作文:70][听力:54][阅读:64][综合:59][总分:298][证书:C]	原文	标注版

共 2 条　10 条/页　到第 1 页

Fig. 6 Results produced searching for the sentence "我们应该帮助别人" *wǒmen yīnggāi bāngzhù biéren* "we should help others"

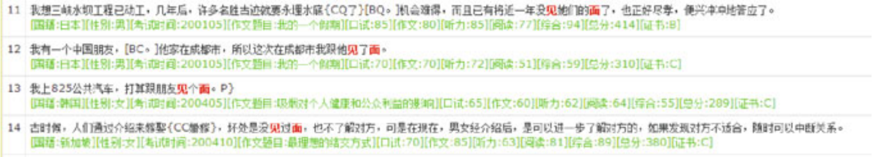
3.1.2 Sentence and Text Search

The HSK corpus offers an exhaustive collection of composition errors made by foreigners during the composition section of the Advanced Chinese Proficiency Test



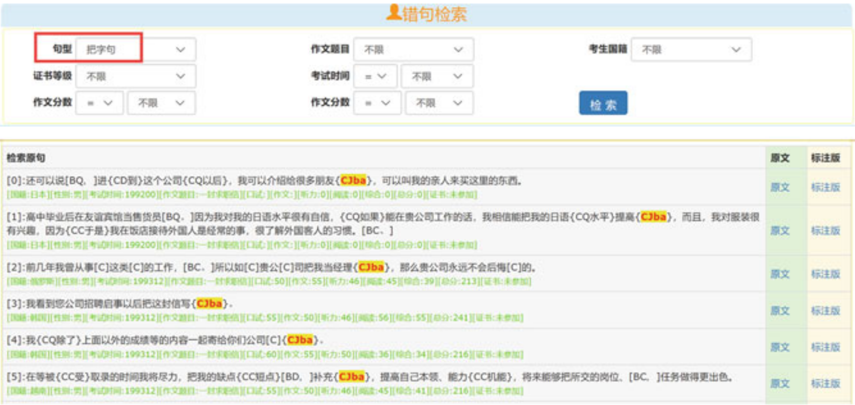
4	不过因为我每次想帮他们的忙，跟他们干活儿，他们明白我不想给他们带来什么麻烦，而只是想跟他交流、交朋友、享受生活。 [国籍:法国][性别:女][考试时间:200105][作文题目:我的一个假期][口试:70][作文:80][听力:57][阅读:61][综合:63][总分:321][证书:C]	原文	标注版
5	从现在{CQ起}我要帮你们的忙。 [国籍:韩国][性别:女][考试时间:200105][作文题目:一封写给父母的信][口试:70][作文:70][听力:38][阅读:42][综合:54][总分:262][证书:无]	原文	标注版
6	我的同学也每天一起学习，帮很多的忙。 [国籍:韩国][性别:女][考试时间:200105][作文题目:一封写给父母的信][口试:65][作文:65][听力:63][阅读:56][综合:61][总分:294][证书:C]	原文	标注版
7	这两天，我感到了教别人的事不是很容易的[BQ。]可是因为可以帮别人的忙，所以我高高兴兴{CJcd}[BQ。]我本来想当一名老师{CD了}嘛，可现在不是。 [国籍:韩国][性别:男][考试时间:200105][作文题目:一封写给父母的信][口试:65][作文:65][听力:44][阅读:25][综合:42][总分:225][证书:无]	原文	标注版
8	爸爸从天{CQ上}看看我，帮我的忙。 [国籍:韩国][性别:女][考试时间:200105][作文题目:一封写给父母的信][口试:70][作文:75][听力:38][阅读:61][综合:56][总分:289][证书:无]	原文	标注版
9	大家都考上了大学，而且都会打工帮自己的忙。 [国籍:韩国][性别:女][考试时间:200105][作文题目:一封写给父母的信][口试:60][作文:65][听力:55][阅读:42][综合:47][总分:252][证书:无]	原文	标注版
10	还有妈妈太辛苦了，[BC。]但是我，甚至爸爸也不能帮妈妈的忙。{CJ-zhuy您} [国籍:韩国][性别:女][考试时间:200105][作文题目:一封写给父母的信][口试:60][作文:75][听力:38][阅读:49][综合:51][总分:259][证书:无]	原文	标注版

Fig. 7 Results produced searching for the separable words “帮忙”



11	我想三峡水坝工程已动工，几年后，许多名胜古迹就要永埋水底{CQ了}[BQ。]机会难得，而且已有将近一年没见她们的面了，也正好尽孝，便兴冲冲地答应了。 [国籍:日本][性别:男][考试时间:200105][作文题目:我的一个假期][口试:85][作文:80][听力:85][阅读:77][综合:94][总分:414][证书:B]
12	我有一个中国朋友，[BC。]他家在成都市，所以这次在成都市我跟他见了面。 [国籍:日本][性别:男][考试时间:200105][作文题目:我的一个假期][口试:70][作文:70][听力:72][阅读:51][综合:59][总分:310][证书:C]
13	我上825公共汽车，打算跟朋友见个面。P} [国籍:韩国][性别:女][考试时间:200405][作文题目:吸烟对个人健康和公众利益的影响][口试:65][作文:60][听力:62][阅读:64][综合:55][总分:289][证书:C]
14	古时候，人们通过介绍来嫁娶{CC婚嫁}，坏处是没见过面，也不了解对方，可是在现在，男女经介绍后，是可以进一步了解对方的，如果发现对方不适合，随时可以中断关系。 [国籍:新加坡][性别:女][考试时间:200410][作文题目:最理想的结交方式][口试:70][作文:85][听力:63][阅读:81][综合:89][总分:380][证书:C]

Fig. 8 Results produced searching for the separable words “见面”



错句检索

句型: 把字句
作文题目: 不限
考生国籍: 不限
证书等级: 不限
考试时间: = 不限
作文分数: = 不限
作文分数: = 不限

检索原句	原文	标注版
[0]:还可以说[BQ。]进{CD到}这个公司{CQ以后}，我可以介绍给很多朋友{CJba}，可以叫我的亲人来买这里的东西。 [国籍:日本][性别:男][考试时间:199200][作文题目:一封求职信][口试:][作文:][听力:0][阅读:0][综合:0][总分:0][证书:未参加]	原文	标注版
[1]:高中毕业后在友谊宾馆当售货员[BQ。]因为我对我的日语水平很有自信，{CQ如果}能在贵公司工作的话，我相信能把我的日语{CQ水平}提高{CJba}，而且，我对服装很有兴趣，因为{CC于是}我在饭店接待外国人是经常的事，很了解外国客人的习惯。[BC、] [国籍:日本][性别:男][考试时间:199200][作文题目:一封求职信][口试:][作文:][听力:0][阅读:0][综合:0][总分:0][证书:未参加]	原文	标注版
[2]:前几年我曾从事[C]这类[C]的工作，[BC。]所以如[C]贵公[C]司把我当经理{CJba}，那么贵公司永远不会后悔[C]的。 [国籍:俄罗斯][性别:男][考试时间:199312][作文题目:一封求职信][口试:50][作文:55][听力:46][阅读:45][综合:39][总分:213][证书:未参加]	原文	标注版
[3]:我看到您公司招聘启事以后把这封信写{CJba}。 [国籍:韩国][性别:男][考试时间:199312][作文题目:一封求职信][口试:55][作文:50][听力:46][阅读:56][综合:55][总分:241][证书:未参加]	原文	标注版
[4]:我{CQ除了}上面以外的成绩等的内容一起寄给你们公司[C]{CJba}。 [国籍:韩国][性别:男][考试时间:199312][作文题目:一封求职信][口试:60][作文:55][听力:50][阅读:36][综合:34][总分:216][证书:未参加]	原文	标注版
[5]:在等被{CC受}取录的时间我将尽力，把我的缺点{CC短点}[BD、]补充{CJba}，提高自己本领、能力{CC机能}，将来能够把所交的岗位、[BC、]任务做得更出色。 [国籍:越南][性别:男][考试时间:199312][作文题目:一封求职信][口试:55][作文:50][听力:46][阅读:45][综合:41][总分:216][证书:未参加]	原文	标注版

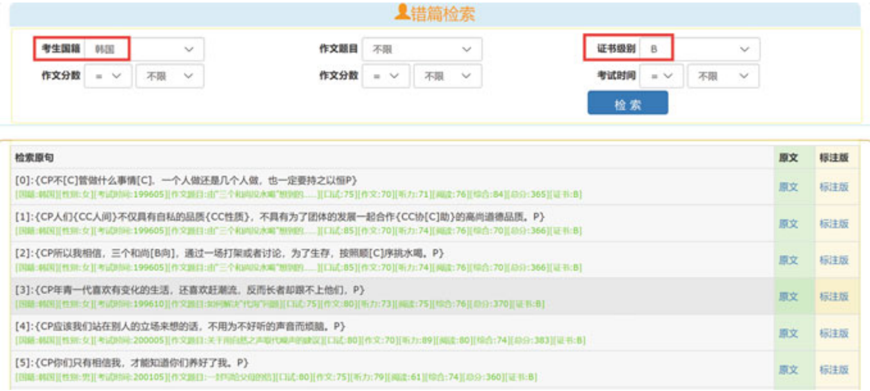
Fig. 9 Results produced searching for the “把” Ba sentence

with a focus on five areas: characters, words, sentences, text, and punctuation marks. Among them, errors in characters, words, and punctuations can be retrieved in either character strings or word and vocabulary lists. Errors in sentences and text can be queried using the sentence and text retrieval function.

Errors sentence retrieval takes “把” Ba sentence as an example (Fig. 9).

See the figure below for the error text (Fig. 10).

The above two search methods are available in the 1.0 and 1.1 versions of the corpus. These methods can solve retrieval problems such as when searching in an error corpus.



检索原句	原文	标注版
[0]:{CP不[C]管做什么事情[C]。一个人做还是几个人做，也一定要持之以恒P} [国籍:韩国][性别:女][考试时间:199605][作文题目:由"三个和尚没水喝"想到的......][口试:75][作文:70][听力:71][阅读:76][综合:84][总分:365][证书:B]	原文	标注版
[1]:{CP人们{CC人间}不仅具有自私的品质{CC性质}，不具有为了团体的发展一起合作{CC协[C]助}的高尚道德品质。P} [国籍:韩国][性别:女][考试时间:199605][作文题目:由"三个和尚没水喝"想到的......][口试:85][作文:70][听力:74][阅读:76][综合:70][总分:366][证书:B]	原文	标注版
[2]:{CP所以我相信，三个和尚[B向]，通过一场打架或者讨论，为了生存，按照顺[C]序挑水喝。P} [国籍:韩国][性别:女][考试时间:199605][作文题目:由"三个和尚没水喝"想到的......][口试:85][作文:70][听力:74][阅读:76][综合:70][总分:366][证书:B]	原文	标注版
[3]:{CP年青一代喜欢有变化的生活，还喜欢赶潮流，反而长者却跟不上他们。P} [国籍:韩国][性别:女][考试时间:199610][作文题目:如何解决"代沟"问题][口试:75][作文:80][听力:73][阅读:75][综合:76][总分:370][证书:B]	原文	标注版
[4]:{CP应该我们站在别人的立场来想的话，不用为不好听的声音而烦脑。P} [国籍:韩国][性别:女][考试时间:200005][作文题目:关于用自然之声取代噪声的建议][口试:80][作文:70][听力:89][阅读:80][综合:74][总分:383][证书:B]	原文	标注版
[5]:{CP你们只有相信我，才能知道你们养好了我。P} [国籍:韩国][性别:男][考试时间:200105][作文题目:一封写给父母的信][口试:80][作文:75][听力:79][阅读:61][综合:74][总分:360][证书:B]	原文	标注版

Fig. 10 Results produced searching for the error text

3.1.3 Advanced Search

In version 2.0 of the HSK corpus, two advanced search features have been added: (1) search parameters for specific conditions, and (2) the ability to search for word collocations. The additions of these features have further enhanced the functionality of the HSK corpus as more results can be retrieved in a single search. In addition, it is also possible to search for the separable word "离" in the new 2.0 version.

The method of generating results based on specific search parameters in the HSK 2.0 corpus is suitable for retrieving specific sentence patterns, semi-fixed structures, and complex sentences with two marker words. The reason for the relatively powerful retrieval capabilities of this type of search is the use of regular expressions. Regular expressions[2] are quite common and general methods for corpus retrieval, yet they are relatively unfamiliar to linguistic professionals with a liberal arts background. The HSK corpus is easy to use and suitable for students of non-STEM majors due to the changes made based on the theoretical and practical background of liberal arts students. These changes include a liberal arts transformation of regular expressions, the simplification of mathematical equations into frame structures, and ensuring successful searching through filling mark words in corresponding positions.

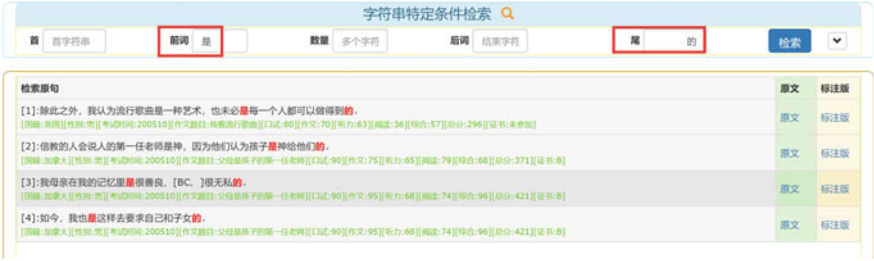
For example, the search of sentences containing the "是……的" construction (Fig. 11) and the "连" construction (Fig. 12).

Fixed structure retrieval with "爱……不……" (Fig. 13) and "一……就……" as examples (Fig. 14).

Take "或者……或者……" *or…or…* as an example for complex sentence retrieval (Fig. 15).

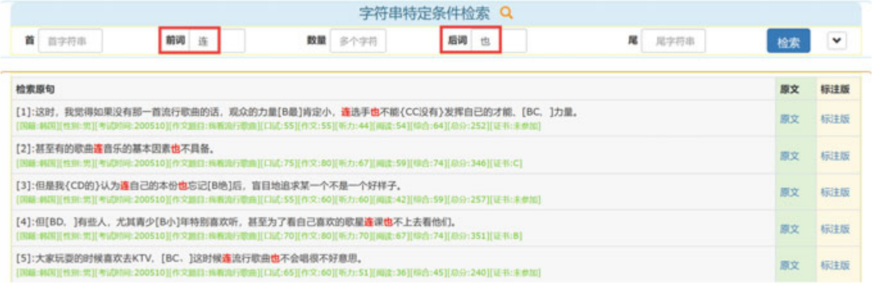
It is important to note that this search method is still based on form. This means that results will be generated as long as there are set search terms in the corpus,

[2] Regular expression is a kind of logical formula for string manipulation. It uses some pre-defined specific characters and the combination of these specific characters to form a "rule string". This "rule string" is used to express the pair of characters and is a kind of filtering logic for strings.



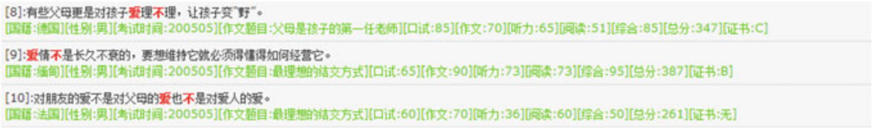
检索原句	原文	标注版
[1]:除此之外，我认为流行歌曲是一种艺术，也未必是每一个人都可以做得到的。 [国籍:美国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:80][作文:70][听力:63][阅读:36][综合:57][总分:296][证书:未参加]	原文	标注版
[2]:信教的人会说人的第一任老师是神，因为他们认为孩子是神给他们的。 [国籍:加拿大][性别:男][考试时间:200510][作文题目:父母是孩子的第一任老师][口试:90][作文:75][听力:65][阅读:79][综合:68][总分:371][证书:B]	原文	标注版
[3]:我母亲在我的记忆里是很善良、[BC、]很无私的。 [国籍:加拿大][性别:男][考试时间:200510][作文题目:父母是孩子的第一任老师][口试:90][作文:95][听力:68][阅读:74][综合:96][总分:421][证书:B]	原文	标注版
[4]:如今，我也是这样去要求自己和子女的。 [国籍:加拿大][性别:男][考试时间:200510][作文题目:父母是孩子的第一任老师][口试:90][作文:95][听力:68][阅读:74][综合:96][总分:421][证书:B]	原文	标注版

Fig. 11 Results produced searching for the “是……的” construction



检索原句	原文	标注版
[1]:这时，我觉得如果没有那一首流行歌曲的话，观众的力量[B量]肯定小，连选手也不能{CC没有}发挥自已的才能、[BC、]力量。 [国籍:韩国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:55][作文:55][听力:44][阅读:54][综合:64][总分:252][证书:未参加]	原文	标注版
[2]:甚至有的歌曲连音乐的基本因素也不具备。 [国籍:韩国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:75][作文:80][听力:67][阅读:59][综合:74][总分:346][证书:C]	原文	标注版
[3]:但是我{CD的}认为连自己的本份也忘记[B绝]后，盲目地追求某一个不是一个好样子。 [国籍:韩国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:55][作文:60][听力:60][阅读:42][综合:59][总分:257][证书:未参加]	原文	标注版
[4]:但[BD、]有些人，尤其青少[B小]年特别喜欢听，甚至为了看自己喜欢的歌星连课也不上去看他们。 [国籍:韩国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:70][作文:80][听力:70][阅读:67][综合:74][总分:351][证书:B]	原文	标注版
[5]:大家玩耍的时候喜欢去KTV，[BC、]这时候连流行歌曲也不会唱很不好意思。 [国籍:韩国][性别:男][考试时间:200510][作文题目:我看流行歌曲][口试:65][作文:60][听力:51][阅读:36][综合:45][总分:240][证书:未参加]	原文	标注版

Fig. 12 Results produced searching for the “连……也……” construction

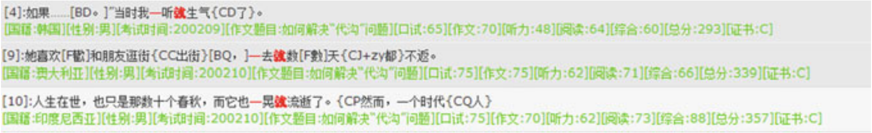


[8]:有些父母更是对孩子爱理不理，让孩子变"野"。
[国籍:德国][性别:男][考试时间:200505][作文题目:父母是孩子的第一任老师][口试:85][作文:70][听力:65][阅读:51][综合:85][总分:347][证书:C]

[9]:爱情不是长久不衰的，要想维持它就必须得懂得如何经营它。
[国籍:缅甸][性别:男][考试时间:200505][作文题目:最理想的结交方式][口试:65][作文:90][听力:73][阅读:73][综合:95][总分:387][证书:B]

[10]:对朋友的爱不是对父母的爱也不是对爱人的爱。
[国籍:法国][性别:男][考试时间:200505][作文题目:最理想的结交方式][口试:60][作文:70][听力:36][阅读:60][综合:50][总分:261][证书:无]

Fig. 13 Results produced searching for the “爱……不……” construction

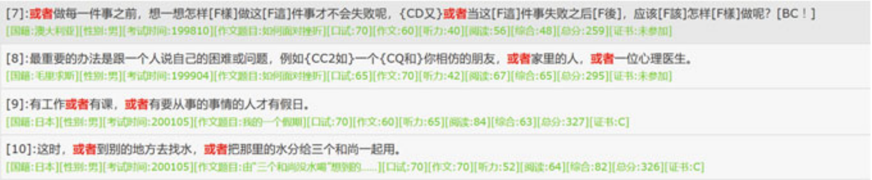


[4]:如果……[BD。]"当时我一听就生气{CD了}。
[国籍:韩国][性别:男][考试时间:200209][作文题目:如何解决"代沟"问题][口试:65][作文:70][听力:48][阅读:64][综合:60][总分:293][证书:C]

[9]:她喜欢[F歡]和朋友逛街{CC出街}[BQ，]一去就数[F數]天{CJ+zy都}不返。
[国籍:意大利亚][性别:男][考试时间:200210][作文题目:如何解决"代沟"问题][口试:75][作文:75][听力:62][阅读:71][综合:66][总分:339][证书:C]

[10]:人生在世，也只是那数十个春秋，而它也一晃就流逝了。{CP然而，一个时代{CQ人}
[国籍:印度尼西亚][性别:男][考试时间:200210][作文题目:如何解决"代沟"问题][口试:75][作文:70][听力:62][阅读:73][综合:88][总分:357][证书:C]

Fig. 14 Results produced searching for the “一……就……” construction



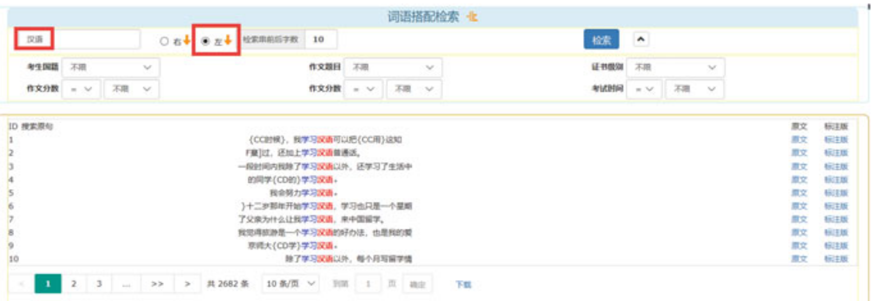
[7]:或者做每一件事之前，想一想怎样[F樣]做这[F這]件事才不会失败呢，{CD又}或者当这[F這]件事失败之后[F後]，应该[F該]怎样[F樣]做呢？[BC！]
[国籍:澳大利亚][性别:男][考试时间:199810][作文题目:如何面对挫折][口试:70][作文:60][听力:40][阅读:56][综合:48][总分:259][证书:未参加]

[8]:最重要的办法是跟一个人说自己的困难或问题，例如{CC2如}一个{CQ和}你相仿的朋友，或者家里的人，或者一位心理医生。
[国籍:毛里求斯][性别:男][考试时间:199904][作文题目:如何面对挫折][口试:65][作文:70][听力:42][阅读:67][综合:65][总分:295][证书:未参加]

[9]:有工作或者有课，或者有要从事的事情的人才有假日。
[国籍:日本][性别:男][考试时间:200105][作文题目:我的一个假期][口试:70][作文:60][听力:65][阅读:84][综合:63][总分:327][证书:C]

[10]:这时，或者到别的地方去找水，或者把那里的水分给三个和尚一起用。
[国籍:日本][性别:男][考试时间:200105][作文题目:由"三个和尚没水喝"想到的……][口试:70][作文:70][听力:52][阅读:64][综合:82][总分:326][证书:C]

Fig. 15 Search results generated for complex sentence retrieval



词语搭配检索

汉语 ○ 右 ● 左 检索串前后字数 10 检索

考生国籍 不限 作文题目 不限 证书级别 不限

作文分数 = 不限 作文分数 = 不限 考试时间 = 不限

ID	搜索原句	原文	标注版
1	{CC时候}，我学习汉语可以把{CC用}这知	原文	标注版
2	F童]过，还加上学习汉语普通话。	原文	标注版
3	一段时间内我除了学习汉语以外，还学习了生活中	原文	标注版
4	的同学{CD的}学习汉语。	原文	标注版
5	我会努力学习汉语。	原文	标注版
6	}十二岁那年开始学习汉语，学习也只是一个星期	原文	标注版
7	了父亲为什么让我学习汉语，来中国留学。	原文	标注版
8	我觉得旅游是一个学习汉语的好办法，也是我的爱	原文	标注版
9	京师大{CD学}学习汉语。	原文	标注版
10	除了学习汉语以外，每个月写留学情	原文	标注版

< 1 2 3 ... >> > 共 2682 条 10 条/页 到第 1 页 确定 下载

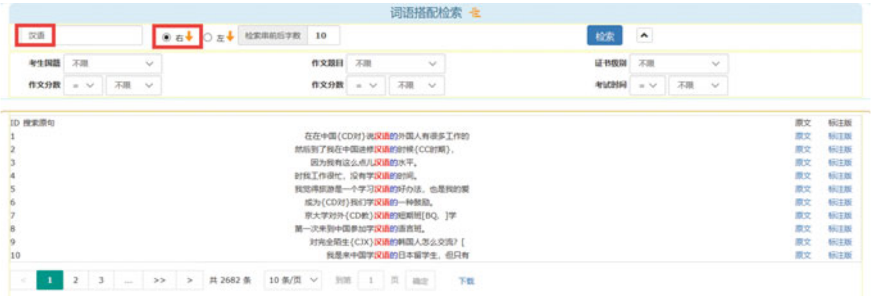
Fig. 16 Collocation retrieval example: “汉语” *Hànyǔ* Chinese left collocation situation

yet these results may not actually be the entire range of linguistic data housed in the corpus. For example, sentences “爱情不是长久不衰的”, “对朋友的爱不是对父母的爱也不是对爱人的爱” and the semi-fixed structure “爱……不……” have nothing to do with each other, even though there are “爱” and “不” in the sentence.

The ability to search for word collocations allows one to search for the co-occurring words preceding or following a word and their frequency. In this way, one can find out what words are collocated to the left or right of a certain word, count the corresponding collocation frequency, and sort them in descending order of frequency. This is a significant data retrieval method because it provides usage of words by demonstrating both the frequency and information obtained before and after a word. The generated results are equivalent to the “Word Collocation Dictionary”, which can serve as an important reference for Chinese language teaching.

Taking “汉语” *hanyu* “Chinese” as an example, the most frequent collocate on the left side is “学习” *xuexi* “study/learn”, with a frequency of 585 (Fig. 16); the second most frequent word is “学” *xue* “study”, with a frequency of 523. These are the two most frequent collocations. It can be seen that “学习汉语” *xuexihanyu* “learn Chinese” and “学汉语” *xuehanyu* “study Chinese” are the two collocations that learners use the most and have the best mastery. From the perspective of Chinese language teaching, these collocations are also the most important words to be taught to learners and should be the focus of teaching. The frequency of “对” *dui* “right” on the left is 48, while “觉得"*juede* “think/feel” only appears 9 times. The most frequent collocate on the right side is the auxiliary “的” *de*, with a frequency of 491(Fig. 17); a comma is used more often at the end of sentences on the right with a frequency of 344; the frequency of “有” *you* “have” after “汉语” *hanyu* “Chinese” is 28; and the frequency of “越来越” *yuelaiyue* “more and more” after “汉语” *hanyu* “Chinese” is only 4.

Details can be seen as follows.



ID 搜索原句 原文 标注版
1 在在中国{CD对}说汉语的外国人有很多工作的 原文 标注版
2 然后到了我在中国进修汉语的时候{CC时期}， 原文 标注版
3 因为我有这么点儿汉语的水平， 原文 标注版
4 时我工作很忙，没有学汉语的时间， 原文 标注版
5 我觉得旅游是一个学习汉语的好办法，也是我的爱 原文 标注版
6 成为{CD对}我们学汉语的一种鼓励。 原文 标注版
7 京大学对外{CD教}汉语的短期班[BQ，]学 原文 标注版
8 第一次来到中国参加学汉语的语言班。 原文 标注版
9 对完全陌生{CJX}汉语的韩国人怎么交流？[原文 标注版
10 我是来中国学汉语的日本留学生，但只有 原文 标注版

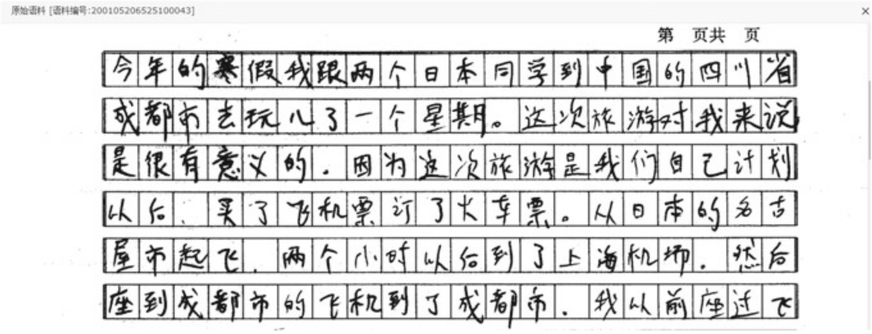
Fig. 17 Collocation Search example: "汉语" Hànyǔ Chinese right collocation situation

3.2 *Corpus Presentation*

In order to provide as much convenience as possible for users' teaching and research, background information regarding the corpus language data and its authors should be presented. In addition, this information should also accompany the search results generated by users when they use the corpus. Furthermore, the retrieved labeled results should conform to the original corpus in terms of form (e.g. composition, audio, video, etc.). In order to adapt to different browsing habits of users, it should be possible for users to adjust the number of results displayed on each page.

The background information of the language data refers to the author's nationality, gender, test time, composition topics, subjective oral and composition test scores, objective listening, reading, and comprehensive expression test scores, their total test results, and obtained certificates. This background information plays an important role in studying and judging learners' acquisition of Chinese. For examples of background information, see the green fonts in Figs. 13 and 14.

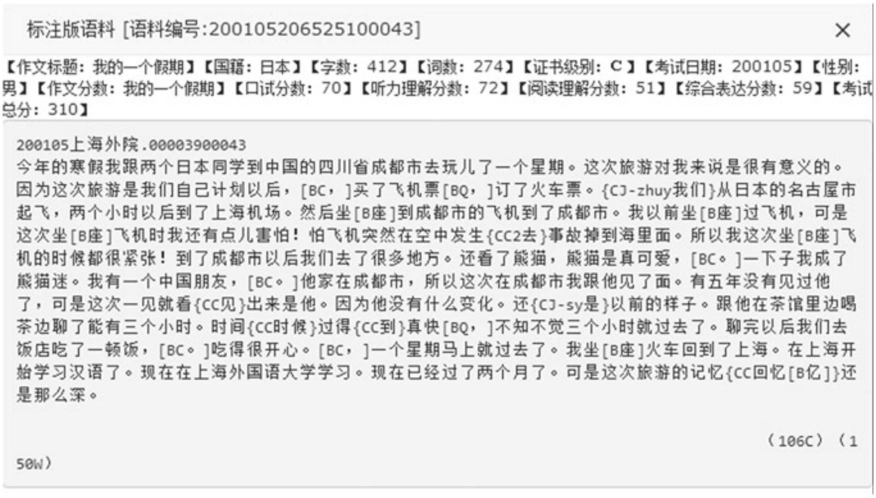
The following is a data entry of an original composition (Fig. 18) followed by the annotated full text in the corpus (Fig. 19).


原始语料 [语料编号:200105206525100043] ×

第 页共 页

今年的寒假我跟两个日本同学到中国的四川省
成都市去玩儿了一个星期。这次旅游对我来说
是很有意义的。因为这次旅游是我们自己计划
以后、买了飞机票、订了火车票。从日本的名古
屋市起飞，两个小时以后到了上海机场。然后
座到成都市的飞机到了成都市。我从前座过飞

Fig. 18 An original composition in the corpus



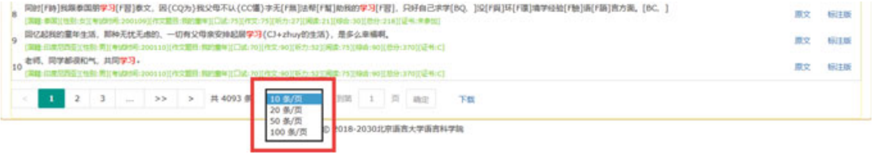
标注版语料 [语料编号:200105206525100043]

【作文标题：我的一个假期】【国籍：日本】【字数：412】【词数：274】【证书级别：C】【考试日期：200105】【性别：男】【作文分数：我的一个假期】【口试分数：70】【听力理解分数：72】【阅读理解分数：51】【综合表达分数：59】【考试总分：310】

200105上海外院.00003900043
今年的寒假我跟两个日本同学到中国的四川省成都市去玩儿了一个星期。这次旅游对我来说是很有意义的。因为这次旅游是我们自己计划以后，[BC，]买了飞机票[BQ，]订了火车票。{CJ-zhuy我们}从日本的名古屋市起飞，两个小时以后到了上海机场。然后坐[B座]到成都市的飞机到了成都市。我以前坐[B座]过飞机，可是这次坐[B座]飞机时我还有点儿害怕！怕飞机突然在空中发生{CC2去}事故掉到海里面。所以我这次坐[B座]飞机的时候都很紧张！到了成都市以后我们去了很多地方。还看了熊猫，熊猫是真可爱，[BC。]一下子我成了熊猫迷。我有一个中国朋友，[BC。]他家在成都市，所以这次在成都市我跟他见了面。有五年没有见过他了，可是这次一见就看{CC见}出来是他。因为他没有什么变化。还{CJ-sy是}以前的样子。跟他在茶馆里边喝茶边聊了能有三个小时。时间{CC时候}过得{CC到}真快[BQ，]不知不觉三个小时就过去了。聊完以后我们去饭店吃了一顿饭，[BC。]吃得很开心。[BC，]一个星期马上就过去了。我坐[B座]火车回到了上海。在上海开始学习汉语了。现在在上海外国语大学学习。现在已经过了两个月了。可是这次旅游的记忆{CC回忆[B亿]}还是那么深。

（106C）（150W）

Fig. 19 Annotated full text in the corpus



8 [illegible]
9 回忆起我的童年生活，那种无忧无虑的、一切有父母亲安排起居学习(CJ+zhuy的生活)，是多么幸福啊。
10 老师、同学都很和气，共同学习。

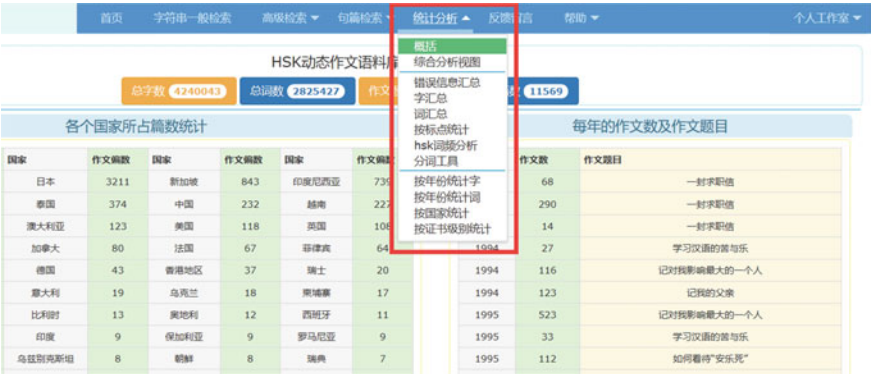
Fig. 20 Users can set the number of results displayed per page

Users can set the number of results displayed per page according to their own reading habits (Fig. 20).

3.3 *Statistical Analysis*

All the language data within the HSK corpus has been statistically analyzed. The data produced by the statistical analyses provide an overview of the corpus, including the total number of characters, words, composition topics, and text (Fig. 21); all kinds of error data related to characters, words, sentences, text, and punctuation (Fig. 22); characters and word information sorted by factors such as year, country, and obtained HSK certificate (Fig. 23). The data is especially useful for studying learners' acquisition of Chinese and can serve as an important source of reference for Chinese language teaching.

Examples of statistical charts are listed as follows.



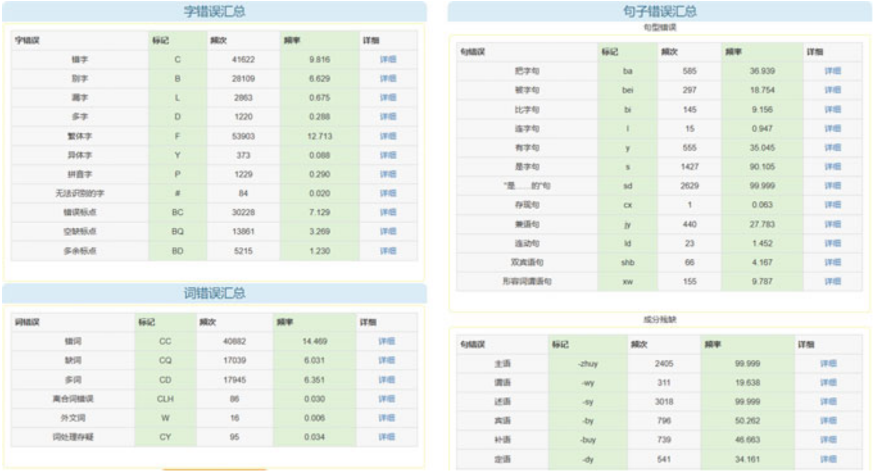
各个国家所占篇数统计

国家	作文篇数	国家	作文篇数	国家	作文篇数
日本	3211	新加坡	843	印度尼西亚	739
泰国	374	中国	232	越南	227
澳大利亚	123	美国	118	英国	108
加拿大	80	法国	67	菲律宾	64
德国	43	香港地区	37	瑞士	20
意大利	19	乌克兰	18	柬埔寨	17
比利时	13	奥地利	12	西班牙	11
印度	9	保加利亚	9	罗马尼亚	9
乌兹别克斯坦	8	朝鲜	8	瑞典	7

每年的作文数及作文题目

	作文数	作文题目
	68	一封求职信
	290	一封求职信
	14	一封求职信
1994	27	学习汉语的苦与乐
1994	116	记对我影响最大的一个人
1994	123	记我的父亲
1995	523	记对我影响最大的一个人
1995	33	学习汉语的苦与乐
1995	112	如何看待"安乐死"

Fig. 21 An overview of the corpus



字错误汇总

字错误	标记	频次	频率	详细
错字	C	41622	9.816	详细
别字	B	28109	6.629	详细
漏字	L	2863	0.675	详细
多字	D	1220	0.288	详细
繁体字	F	53903	12.713	详细
异体字	Y	373	0.088	详细
拼音字	P	1229	0.290	详细
无法识别的字	#	84	0.020	详细
错误标点	BC	30228	7.129	详细
空缺标点	BQ	13861	3.269	详细
多余标点	BD	5215	1.230	详细

词错误汇总

词错误	标记	频次	频率	详细
错词	CC	40882	14.469	详细
缺词	CQ	17039	6.031	详细
多词	CD	17945	6.351	详细
离合词错误	CLH	86	0.030	详细
外文词	W	16	0.006	详细
词处理存疑	CY	95	0.034	详细

句子错误汇总

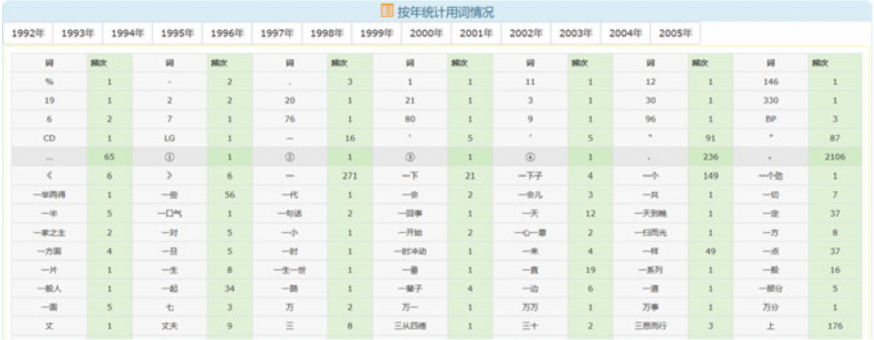
句型错误

句错误	标记	频次	频率	详细
把字句	ba	585	36.939	详细
被字句	bei	297	18.754	详细
比字句	bi	145	9.156	详细
连字句	l	15	0.947	详细
有字句	y	555	35.045	详细
是字句	s	1427	90.105	详细
"是……的"句	sd	2629	99.999	详细
存现句	cx	1	0.063	详细
兼语句	jy	440	27.783	详细
连动句	ld	23	1.452	详细
双宾语句	shb	66	4.167	详细
形容词谓语句	xw	155	9.787	详细

成分残缺

句错误	标记	频次	频率	详细
主语	-zhuy	2405	99.999	详细
谓语	-wy	311	19.638	详细
述语	-sy	3018	99.999	详细
宾语	-by	796	50.262	详细
补语	-buy	739	46.663	详细
定语	-dy	541	34.161	详细

Fig. 22 All error data types related to characters, words, and sentences



词	频次	词	频次	词	频次	词	频次	词	频次	词	频次	词	频次
%	1	-	2	.	3	1	1	11	1	12	1	146	1
19	1	2	2	20	1	21	1	3	1	30	1	330	1
6	2	7	1	76	1	80	1	9	1	96	1	BP	3
CD	1	LG	1	—	16	'	5	'	5	"	91	"	87
…	65	①	1	②	1	③	1	④	1	,	236	。	2106
《	6	》	6	—	271	一下	21	一下子	4	一个	149	一个劲	1
一举两得	1	一些	56	一代	1	一会	2	一会儿	3	一共	1	一切	7
一半	5	一口气	1	一句话	2	一回事	1	一天	12	一天到晚	1	一定	37
一家之主	2	一对	5	一小	1	一开始	2	一心一意	2	一扫而光	1	一方	8
一方面	4	一旦	5	一封	1	一时冲动	1	一来	4	一样	49	一点	37
一片	1	一生	8	一生一世	1	一番	1	一直	19	一系列	1	一般	16
一般人	1	一起	34	一路	1	一辈子	4	一边	6	一道	1	一部分	5
一面	5	七	3	万	2	万一	1	万万	1	万事	1	万分	1
丈	1	丈夫	9	三	8	三从四德	1	三十	2	三思而行	3	上	176

Fig. 23 Characters information sorted by year

3.4 Others Functions

In addition to the above functions, the corpus also provides functions such as crowdsourcing maintenance, message feedback, personal workspace, automatic downloading, and adding related resources to further enhance the ability to serve teaching and research.

(1) Crowdsourcing maintenance

Large-scale interlanguage corpora are usually annotated manually and rely on hundreds of annotators. Inconsistent annotation and labeling, and even errors and omissions, are inevitable. Although quality monitoring is carried out during the process, it still cannot solve the problem completely. According to the concept of crowdsourcing, allowing users to correct errors and omissions in the input version and marked version of a corpus is an effective way to improve the quality of that corpus.

Specific instructions: double-click on the corpus entry to open the dialog box for modification and editing. Edit the entry then click submit and wait for it to update. It will then go through backend screening and once finished will be published and replace the original entry. The screening process is carried out by the corpus administrator who decides whether the modification made by the user should replace the original entry. This is an especially important step in the process which helps avoid inadvertent and incorrect changes made by users. In this way, crowdsourcing can effectively improve the quality of a corpus' transcriptions and annotations so that it can better serve the majority of users.

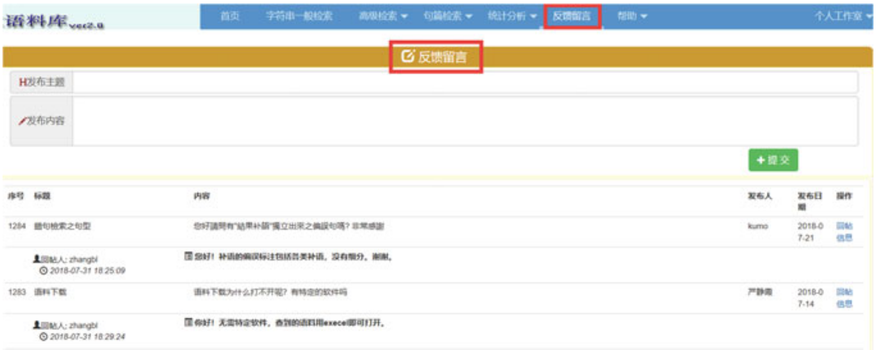
(2) Feedback

Users will inevitably encounter various questions that need to be answered in order to better use the corpus. In addition, they may have comments and suggestions that are important for the construction and improvement of the corpus. Thus, a medium for communication and feedback is necessary for effective communication between corpus constructors and users. The approach taken by the HSK corpus was to add a "feedback message" function to facilitate communication, exchange, and discussion between users and the corpus builder (Fig. 24).

From a practical point of view, this function serves a good communicative purpose.

(3) Personal workspace

Setting up a "personal workspace" in a corpus is a good idea and can have many practical functions. For example, users can maintain their own information, input clerks can extract corpus entries for input and transcription, annotators can label entries, and users can perform corpus-assisted analyses, research, and even write papers. In short, it can be a working platform for builders to build a corpus and for users to conduct related research. At present, the functions of the personal workspace in the HSK corpus are still inadequate and should be enriched.



序号	标题	内容	发布人	发布日期	操作
1284	整句检索之句型	您好请问有"结果补语"独立出来之偏误句吗? 非常感谢	kumo	2018-07-21	回帖 信息
回帖人: zhangbl 2018-07-31 18:25:09		您好! 补语的偏误标注包括各类补语，没有细分，谢谢。			
1283	语料下载	语料下载为什么打不开呢? 有特定的软件吗	严静雨	2018-07-14	回帖 信息
回帖人: zhangbl 2018-07-31 18:29:24		你好! 无需特定软件，查到的语料用execel即可打开。			

Fig. 24 Sample feedback message

(4) Automatic language data download

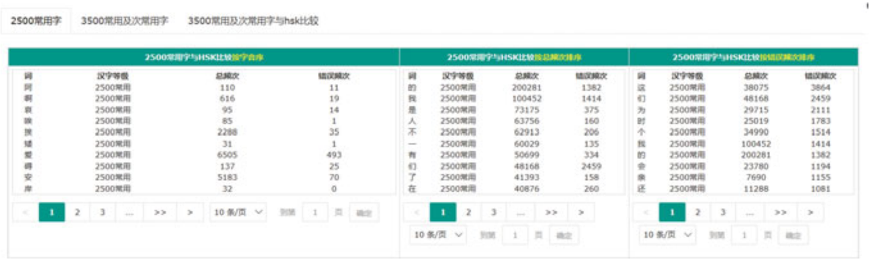
In response to the inconvenience reported by users of the corpus in the past, version 2.0 of the HSK corpus has been equipped with an automatic download feature. The generated results can be downloaded automatically (limited to 500). This fast and convenient feature forgoes the labor of manual copying and downloading each entry.

(5) Increase and accumulation of the relevant resources

The inclusion of practical resources closely related to Chinese language teaching in the corpus can provide users with great convenience in teaching and research. For example, in our study, we compared the characters used in the corpus with the characters used in the "Chinese Proficiency Vocabulary and Chinese Character Level Syllabus". We found that there are 2905 Chinese characters in the syllabus, and 3904 different Chinese characters in the HSK corpus. Learners mastered 999 more Chinese characters than stated in the syllabus. Among the 3905 Chinese characters, there are 2778 characters in the syllabus accounting for 71.16% and 1126 characters outside the syllabus accounting for 28.84%.

The comparison with the "List of Frequently Used Modern Chinese Characters" (1988) for native speakers shows that there are 3500 frequently used characters in the word list, which are divided into 2500 frequently used characters and 1000 sub-frequently used characters. Comparing the 3904 characters in the corpus with the "Frequently Used Characters List", there are a total of 3153 characters in the table, with 2452 frequently used characters accounting for 98.08% of 2500 frequently used characters and 701 sub-frequently used characters accounting for 70.1% of the 1,000 sub-frequently used characters.

Based on these studies and findings, we have compiled "A Comparison List of 2500 Frequently used Characters with HSK by Phonetic Errors", "A Comparison List of 2500 Frequently used Characters with HSK by Total Frequency ", and "A Comparison List of 2500 Frequently used Characters with HSK by Error Frequency", which have been put into the statistical information as an important reference for teaching Chinese characters (Fig. 25).



2500常用字与HSK比较按字音序

词	汉字等级	总频次	错误频次
阿	2500常用	110	11
啊	2500常用	616	19
哀	2500常用	95	14
唉	2500常用	85	1
挨	2500常用	2288	35
矮	2500常用	31	1
爱	2500常用	6505	493
碍	2500常用	137	25
安	2500常用	5183	70
岸	2500常用	32	0

2500常用字与HSK比较按总频次排序

词	汉字等级	总频次	错误频次
的	2500常用	200281	1382
我	2500常用	100452	1414
是	2500常用	73175	375
人	2500常用	63756	160
不	2500常用	62913	206
一	2500常用	60029	135
有	2500常用	50699	334
们	2500常用	48168	2459
了	2500常用	41393	158
在	2500常用	40876	260

2500常用字与HSK比较按错误频次排序

词	汉字等级	总频次	错误频次
这	2500常用	38075	3864
们	2500常用	48168	2459
为	2500常用	29715	2111
时	2500常用	25019	1783
个	2500常用	34990	1514
我	2500常用	100452	1414
的	2500常用	200281	1382
会	2500常用	23780	1194
来	2500常用	7690	1155
还	2500常用	11288	1081

Fig. 25 Statistics example graph

4 Conclusions

Through the review of the construction of the Chinese interlanguage corpora and the discussion of existing problems, we put forward some new concepts and functions of corpora construction in order to improve the level of corpus construction and better serve teaching and research.

1. The objective and fundamental purpose of building corpora are to serve Chinese language teaching and scientific research all around the world. The premise of ensuring this function is to make sure that the corpora are always open to the public. This requires the corpora systems to be secure without any high-or medium-risk vulnerabilities. This is a new situation and a new problem brought about by the development of new information technology, which must be paid great attention to by corpora builders.
2. Improvements in corpora software systems can enhance the functionality of corpora and can better meet users' needs. The improvement and enrichment of search and retrieval methods to enable users to query some words, phrases, and sentences that were previously unavailable are one such example of functionality enhancement. The rich and practical statistical information has important reference value for teaching and research. A user-friendly interface and the design of certain humanized functions, such as the autonomous setting of the number of corpora presentations and automatic downloading, can provide users with convenience and improve their user experience. The function which allows users to modify the transcriptions and annotations by crowdsourcing allows for the continuous improvement in the quality of corpora annotations.
3. Users have the most say on what kind of functions a corpus should have. Their questions, comments, and suggestions in the process of using a corpus are of great significance to corpus construction and should be understood in time and given feedback as soon as possible. Therefore, it is particularly important for corpus builders to communicate with users and to maintain a smooth and effective medium for communication as provided by the "feedback message" function.
4. In the past, corpora were designed and built in a simple and extensive way, which was in the initial stage of corpora construction, or Corpora Generation 1.0. The development of the HSK Corpus 2.0 has made us realize the important

role of software systems. A good software system can make the corpus powerful, easy to use, and possess "fine and rich" characteristics. It also promotes corpora construction into Generation 2.0. The development and transition from the simple Generation 1.0 to the refined Generation 2.0 reflect the developing progress of Chinese interlanguage corpora construction. This is also an inevitable result of the technological progress of the times.

There are some important characteristic differences between Generation 1.0 to Generation 2.0. as noted below:

Corpora labeling: individual layer labeling → comprehensive labeling
Labeling mode: error labeling → error labeling + basic labeling
Search method: simple search → advanced search
Construction concept subcontracting → crowdsourcing
Research Paradigm: Error Analysis → Comprehensive Investigation of Interlanguage
Data view: individual data → big data

It can be said that with HSK corpus version 2.0, the construction of Chinese interlanguage corpora has entered Generation 2.0 from Generation 1.0, with 2018 being regarded as the first year of Generation 2.0.

References

Chen, X. (1996). *Introduction to "Chinese interlanguage corpus system*"[C]. *Selected Papers of the Fifth International Chinese Language Teaching Conference*, Beijing: Peking University Press, pp. 459–467.

Chu, C., & Xiaohe, C. (1993). The basic idea of establishing the "Chinese interlanguage corpus system." *World Chinese Teaching, 3*, 199–205.

Cui, X., & Baolin, Z. (2011). "Global Chinese language learners corpus" construction plan. *Language Application, 2*, 100–108.

Xiao, X. et al. (2009). *Research on the difficulty and classification of Chinese sentence learning for foreign students.* Beijing: Higher Education Press.

Xiao, X., & Wenhua, Z. (2014). The comprehensiveness and classification of Chinese interlanguage corpus labeling. *World Chinese Teaching, 3*, 368–377.

Zhang, B. (2019). From 1.0 to 2.0_construction and development of chinese interlanguage corpus. *International Chinese Language Teaching Research* (4), 84–95.

Zhang, B. et al. (2019). *A Study on the Standardization of Chinese Interlanguage Corpus Annotation.* Beijing: Peking University Press.

Zhang, B., & Xiliang, C. (2015). On the construction standards of Chinese interlanguage corpus. *Language Application, 2*, 125–134.

Zhang, B., & Xiliang, C. (2018). New thoughts on the research of chinese interlanguage corpus annotation standards—also on the design of the "global Chinese interlanguage corpus, annotation standards[C]. *Selected Papers of the Third International Symposium on the Construction and Application of Chinese Interlanguage Corpus*, Beijing: World Book Publishing Company.

Zhang, B. et al. (2014). *A corpus-based study of Chinese sentence acquisition for foreigners*, Beijing: China Book Publishing House.

Zhang, B. et al. (2008). *Thematic research on Chinese vocabulary based on interlanguage corpus.* Beijing: Peking University Press.

Zhao, J. et al. (2008). *A study of Chinese syntax based on interlanguage corpus.* Beijing: Peking University Press.