ON THE USE OF CORPORA IN SECOND LANGUAGE ACQUISITION — CHINESE AS AN EXAMPLE

Mária IŠTVÁNOVÁ
Comenius University in Bratislava, Slovakia
istvanova6@uniba.sk

Abstract

This paper aims to introduce the language corpora and the advantages of their use in the process of Chinese language acquisition. We provide practical examples of the corpora’s direct and indirect use for teaching and learning Chinese as a second language. The exploratory approach towards Chinese by using various types of corpora is applicable for general language seminars as well as specialized translation seminars. The indirect use is mainly linked to the preparation of teaching materials and facilitates the curriculum design.

Keywords: Chinese; corpora; teaching methodology; second language acquisition; linguistic research

Povzetek


Ključne besede: kitajščina; korpusi; metodologija poučevanja; učenje drugega jezika; jezikoslovno raziskovanje
1 Introduction

The corpora play an important role in the teaching and learning of a second language and although their use in language analysis is regarded as a relatively new approach, we already perceive a dynamic relationship established between corpora and linguistic research (Casas-Pedrosa et al., 2013, p. 2). The increasing number of studies embedded in the field of corpus linguistics and numerous published articles reflect the rising interest in the use of corpora for second language acquisition and the study of a learner’s interlanguage. The emergence of corpora and their use in language learning transforms the traditional teaching approaches and methods used for the compilation of teaching materials.

2 Direct use of corpora in second language acquisition

The direct use of corpora enhances the process of second language acquisition, however, specialized training in corpus analysis for teachers and students is necessary to ensure a successful application of corpora in classroom instruction. Apart from that, the direct use of corpora requires certain adjustments in the curriculum design, therefore its incorporation in classroom activities is more common at higher educational institutions such as universities (McEnery & Xiao, 2011, p. 374). The introduction to corpus linguistics and familiarity with the user interface of the selected corpus facilitates the interpretation of the query, therefore in cases when students are not familiar with the use of corpora, the practical preparatory sessions would improve the efficiency of the corpora's use in the instruction.

Following the development of teaching methodology of a second language, the teacher's focus shifts from a linguistic theory to a student, and the student's ability to work independently is emphasized. We may therefore expect students to have mastered learning strategies, and find corpora a valuable source of information about the target language. At such a level, the students are able to easily compare various characteristics between their mother tongue and the studied language (Benko et al., 2019, p. 14). From the learners' perspective, the direct use of corpora in classroom instruction is beneficial because of their limited exposure to the target language, and the teacher's guidance is crucial especially for elementary or intermediate learners due to their limited knowledge of the target language (Tsui, 2004, p. 40). The exploration of a large amount of language material compiled in a corpus creates a suitable environment for the so-called learning as discovery approach. In this way, we enhance learner's individual research interests and improve their language awareness (Bernardini, 2004, p. 22). The use of corpora in instruction is closely linked to the data-driven learning method (DDL). This method is based on student's independent work with language data to explore the regularities and patterns of the target language (the corpus is perceived as the core teaching material). We shift the learner's passive
position of receiving information into an active position of a researcher of the second language (e.g., by using the concordance search feature or keyword in context – KWIC) (Johns, 1991, pp. 9-13). The integration of corpora and language instruction leads to the active participation of students, which further enables them to discuss the difficulties and learning strategies among each other. Apart from that, it strengthens the learner’s focus on the form and meaning as well as the relation between the two (Bernardini, 2004, p. 28).

Regarding the use of corpora with the objective to increase the learner’s language proficiency, the study of collocations represents an efficient way for acquiring the prefabricated units that are necessary for achieving native-like language proficiency, because the number of acquired collocations facilitates language comprehension and determines the learner’s language production (Cowie, 1994, p.3168). The perception of collocations differs among linguists and most linguists are convinced that the use of corpora to extract language data is the most reliable way of identifying collocations (Harris, 2006; Sinclair et al., 2004). The collocation boundaries are indistinct because collocations are either perceived as constructions composed of elements in direct proximity or in a broader sense, and the individual units are not necessarily followed one after another but they occur near the selected language unit (McEnery & Hardie, 2012, p. 123). Therefore, great importance is placed on the knowledge of lexical units with their co-occurring items, including their position in a sentence, instead of on individual words (McEnery & Xiao, 2011, p. 368).

3 Direct use of Chinese corpora

The use of Chinese corpora for teaching and learning purposes is favorable because of their quantitative approach in language study (Xu, 2019, p. 49). The corpora are suitable for displaying positive examples in a given context when focusing on the structures that need to be acquired by the learners (Chen & Tao, 2019, p. 75). The teaching approaches which base on information mined from the corpus concentrate on explanations of commonly used language forms in Chinese, and therefore the instruction time is focused mainly on the most relevant knowledge (Zhao & Kang, 2017, p. 86). An example of direct use of a corpus in the instruction has been introduced by Gajdoš in Chinese Legal Texts – Quantitative Description (Gajdoš, 2017). The subcorpus of Chinese legal texts as an integral part of the Hanku corpus enables teachers to concentrate on a specific field of language instruction. One of the prerequisites for the successful inclusion of corpora in the instruction is the learner’s knowledge of CQL query and regular expressions, which is necessary for the retrieval of the language data (Gajdoš, 2017, p. 86). The query functionalities and search options vary based on the user interface of a chosen corpus. Not only the knowledge of the use of corpora but also a refined knowledge of grammatical rules of the target language is an important
prerequisite that allows a corpus to be successfully included in the process of language learning.

Incorporation of corpora in the translation seminars requires revisions of the existing curricula and training of teachers in corpus linguistics focused on the acquisition of necessary skills related to software use, statistical analysis, and data interpretation (Hu, 2011, p. 184). The main purpose of the translation teaching methodology is the development of the learner's translation competence and translation awareness. The translation competence represents a student's ability to transfer meaning of the source language into meaning in the target language by employing suitable translation strategies. The translation awareness reflects the student's understanding of the nature of the translation subject, object, or receptor and the relation between them (Liu, 2003, p. 1). Corpora is a useful training tool for future translators because it enhances understanding of typical translation problems solved by experienced translators when working for instance with parallel corpora (Bernardini, 2004, pp. 20-21). Figure 1 (Hu, 2011, p. 148) displays a variation in the translation of the word *shehui 社会* into English as a target language. Based on the fixed terms in Chinese, it is necessary to select an appropriate translation and the query in the corpus facilitates the decision-making process, especially in the case of inexperienced learners and novice translators.

![Figure 1: Translation of word *shehui 社会* in different contexts](image)

As mentioned in the previous section, the study of collocations improves learner's language production and comprehension skills. In this way, corpora represent a powerful tool when getting acquainted with linguistic conventions in Chinese (Jing-Schmidt, 2019, p. 23). Collocation functionalities enable learners to extend their knowledge of prefabricated units and increase their ability to identify collocations in the context or to employ them in language production. Figure 2 displays a collocation of the verb *zuzhi 组织* followed by a noun by using Lancaster Corpus of Mandarin Chinese accessible on <corpus.leeds.ac.uk/query-zh.html>. The corpus functionality
among other statistical data (MI, Dice, and LL score) provides a T score and offers the possibility to display the chosen expression in a context with all corresponding entries.

**T score**

<table>
<thead>
<tr>
<th>Collocation</th>
<th>JointFreq1</th>
<th>Freq2</th>
<th>T score</th>
<th>Concordance</th>
</tr>
</thead>
<tbody>
<tr>
<td>组织结构</td>
<td>15</td>
<td>463</td>
<td>3.81</td>
<td>Examples</td>
</tr>
<tr>
<td>组织机构</td>
<td>7</td>
<td>207</td>
<td>2.60</td>
<td>Examples</td>
</tr>
<tr>
<td>组织形式</td>
<td>7</td>
<td>269</td>
<td>2.59</td>
<td>Examples</td>
</tr>
<tr>
<td>组织理论</td>
<td>7</td>
<td>344</td>
<td>2.57</td>
<td>Examples</td>
</tr>
<tr>
<td>组织群众</td>
<td>4</td>
<td>293</td>
<td>1.92</td>
<td>Examples</td>
</tr>
<tr>
<td>组织领导</td>
<td>4</td>
<td>382</td>
<td>1.90</td>
<td>Examples</td>
</tr>
<tr>
<td>组织程度</td>
<td>3</td>
<td>208</td>
<td>1.67</td>
<td>Examples</td>
</tr>
<tr>
<td>组织学生</td>
<td>3</td>
<td>469</td>
<td>1.58</td>
<td>Examples</td>
</tr>
<tr>
<td>组织职业</td>
<td>2</td>
<td>8</td>
<td>1.41</td>
<td>Examples</td>
</tr>
<tr>
<td>组织民兵</td>
<td>2</td>
<td>9</td>
<td>1.41</td>
<td>Examples</td>
</tr>
<tr>
<td>组织人力</td>
<td>2</td>
<td>17</td>
<td>1.41</td>
<td>Examples</td>
</tr>
<tr>
<td>组织城镇</td>
<td>2</td>
<td>22</td>
<td>1.41</td>
<td>Examples</td>
</tr>
</tbody>
</table>

**Figure 2:** Collocation of zuzhi 组织 followed by a noun

Corpora are useful for improving learner's knowledge of the general vocabulary, but it is also possible to concentrate on a specific field based on a learner's future specialization. Such specializations are mainly required for academic or any other professional purposes with the objective to achieve native speaker-like language proficiency in a particular field (Zhao & Zhang, 2015, p. 38). Monolingual, as well as parallel corpora, are useful reference sources. Figure 3 (Guan & Tao, 2017, p. 178) displays different translations of the word tiaojie 调解 into English. The provided information is extracted from the specialized legal corpora Beida fabao falv faqui shujuku 北大法宝法律法规数据库 and Wanshi falv xinxi shujuku 万事法律信息数据库 (Guan & Tao, 2017, p. 178). Its translation into the target language varies based on the context and particular use of the expression. In the case of an existing specialized corpus in the language combination of the learner’s mother tongue and Chinese, the corpora in the Chinese – English language combination might be employed as a supplementary reference. Translation of this term in a general web-based dictionary of Chinese is *to mediate* and *to bring parties to an agreement, bring together to an agreement* (retrieved from <mdbg.net> and <www.chinese-dictionary.org>). Specialized legal dictionary – *Chinese-English/English-Chinese Pocket Legal Dictionary* (Chen, 2008, p.19) also provides the translation of tiaojie 调解 as *mediation, to mediate*. In comparison to entries in the dictionaries, the query in the corpora provides another possibility of a translation in English – *to conciliate, conciliation*. 
Apart from the use of parallel corpora in the field of translation didactics, we generally consider the parallel corpus useful for explaining studied expressions within the context, because it provides a conceptual framework for their understanding in the learner’s native language as well as in their target language, Chinese in this case (Bluemel, 2019, pp. 85-86). Figure 4 displays query of *suzhi* 素质 extracted from *CCL Chinese-English parallel corpus* accessible on the website of Center for Chinese linguistics PKU (*Beijing daxue Zhongguo yuyanxue yanjiu zhongxin* 北京大学中国语言学研究中心, <ccl.pku.edu.cn:8080/ccl_corpus/index_bi.jsp>). We selected the first ten entries to display the variability in the English translation. The prevailing translation is *quality*, but there are also two entries translated into English as *competence* and one entry translated as *consciousness*. The chosen expression reflects a specific term in Chinese, and the context, as well as several translation options, enhance its correct understanding by learners. In comparison to query in the parallel corpus, the translation of *suzhi* 素质 in the web-based dictionaries (<mdbg.net> and <chinese-dictionary.org>) is *inner quality* and *basic essence*. In this case, the use of corpora provides valuable guidance for the selection of an appropriate translation in the target language.
The incorporation of corpora in classroom instruction is not a trouble-free process. It is on a teacher to decide whether it is necessary to compile a new corpus corresponding to the requirements of the teaching objectives, or whether the existing one would fulfilling such prerequisites. Despite the fact that most of the commonly taught second languages already have several existing corpora, there are nevertheless cases when creating a new corpus is necessary. In such a case, a teacher has to deal with curriculum adjustments and would face general issues related to the creation of a corpus including the compilation procedure, representativeness, or source of language material (Casas-Pedrosa et al., 2013, p. 2). It is therefore necessary to create more corpora suitable for didactical use in the frame of their content and design to enhance their use in language instruction (McEnery & Xiao, 2011, pp. 374-375). The direct use of corpora in the instruction is still less frequent in comparison to its indirect implications, therefore it is necessary to increase the extent of the employed data-driven learning approaches in the process of Chinese language acquisition and start using more computer-assisted software tools that facilitate the learning process (Xu, 2019, p. 47). When working with corpora, it is important to keep in mind that the search results are restricted to the findings in that particular corpus, which is processed by different tools (e.g., the text is tokenized and annotated). These factors also have an impact on the results extracted from the corpus (Gajdoš, 2020, p. 122).

Based on the variety of introduced approaches aimed at a direct use of corpora, we expect that the use of different Chinese corpora becomes an inseparable part of a teaching and learning practice in the near future, especially with successive attempts to simplify procedures related to the creation of corpora thanks to fully automated or semi-automated processes.

### 4 Indirect use of Chinese corpora

The indirect use of corpora is often linked to the revision of an existing curriculum or to designing a new one, as well as the creation of teaching materials. The advantage of corpus-based teaching materials is that the examples reflect real language utterances.
We expect that learners improve their communication skills when using this kind of teaching materials because they get used to the language utterances of the native speakers and the distinction of specific nuances in Chinese becomes more evident (McEnery & Xiao, 2011, pp. 367-368). The traditional compilation of teaching materials used to be based mainly on the individual preference, language intuition, and teaching experience of the author. In comparison to the teaching materials created with the use of corpora, the research aimed at the content of the existing teaching materials compiled with the traditional method showed that in many cases, descriptions of words or expression utterances are inconsistent with the natural language usage (Luo, 2008, p. 48). The frequency lists of morphemes, expressions, and grammar structures are one of the main determining features of corpus-aided teaching materials. The textbook's difficulty level is easily controllable using corpus functionalities, and the arrangement of the content is considered more reasonable. We determine the textbook's difficulty level by the text's length, the core grammar structures to be acquired as well as by the ratio of the new vocabulary and the total word count. In addition, it is possible to adjust the occurrence rate of more common words or increase the reoccurrence rate of new words (Zhao & Kang, 2017, p. 85).

Chi-Editor (Hanyu yuedu fenji zhinanzhen 汉语阅读分级指难针) accessible on <langaugedata.net/editor> represents a useful tool for assessing the difficulty of the textual material chosen for the instruction practice. Its functionality enables a user to retrieve language material that is sorted out based on individual proficiency levels together with the vocabulary lists corresponding to the requirements of HSK, as well as Chinese syntactic and lexical structures (Jin et al., 2018). Figure 3 (Jin et al., 2018, p. 7) displays a sample text with different colors according to varying difficulty of the vocabulary. As the authors of Chi-Editor state, the proficiency levels are in accordance with International Curriculum for Chinese Language Education published by Confucius Institute Headquarters in 2015 (Jin et al., 2018). As revisions of the HSK exams have been announced, we expect certain adjustments and updates to follow. At this point, we consider Chi-Editor as a suitable tool for the regular preparation of teaching materials, and its use is not restricted by the potential absence of an update in the future.
When focusing on advanced learners of Chinese, there is still a rather limited number of studies devoted to them and their academic Chinese as a target language. Learners’ corpora composed of texts written by students at the advanced level depicts learners' needs in the frame of creating new teaching material (Chen & Tao, 2019, p. 58).

Not only the Chinese learner corpus but also the corpus used for the analysis of existing teaching materials such as Duiwai Hanyu jiaocai yuliaoku 对外汉语教材语料库 created by National Language Resources Monitoring and Research Center at the Xiamen University represents a helpful tool. It contains texts taken from widely used teaching materials for international students of Chinese that were published between 1992 and 2006. Based on the above-mentioned corpus, researchers or teachers are able to study and assess the content of the existing teaching materials (Su et al., 2010) and conclude whether rearrangements are necessary to increase teaching efficiency. The corpus is accessible on the website of the research center and provides a complete list of included teaching materials. The study of the existing materials is necessary for the verification of the teaching materials’ compatibility with the curriculum requirements. The corpus composed of texts from existing teaching materials is also useful for the retrieval of studied vocabulary or structures to display all entries in the corpus with corresponding examples in the context to strengthen the learner’s knowledge (Luo, 2008, p. 48).

In the more recent corpus linguistics research, there are increasing tendencies aimed at the automation of processes facilitating the use of corpora. It is possible to sort out the language material in the corpus based on the proficiency level. An evident benefit of corpus-aided teaching materials is the natural language providing examples for vocabulary lists enhancing the acquisition of authentic language by the learners (Zhao & Kang, 2017, pp. 85-86). In addition to traditional methods, it is possible to supplement the exercise design based on the findings from
the corpora which focus on the use of collocations and lexical bundles (Chen & Tao, 2019, p. 64).

The indirect use of corpora is also applicable in the field of lexicography, and numerous existing dictionaries compiled by the use of corpora continue to grow. One of the examples is *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners* compiled by Xiao et al. (2009), *Frequency-based HSK vocabulary* by Yang Ying (2016), covering all six proficiency levels in the individual volumes, and the vocabulary is supplemented by example sentences.

The indirect use of corpora for the creation of teaching materials or frequency dictionaries not only spares the time needed for their compilation but also increases the relevance of the selected content. Corpora composed of teaching materials such as the one created at the Xiamen University, or the tool *Chi-Editor*, facilitate the selection of appropriate teaching materials. In case that the employed teaching materials are not identical with the processed textbooks in *Duiwai Hanyu jiaocai yuliaoku* 对外汉语教材语料库, it might serve as a suitable model for teachers who create corpora composed of their own teaching materials to facilitate the analysis.

5 Conclusion

The direct use of corpora in the instruction is beneficial to teachers and students because they are transformed into active participants of the instruction practice. Students learn how to work with corpora and its features and gain experience in interpreting the query results. As experienced corpus users, learners obtain the opportunity to improve their proficiency level based on the examples of native speakers reflecting the natural language production of Chinese. The corpora as a part of translation seminars enhance the students' translation skills and quality of translation. Corpora besides enables learners to concentrate on a specific field of Chinese and teach how to explore the regularities of the target language. The indirect use of corpora is mainly employed for the creation of teaching materials, dictionaries, and curricula, providing the students with a reliable source of native speaker-like language utterances. From the perspective of the authors of teaching materials, the processing time is relatively short and the contents are easily controlled using corpus functionalities with the objective to increase the efficiency of the acquisition process. Taking into consideration a limited number of teaching materials in the Slovak – Chinese language combination, the continuous development of the use of corpora (both direct and indirect) not only facilitates linguistic research but improves the instruction practice, and as such represents an important source for reference.
Acknowledgments

This work was supported by the Comenius University in Bratislava as a Grant for young researchers under grant number UK/306/2021.

Abbreviations

DDL Data Driven Learning
KWIC Key Word in Context
CQL Corpus Query Language
HSK Hanyu shuiping kaoshi

References


