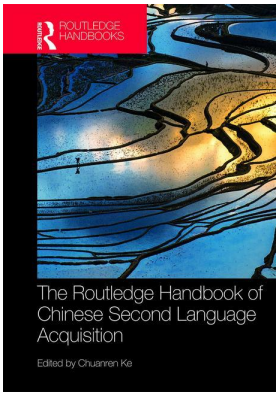

The Routledge Handbook of Chinese Second Language Acquisition

Chuanren Ke

Corpus-based research in Chinese as a second language

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315670706-3
Jie Zhang, Hongyin Tao
Published online on: 03 Apr 2018

3

Corpus-based research in Chinese as a second language

Jie Zhang and Hongyin Tao

Historical perspectives

Over the past several decades, the field of linguistics has witnessed a significant paradigm shift from the study of language as an abstract mental representation to the study of language in actual use. Corpus linguistics analysis, which is based on electronically stored and automatically processed large collections of language samples, makes it possible to systematically study patterns of natural language use. John Sinclair, a pioneer of corpus linguistics, defined a corpus as "a collection of naturally-occurring language text, chosen to characterize a state or variety of a language" (1991, p. 171). Compared with other linguistic approaches, corpus linguistics has several advantages. It bases linguistic analysis on naturally occurring data rather than intuition or introspection. Corpus data are empirical, constituting a rich resource for uncovering patterns of language use in natural contexts. Corpus linguistics utilizes a large and principled collection of texts, making it easier to compare different varieties and genres of a language or languages. With the help of data processing software and text retrieval programs, users can work with large-scale data using both automatic and interactive techniques (Biber, Conrad, & Reppen, 1998).

As a specific type of corpora, learner corpora (LC) are "digital representations of the performance or output, typically written, of language learners" (Barlow, 2005, p. 335). Learner corpora collect language data from second language (L2) learners and native speakers (NSs). Learner data can represent learners of the same first language (L1), learners of different L1s, and learners at different proficiency levels. NS data are preferably collected from those who are comparable to the learners in age, education, and sociocultural status. Depending on the media in which learner language is collected, LC can be classified into three broad types: written LC, spoken LC, and multi-modal LC. Written LC collect essays or other types of writing that learners produce. Spoken LC collect learners' spoken language in various learning and testing scenarios. Written LC are more common than spoken LC due to practical reasons. Written LC are less labor-intensive and more controllable in data collection, and therefore often take comparatively less time. When building spoken LC, there are more stringent demands in terms of collecting, storing and processing data, so this process often takes longer and involves more manual work. Multi-modal LC represent a new model of LC, which collect written or spoken data together with the context in which the learning takes place. This could include, for example, course materials, language input from the instructors, interactions with peer classmates, activities performed in the classroom, and so forth.

Similar to a native corpus, a learner corpus can be composed of 'raw' or plain text data; it can also be annotated. A raw learner corpus can be annotated to show the POS (part-of-speech) of each word by adding grammatical labels. A POS-tagged corpus can be linguistically parsed by identifying and labeling the function of a word or group of words in a sentence. It may be further parsed to show the sentence structure and the function of the different word classes. Using a text retrieval program, researchers can generate the wordlist of a corpus, search for keywords in context (KWIC) and collocations, and compare the wordlist of a corpus against that of a reference corpus.

A unique feature of LC is error annotation. Using a text retrieval program, an error-tagged learner corpus enables researchers to search for any error type, sort errors in different ways, and analyze them in their context. There are two main models of error tagging. The traditional model first develops an error tagging system based on a pilot analysis of the LC and then uses human coders to identify and apply error tags. The more promising way is computer automatic error detection and annotation. Even with automatic annotating programs, however, researchers oftentimes need to develop their own tagging schemes and manually annotate errors to address their specific research questions.

The most well-known learner corpus is the International Corpus of Learner English (ICLE), which was initiated in the early 1990s by Sylviane Granger and her associates at the Université catholique de Louvain in Belgium. The corpus has collected essays written by higher intermediate to advanced learners of English as a Foreign Language (EFL). As a product of international collaborative efforts, the ICLE v.2 now contains 3.7 million words of EFL writing from over 3,000 learners representing 16 L1 backgrounds. It systematically documents more than 20 tasks and learner variables (Granger *et al.*, 2009). According to Granger (2003), ICLE is best suited for analyzing high-frequency linguistic phenomena at the morphology, grammar, lexis, and discourse levels.

Over the past 20 years, a considerable body of literature based on LC has been produced across a wide range of lexical and grammatical topics, including tenses (Granger, 1999), lexical bundles (Biber, 2006), collocations (Nesselhauf, 2005), and phraseology (Granger & Meunier, 2008), to name a few. The inaugural meeting of the biannual conference of Learner Corpus Research (LCR) in 2011, the establishment of the Learner Corpus Association in 2013, and the publication of the first issue of the *International Journal of Learner Corpus Research* in 2015 all attest to the growing importance of this research tradition to the field.

In the context of Chinese, LC have been known as 学习者语料库 'learner corpus' or as 中介语语料库 'interlanguage corpus'. Chinese as a second language (CSL) learner corpus research first began to appear in the late 1990s and has blossomed over the past ten years. With more LC becoming available, the scope of CSL corpus studies has expanded tremendously. Researchers have explored a wide range of inquiries regarding how CSL learners acquire the different levels and aspects of Chinese. This research tradition is further strengthened by the convening of the biennial CSL learner corpus symposium, which first convened in 2012, and the subsequent conference proceedings reporting exclusively on CSL learner corpora construction and applied research. Following this lead, the first international conference on corpora of Chinese spoken interlanguage was convened in 2015.

Core issues and key findings

Learner language

The description of learner language, in particular learner errors, has been a central theme of CSL LCR. Early learner corpora, such as the L2 Chinese Interlanguage Corpus and the HSK Dynamic Composition Corpus, which we will introduce in the next section, were only tagged for learner

errors rather than language use in totality. Therefore, early CSL LCR used error analysis as its primary analytical framework. Researchers described the taxonomies of overuse, underuse, and misuse of a target linguistic feature, counted the frequency of the error types, and then tried to find explanations for these errors. Later research using error analysis as the analytical lens moved from pure description of error taxonomy to exploring motivations for interlanguage errors. An excellent example is B. L. Zhang's (2010a) investigation of CSL learners' use of the *ba*-sentence using the HSK Dynamic Composition Corpus. It has been a standing belief of CSL practitioners that CSL learners have tended to avoid using the *ba*-sentence. However, using large-scale corpus data, B. L. Zhang found that the avoidance of the *ba*-sentence is not as serious an issue as people had thought, and learners did not intentionally avoid the *ba*-sentence. The underuse of the *ba*-sentence was mainly because learners failed to grasp when to use the *ba*-sentence due to the lack of counterparts in their L1s. B. L. Zhang further pointed out that learners tended to overgeneralize the *ba*-sentence in contexts where other prepositions should be used. He found that the co-existence of underuse and overuse was due to the complexity of the target structure and students' inaccurate understandings of the semantic and pragmatic meanings of the structure, as well as the excessive instruction on the *ba*-sentence in the classroom (also see Chapter 8, Lu & Ke, this volume). This study showcases the advantage of large-scale learner corpora in investigating and evaluating speculations about CSL development using empirical data.

It was quickly realized that investigating errors only could not provide an adequate picture of how well a linguistic feature is acquired by learners. To understand the characteristics of learner language, one needs to study both errors and correct usages. The norm of current CSL LCR is to look at language use in its totality. Using the L2 Chinese Interlanguage Corpus, Cui (2005) examined the acquisition of 20 prepositions by European and American learners. He not only looked at learner errors but compared the frequency of these prepositions by learners with different L1s. He found that European and American learners, regardless of their respective L1s, tended to use prepositions more frequently than Japanese and Korean learners, as well as Chinese native speakers (also see Chapter 8, Lu & Ke, this volume). Drawing on a self-compiled written corpus, J. Zhang (2011, 2014) investigated intermediate and advanced CSL learners' lexical development of resultative verb compounds (RVCs). In addition to linguistic accuracy, she looked at the frequency of use and the component versatility of RVCs. Her examination of the three dimensions of RVCs revealed that learners acquired RVCs in three phases: the whole-word formula phase, the emergence of compound awareness phase, and the solidified compound awareness and lexical development phase. In addition, she found that different types of RVCs (directional RVCs, completive RVCs, and result-state RVCs) showed different patterns of development and posed different acquisition difficulties for learners (also see Chapter 8, Lu & Ke, this volume).

First language (L1) transfer

The effect of L1 influence or transfer is a topic of standing interest in SLA. CSL researchers are interested in uncovering the L1 influences in CSL learner language. Huang (2013), for example, looked at whether L1 backgrounds have an effect on advanced learners' character acquisition. Using the HSK Dynamic Composition Corpus, he found there was no L1 effect on the writing of simple characters, but there was an L1 effect on the writing of compound characters. The error rates ranking from high to low were Koreans, Europeans and Americans, and Japanese. Huang noted that Koreans demonstrating a higher error rate than the Europeans and Americans might be due to the relatively smaller data sample of European and American learners. In addition to the L1 effect, the structure of characters and stroke number were also important

factors. The outside-inside structure induced significantly more errors than the left-right or top-bottom structures, and characters with more components induced more errors (also see Chapter 6, Zhang & Ke, this volume).

Acquisition order and developmental sequences

The field of SLA has found substantial evidence for the claim that learners seem to acquire grammatical structures in a relatively fixed order. Also, when acquiring a grammatical structure in a second language, learners seem to follow predetermined stages of development. In this regard, learner corpora have been used to identify CSL learners' acquisition order and developmental sequences. Using the L2 Chinese Interlanguage Corpus, Shi (1998) identified the acquisition order of 22 typical Chinese syntactic structures by Korean- and English-speaking learners. She used the percentage of correct instances in total usages as the criterion of successful acquisition. The study identified the acquisition order of the 22 syntactic structures, which was roughly consistent among learners of different L1s. The study also revealed some degrees of variation among certain learner groups and individuals. Using the same corpus, Yang (2003a, 2003b, 2004) examined how American, Korean and Japanese learners at different proficiency levels acquired directional complements (DCs). He found a similar acquisition order regardless of learners' L1 backgrounds: (1) verb + simple DC (literal meaning), (2) verb + simple DC (extended meaning), (3) verb + compound DC (literal meaning), (4) verb + simple DC (extended meaning) with an object, (5) verb + DC1 + object + DC2 (literal meaning), (6) verb + DC1 + object + DC2 (extended meaning), (7) verb + compound DC (extended meaning), (8) verb + compound DC (extended meaning) + object, (9) verb + simple DC (literal meaning) with an object, and (10) verb + compound DC (literal meaning) + object.

Learner variability

An important inquiry of SLA is to understand learner variability in second language development. Large-scale and longitudinal corpora with dense data points that trace learner development over time provide an ideal resource to study learner variability. J. Zhang and Lu (2013), using a self-compiled written corpus, investigated CSL learners' use of Chinese numeral classifiers from a dynamic systems approach (de Bot, Lowie, & Verspoor, 2007). The corpus was a longitudinal corpus of 657 essays written by CSL learners at lower and higher intermediate levels. They first analyzed the inter- and intra-individual variability in learners' development of the fluency, diversity, and accuracy of numeral classifiers. They then closely examined the use of numeral classifiers by four focal learners at the higher immediate level. Learners were found to exhibit varied, nonlinear development for all three dimensions, accompanied by different degrees of fluctuation and regression in the process. They also reported that different dimensions of language development (linguistic fluency, diversity, and accuracy) did not develop in parallel, but rather interacted with each other in divergent ways.

Effects of language backgrounds on CSL development

A recent advancement of CSL LCR is the study of the language development of learners from different L1 backgrounds and heritage language backgrounds to uncover the special impacts of specific language backgrounds on language development. Jing-Schmidt (2011), for instance, used small-scale self-compiled corpora to understand CSL learners' acquisition of a salient Chinese discourse feature, zero anaphora, which refers to the omission of pronouns when references are

clear in context. The learner corpora comprised 53 compositions written by advanced CSL students from three language backgrounds: China-born Chinese heritage language learners, US-born Chinese heritage language learners, and non-heritage language learners. Native Chinese and native English corpora of similar argumentative writing were used as baseline data. The study found significant differences among the three learner groups in their frequency of use of zero anaphora and pronouns. China-born heritage language learners used zero anaphora in a way more akin to Chinese native speakers, whereas non-heritage language learners used pronouns in a way more like English native speakers. The findings suggested that language backgrounds and Chinese learning experience both played a role in acquiring Chinese discourse patterns. This research called for explicit instruction of Chinese discourse structures.

Research approaches

Construction of CSL learner corpora

To conduct a learner corpus study, one needs to have access to a systematically designed, well balanced, carefully annotated learner corpus. Thanks to the endeavors of several research teams in mainland China and Taiwan, researchers are now able to access a handful of CSL LC. We here provide brief summaries of a few widely used LC in the CSL LCR literature.

L2 Chinese Interlanguage Corpus (汉语中介语语料库系统**).** From 1993 to 1995, the research team led by Chengzhi Chu and Xiaohe Chen at the Beijing Language and Culture University constructed the first CSL learner corpus, the L2 Chinese Interlanguage Corpus (Chu *et al.*, 1995). The corpus collected 5,774 essays and written materials by 1,635 CSL students from 96 different countries studying Chinese at nine universities in China. The raw text consisted of approximately 3,528,988 characters, among which 1,731 essays (totaling 1,041,274 characters) were annotated and later included in the corpus. All essays were POS tagged, parsed, and error tagged. Learner metadata were carefully documented for 23 properties ranging from biographic information and Chinese learning experiences and motivations to topics, types and time of writing (Chu & Chen, 1993). The system is supported by a retrieval system allowing users to search by character, word, sentence, and discourse levels, or by learner metadata. This corpus, however, remains unavailable for public use.

HSK Dynamic Composition Corpus (HSK 动态作文语料库**).** The HSK Dynamic Composition Corpus Version 1.1 is a free, online, searchable database constructed and managed by the International Research and Development Center for Chinese Education at the Beijing Language and Culture University. It collects essays written by non-native Chinese speakers who took the HSK Test (Advanced Level) from 1992 to 2005. The corpus has collected 11,569 essays with approximately 4.24 million characters. 88.81% of the learners who contributed to the corpus are from an Asian country or region, with Korean and Japanese learners contributing 63.9% of the data (Huang, 2012). The language levels represented were intermediate to advanced based on the HSK scale. Each composition was annotated with a header that provided learner metadata including gender, nationality, HSK written score, HSK spoken test score, HSK listening score, HSK reading score, HSK comprehensive expression score, HSK total score, and certificate awarded. This corpus was error tagged at the levels of character, punctuation, lexicon, grammar, and discourse. The genres were mainly narrative and argumentative. With registration, users can search the corpus and access the original scanned copies of the compositions at http://202.112.195.192:8060/hsk/login.asp.

National Taiwan Normal University (NTNU) Chinese Character Errors Corpus (國立台灣師范大學漢語學習者漢字偏誤數據資料庫**), Chinese as a Second Language**

Spoken Corpus (華語為第二語口語語料庫), and TOCFL Learner Corpus (華語學習者語料庫). The research teams at National Taiwan Normal University developed the Chinese Character Errors Corpus (CCEC), the first of its kind, that exclusively compiles learner errors in the writing of Chinese (traditional) characters. This corpus contains 2,536 erroneous characters by students from 15 different L1 backgrounds (English, French, German, etc.). Because of the abundance of richer data sources of learners from Japan and Korea in other CSL LC, the CCEC decided not to include learners from these two countries (R. Zhang, 2013). The data were mainly students' writing at the beginner, intermediate, and advanced levels. Erroneous characters were scanned and included in the corpus. The corpus was annotated only for character errors.

The spoken learner corpus (華語為第二語口語語料庫), developed at NTNU and sponsored by the ROC government, is a rare collection of spoken learner data. It is based on the standard Mandarin test, called Test of Chinese as a Foreign Language (TOCFL, 華語文能力測驗, Chang, 2016), that has been used in Taiwan and overseas. Test takers are grouped into the basic and advanced categories and have come from various countries. However, the corpus as of 2016 includes learner data only from three language backgrounds: English, Japanese, and Korean, with 450 people/tests and 773,000 characters. The corpus is online and searchable at http://140.122.83.243/mp3c/.

The TOCFL Learner Corpus, built at the same institution, included essays that students wrote for TOCFL (Chang 2013, 2014a, 2014b, 2016). Since 2006, 5,092 essays (totaling about 1,740,000 characters and 1,140,000 words) by learners from 42 different L1 backgrounds have been collected, among which 2,837 essays with 989,045 characters are error tagged. Following the CEFR (Common European Framework of Reference) standards, this corpus covered the levels of A2, B1, B2, and C1 (Chang, 2013). Student metadata included native language, source, text genre, text function, text length, text type, score, and CEFR level. An online searchable interface of this corpus can be accessed at http://tocfl.itc.ntnu.edu.tw. A parallel composition corpus that collects student writing in non-testing situations can be searched at http://kitty.2y.idv.tw/~hjchen/cwrite-mtc/main.cgi.

UCLA Heritage Language Learner Corpora. Chinese heritage language (HL) learner refers to students with Chinese family backgrounds. They constitute a specific group of CSL learners because they usually have acquired some degree of the Chinese language at a young age and have an advantage in listening to and speaking Chinese. In many institutions Chinese HL learners are placed in a separate track from non-heritage learners. These learners have different needs and follow different developmental paths in Chinese learning. To better our understanding of HL learners, researchers have been calling for "large-scale empirical studies in HL acquisition" (Ming & Tao, 2008, p. 168).

The team at the University of California, Los Angeles, later with contributions from the first author of this chapter, constructed the Chinese Heritage Language (HL) Corpus (ibid.). The corpus has collected written essays by Chinese HL learners at the intermediate level attending elementary heritage Chinese classes in 2006 and 2007. The corpus comprised about 1,000 samples of essays and compositions students wrote as homework assignments, with a total of about 200,000 characters. The text types covered a wide range, from argumentative and narrative, to descriptive. The corpus was POS tagged using the ICTCLAS POS tagger. The pilot corpus was error tagged following a coding system with 10 major categories and 36 subcategories that the team developed specifically for HL learner error annotation (ibid.). This is, to the best of our knowledge, the first Chinese learner corpus built in North America.

Ongoing efforts in CSL learner corpus construction. Since constructing CSL LC, as with LC of many other languages, remains a challenging research topic for researchers in the field, there is no lack of effort in discussing ways to construct LC. After the publication of

perhaps the first explicit methodology paper on CSL learner corpus in Chu and Chen (1993), however, there was a noticeable gap in the following decade, as the next wave of methodological discussions did not appear until the 2000s (e.g., Yang *et al.*, 2006). However, since 2010, there has been a major surge of interest in CSL LC, reflecting what we believe to be a sharp shift of attention to corpora in the field of CSL. Many of the methodological discussions since 2010 have focused on specific ways to improve corpus construction, including processing, annotation, and user interface (e.g., Ren, 2010; R. Zhang, 2013).

B. L. Zhang (2010b, 2016) discussed several challenges of current CSL LC. First is the corpus size and limited representation of learners. There are only a handful of CSL learner corpora, which are relatively smaller in scale compared to LC of other languages, especially English. These corpora mostly include written essays by advanced CSL learners, whereas novice and intermediate learners are underrepresented. The currently available corpora are primarily cross-sectional and therefore unsuitable for studying learner development over time. The corpora are unbalanced, with data mainly from Asian learners (specifically Korean and Japanese); there are far less data from English-speaking regions. Second, the corpora were not built based on the same criteria. Different research teams operate on their own standards regarding corpus size, data collection, topics and text types, data computerization, data storing and annotation, making it extremely difficult to compare findings reported in different studies. The third problem is the limited functions these corpora offer, with some even lacking the basic KWIC search function. Fourth, annotation is not standardized, and many problems exist with the different annotation conventions these corpora use. Last but not least, although more and more corpora are working towards data sharing, most of the current CSL learner corpora only provide limited access to these valuable resources. B. L. Zhang (2016) further highlighted the needs for standardization of CSL learner corpora.

To overcome these problems, the team led by Cui and B. L. Zhang at the Beijing Language and Culture University initiated an effort to construct an International Corpus of Learner Chinese (ICLC) (B. L. Zhang & Cui, 2013). The ICLC aims to include 50 million characters with 45 million written data (25 million raw data and 20 million annotated data) and 5 million spoken data (3 million raw data and 2 million annotated data). It will represent a much wider and more balanced range of learners of L1 backgrounds, geographic locations, Chinese proficiency levels, and learning contexts. It will comprise five sub-corpora: raw corpus, annotated corpus, statistical information corpus, metadata corpus, and Chinese native speaker primary and middle school student corpus. Upon completion the corpus will grant online public access to interested researchers and educators. This global effort of constructing a large-scale, balanced CSL learner corpus signifies a brand new stage of CSL learner corpus construction and will have a profound impact on the scale and rigor of CSL LCR.

Processing and analyzing learner corpus data

With a learner corpus at hand and using a concordance program, one can study learner language on multiple levels, from the patterns of particular words or phrases and the co-occurrence of words, to sets of words that are syntactically or semantically associated. The common techniques in corpus linguistics are frequency list, concordance, and collocation (Kennedy, 1998). Frequency is the token count of a single item (a word, phrase, or structure). Using a concordance program, words in a corpus can be arranged into a frequency list so that comparisons can be made between corpora of different genres, registers and linguistic variations. The frequency of given words can be compared across corpora to determine differences in use. Another useful function of a concordance program is to retrieve concordance lines of all instances containing the word or

structure in focus, which provide contextualized examples of the more typical usage and the less typical usage of a word or structure. When synonymous words are compared, concordance lines may reveal the subtle differences in the linguistic contexts and the meanings of these words. Finally, corpora can be used to retrieve collocation, which is the statistical tendency of co-occurring words. It can indicate pairs of lexical items as well as pairs of lexical and grammatical items. Some commonly used concordance programs are *AntConc 3.4.4* (Anthony, 2015) and *WordSmith Tools* 6.0 (Scott, 2015).

Using the KWIC search function of a concordance program, a wealth of research has been generated on the grammatical and lexical aspects of CSL acquisition covering almost all important aspects of Chinese grammar and vocabulary. As the levels of annotation move beyond the lexical and syntactical levels to the discourse level, LC can be used to investigate CSL learners' use of discourse features, writing development, and idiomatic expressions. Corpora annotated specifically for misused or miswritten characters provide valuable resources for investigating Chinese character acquisition. Phonological studies that were made possible by recent efforts of building spoken LC represent another new direction in CSL corpus research.

Contrastive Interlanguage Analysis (CIA)

The most influential model of LCR is probably the Contrastive Interlanguage Analysis (CIA) developed by Granger (1998a, 1998b, 2002, 2009). CIA makes use of both quantitative and qualitative comparisons of the L1 and the L2, as well as different L2s. The comparisons between NSs and learners aim to uncover features that distinguish learners and NSs. Besides identifying plain errors, LCR can identify the overuse and underuse of linguistic features, revealing the non-native aspects of learner language. The comparisons between learners of different L1s highlight aspects of language use and learner development. By comparing learner corpora covering different variables (age, proficiency level, L1 background, task type, learning setting, and so forth), one can evaluate the effects of these variables on learner language. The interlanguage characteristics can be explained by such factors as L1 transfer, general learner strategies, inter-language development, intra-lingual overgeneralization, input bias, or genre/register influences (Barlow, 2005).

Using the CIA model, Chang (2014a) investigated differences in the use of Chinese lexical items between CSL learners and NSs, and between learners from different L1 backgrounds (Japanese, English, Korean, Vietnamese, Indonesian, and Thai). The native speaker data came from the Academia Sinica Balanced Corpus; the learner data were drawn from the TOCFL Learner Corpus. Using the keyword-keyness analysis, the study first generated a list of 20 most frequent words for each sub-corpus, and then compared the overuse or underuse of these items. The CIA analysis uncovered several salient features in lexical use. For example, compared with learners of other L1 backgrounds, English CSL learners were found to overuse pronouns but underuse sentential final particles. Japanese and Korean CSL learners seemed to overuse '*suoyi*' in expressing a cause-effect relationship. For the Chinese '*if*' sentence, '*ruguo. . .de hua*', English learners tended to use '*ruguo*' while dropping the post form '*dehua*', a pattern similar to Chinese native speakers, whereas Korean and Japanese learners preferred to use the post form '*dehua*' alone or the full pattern. Explanations were offered in terms of L1 transfer, linguistic structures, and cultural influences.

As another example, Xiao and Huang (2013) studied Korean-speaking learners' development of the syntactic complexity of relative clauses. They first looked at learners' performance from novice to advanced proficiency levels. It showed that learners progressed in syntactic complexity with language proficiency, and learners remained relatively stable within a proficiency level.

They then compared learners and NSs. They found that learners progressed towards NSs' performance. The differences between learners and NSs manifested at the intermediate and advanced levels, where the percentage of sentences with one or more relative clauses was higher than NSs. They also compared learners' syntactic complexity against the textbook input. Learner language production was found to be correlated with the type and intensity of textbook input. However, they also found that instead of mirroring textbook input, learners' language development seemed to follow its own development path.

Multi-method approach

A noteworthy improvement of LCR in recent years is the multi-method approach (Gilquin, 2007). As Gilquin (2007) proposed, an investigation of the errors found in a learner corpus "should ideally be complemented by two other types of analyses, namely a comparison of the learner corpus data with native data, which highlights phenomena of overuse or underuse, and elicitation tests, which focus on competence rather than performance" (p. 273). Put plainly, it means learner corpus method needs to be supplemented and verified by experimental elicitation methods. In fact almost ten years before Gilquin's (2007) proposal, Shi (1998) had used the multi-method approach in investigating CSL learners' acquisition order of 22 syntactic structures, in which she combined learner corpus method with an elicitation test, questionnaire survey, and a case study. Recent studies have advanced towards combining the corpus-based method for analyzing learner performance (overuse or underuse) and the experimental method for investigating learner competence. Qu (2013), for example, used the multi-method approach to investigate CSL learners' acquisition of *gei* as a preposition or a verb. She first used the HSK Dynamic Composition Corpus to generate the error types of the *gei* sentences, which revealed that learners omitted *gei* as a preposition. Based on corpus analysis, she hypothesized that learners were not as familiar with *gei* as a preposition as they were with *gei* as a verb. To test her hypothesis, she administered a grammaticality test followed by interviews with selected participants. These studies show that the multi-method approach can more effectively examine both learner performance and competence, and should serve as a rigorous methodological model for future LCR.

Pedagogical implications

When it comes to pedagogical implications, LC can be useful in two ways: the indirect use and the direct use (Leech, 1997). The indirect approach, which is more commonly seen in the LCR literature, refers to the delayed use of LC to guide the writing of textbooks, pedagogical materials, and dictionaries. The rich understandings gained from LCR research, including the acquisition orders and developmental sequences of different linguistic features, the typical errors learners tend to commit at different levels, and the desirable and less desirable L1 effects, should all be taken into account by CSL textbook writers and pedagogical material developers. It provides important information about the selection, description and sequencing of linguistic forms and structures (Granger, 2015). Another area in which CSL corpus research is particularly useful is dictionary compilation. Xiao, Rayson, and McEnery's (2009) corpus-based dictionary of Chinese core vocabulary sets a good model for corpus-informed reference materials development. The *Cambridge Advanced Learner's Dictionary* (2003) used a learner corpus to include information on frequent learner errors. In future we hope to see CSL learner dictionaries with similar information on learner errors together with their statistical distributions based on analyses of CSL LC.

The other use of LC from a pedagogical perspective is the direct use of LC in classrooms by teachers and students. The LC data in this approach could be collected from the same group of students who are to use the data or the teacher could use some general LC for similar instructional purposes. This approach is often referred to as a data-driven learning approach (Johansson, 2009). When learning a word, phrase, or grammatical structure, the teacher can ask the class to search for examples of student use from a native corpus or a learner corpus. Students are guided to examine both the correct and erroneous usages of an expression and form hypotheses about its usage patterns. The data-driven learning approach based on corpus analysis has clear advantages in attention, input, awareness raising, and hypothesis formation, all of which are indispensable factors for successful acquisition of the L2. Admittedly, data-driven learning can be quite time-consuming, and many teachers have doubts about student gains from engaging in such linguistic analysis. Nonetheless, if used in the right way, it can be an alternative to the traditional instructional approaches for vocabulary, grammar, and language use.

There are, in addition, obvious benefits of using learner corpora to inform CSL assessment. Most importantly, it helps establish the benchmarks of students' language proficiency at different levels both in writing and speaking. LC can serve as critical resources by providing quantitative, empirical information that can guide the development of assessment measures, such as placement tests, exit tests, and other types of proficiency assessment.

Future research directions

Since the birth of the first CSL learner corpus in 1995, CSL LC construction and applied research have achieved a great deal in the short course of 20 years. CSL LC constitute an important resource for CSL research, language teaching, and pedagogical material development. With the advances in corpus tools, CSL corpus research is believed to play an even more important role in understanding CSL acquisition. Looking forward, we propose a few suggestions for future research.

The L2 acquisition of phraseology, also called multi-word units, lexical bundles, or formulaic sequences in the SLA literature, is assuming a more central role in LCR. This line of research studies L2 learners' use of linguistic units that are holistically stored and retrieved whole from memory at the time of use (Granger & Paquot, 2008). The new corpus tools enable the analysis of N-grams, i.e., the co-occurrence of words, thus generating a large number of phraseology studies (e.g., Granger & Meunier, 2008; Hasko & Meunier, 2013). Our review of CSL corpus studies suggests a lack of research in this growing area (for a sample research on Chinese lexicon, see B. Zhang, 2008). With the distinctive characteristics of the Chinese language (Tao, 2015), research on phraseology is expected to uncover the unique characteristics of CSL acquisition.

As one may have noticed, the scope of current CSL corpus research is limited to investigating individual linguistic features. It is understandable for researchers to focus on areas in which learners tend to make errors or, in other words, find difficult to learn. However, as the field matures, it does not suffice to only look at discrete linguistic features. Comprehensive assessment of learner language is needed. The field of SLA uses linguistic complexity, accuracy, and fluency (CAF), with each category measured by several indices. The CAF measures of English have seen great development. Lu (2010, 2012), for example, developed computer programs that automatically analyze a dozen indices of syntactic and lexical complexity, which serve as the foundation for targeted instruction. CSL learner development should be measured by similar comprehensive measures. However, due to the unique characteristics of Chinese, tools developed for English cannot be readily applied to analyzing Chinese. It is our hope that Chinese corpus linguists and computational linguists will collaborate to develop ways of automatically assessing

CAF measures for Chinese. Using these tools, CSL learner language development will be researched more effectively.

The status quo of CSL LCR is essentially that of researchers conducting research using the corpora to which they have access. Due to the variations in corpora sizes, learner proficiency levels, contexts of data collection, types of tasks, and so forth, different studies on the same topic can generate very different, sometimes even contradictory, findings. This causes a baffling situation for language teachers and practitioners who want to be informed about the acquisition of a particular structure to guide teaching and material development. To put it differently, although much understanding has been gained in a wide range of linguistic areas using LC, there is no synthesis of research findings, making it difficult to outline a full picture of CSL learners' development. In SLA, meta-analysis has been widely used to synthesize primary research findings of different studies on a similar topic (Norris & Ortega, 2006, 2007). With the anticipated blossoming of CSL corpus studies made possible by the ongoing construction of the ICLC, we believe that a synthesis approach is needed in CSL to organize learner corpus findings so as to provide useful guidance for future CSL learning and teaching. In addition, LCR findings must be verified by other research methodologies, including more controlled experimental methods.

Methodologically, in addition to the multi-method approach, we advocate incorporating more rigorous statistical methods into CSL corpus studies. Current CSL corpus research is mainly based on descriptive statistics of token and type frequencies. While frequency can tell us the overall distribution of a linguistic phenomenon, it does not give enough details about how exactly the phenomenon is distributed in an aggregated data set. CSL corpus linguists should seek support from statisticians in future research. As excellent starting points, Gries (2009) introduces the R program and Lu (2014) provides a useful introduction to some accessible computational tools for LCR.

As CSL corpora of different modes (written, spoken, multi-modal, cross-sectional, and longitudinal) mature, more research is expected across a wider spectrum of learner development. Spoken corpora will make possible investigations of CSL phonological and prosodic development. Multi-modal LC not only document learners' language production but the contexts in which it takes place. They can be used to investigate the simultaneous development of writing and speaking, the effects of input on language use, and the use of non-linguistic cues such as gestures in language acquisition. To date, we have a better understanding about learner development cross-sectionally than we do about their development over time. Studies based on longitudinal corpora that trace the development of individual learners can offer unique perspectives into learner variations.

A related issue is cross-regional analysis of learner language development based on corpora. As is well known, immersion in the target language environment often results in rapid learner language development (Du, 2013; Freed, 1995; Freed, Segalowitz, & Dewey, 2004; Segalowitz & Freed, 2004). In the case of Chinese, however, the target language environment includes a number of Chinese speaking communities: mainland China, Taiwan, Hong Kong, and Singapore, to name a few. Especially interesting is the case of mainland China and Taiwan. Due to the systematic differences between the Mandarin varieties spoken and written in these two regions (Beijing Yuyan Daxue/Zhonghua Yuwen Yanxisuo, 2003), it would be interesting to investigate how learner language development is impacted by regional differences. Chang (2014b, pp. 46–47) provided a comparative analysis of *ba*-sentences based on learner corpora from the mainland and Taiwan. More work along the same lines can be done. Thus, it would be of interest to see if learner language developments in these regions share similar patterns or diverge in some ways with regard to certain linguistic and sociocultural features. Corpus-based cross-regional comparative analyses should be possible given the availability of large-scale learner corpora from institutions on both sides of the Taiwan Strait, such as those from the Beijing

Language and Culture University and the National Taiwan Normal University, which we outlined previously.

Acknowledgements

This article was developed with the support of the grant #P229A140026 from the U.S. Department of Education in connection with the Center for Advanced Language Proficiency Education and Research (CALPER) at the Pennsylvania State University. However, the contents of this article do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal Government. The second author also acknowledges the support of UCLA Academic Senate Faculty Research grants (2014–15, 2015–16) and the National Taiwan Normal University. Thanks are also due to Liping Chang for providing critical references concerning NTNU resources and research, and to Elizabeth Carter for invaluable editorial assistance. All remaining shortcomings are of course our own responsibility.

Additional references

Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen, *Analysing learner language* (pp. 335–357). Oxford & New York: Oxford University Press.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

Zhang, B. [Bo]. (2008). *Jiyu zhongjieyu yuliaoku de Hanyu cihui zhuanti yanjiu* [*An interlanguage corpus-based study on Chinese vocabulary*]. Beijing: Beijing University Press.

Zhang, B. L. (2016). Zai tan Hanyu zhongjieyu yuliaoku de jianshe biaozhun [Standardization of Chinese interlanguage corpora revisited]. *Yuliaoku Yuyanxue* [*Corpus Linguistics*], *3*(1), 21–30.

References

Anthony, L. (2015). *AntConc (version 3.4.4)*. Tokyo, Japan: Waseda University.

Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen, *Analysing learner language* (pp. 335–357). Oxford & New York: Oxford University Press.

Beijing Yuyan Daxue/Zhonghua Yuwen Yanxisuo (2003). *Liang'an xiandai Hanyu changyong cidian* [*Modern Chinese cross-strait dictionary*]. Beijing: Beijing Language and Culture University Press / Taipei: Taipei Language Institute.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam & Philadelphia: John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Cambridge advanced learner's dictionary. (2003). Cambridge: Cambridge University Press.

Chang, L. (2013). TOCFL Zuowen yuliaoku de jianzhi yu yingyong [Construction and applications of the TOCFL Composition Corpus]. In X. Cui & B. Zhang (Eds.), Di'er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji [*Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora*] (pp. 143–154). Beijing: Beijing Language and Culture University Press.

Chang, L. (2014a). Butong muyu Beijing Huayu xuexizhe de yongci tezheng: Yi yuliaoku weiben de yanjiu [Salient linguistic features of Chinese learners with different L1s: A corpus-based study]. *Computational Linguistics and Chinese Language Processing*, *19*(2), 53–72.

Chang, L. (2014b). Huayu xuexizhe jushi shiyong qingkuang fenxi [The usage of some sentence patterns by L2 Chinese learners: A corpus-based study]. *Taiwan Huayu Jiaoxue Yanjiu* [*Taiwan Journal of Chinese as a Second Language*], *8*(2014.06), 41–57.

Chang, L. (2016). TOCFL xuexizhe yuliaoku de pianwu biaoji [Error annotation for the TOCFL learner corpus]. In X. Lin, X. Xiao, & B. Zhang (Eds.), Disanjie Hanyu zhongjie yuliaoku jianshe yu yingyong

guoji xueshu taolunhui lunwen xuanji [*Selected papers from the 3rd International Conference on the Construction and Applications of Chinese Learner Corpora*] (pp. 131–159). Beijing: World Books.

Chu, C., & Chen, X. (1993). Constructing a Chinese Interlanguage Corpus. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], 7(3), 199–205.

Chu, C. [Chengzhi], Chen, X., Zhang, W. X., Zhang, W., Wei, P., & Zhu, Q. (1995). "Hanyu zhongjieyu yuliaoku xitong" yanzhi baogao [*Research report of "The Corpus of Chinese Interlanguage (CCI 1.0)*]. Beijing: Beijing Language and Culture University Press.

Cui, X. (2005). Oumei xuesheng Hanyu jieci xide de tedian ji pianwu fenxi [The acquisition of Chinese prepositions by European and American learners and analysis of their errors]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *19*(3), 83–95.

de Bot, K., Lowie, W., & Verspoor, M. H. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, *97*(1), 131–141.

Freed, B. (Ed.). (1995). *Second language acquisition in a study abroad context*. Amsterdam & Philadelphia: John Benjamins.

Freed, B., Segalowitz, N., & Dewey, D. (2004). Contexts of learning and second language fluency in French: Comparing regular classrooms, study abroad, and intensive domestic programs. *Studies in Second Language Acquisition*, *26*(2), 275–301.

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *ZAA*, *55*(3), 273–291.

Granger, S. (Ed.). (1998a). *Learner English on computer*. London & New York: Addison Wesley Longman.

Granger, S. (1998b). The computerized learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on compute*r (pp. 3–18). London & New York: Addison Wesley Longman.

Granger, S. (2002). Bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538–546.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–32). Amsterdam & Philadelphia: John Benjamins.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 486–510). Cambridge: Cambridge University Press.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (2009). *International Corpus of Learner English v2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam & Philadelphia: John Benjamins.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam & Philadelphia: John Benjamins.

Gries, S. (2009). *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Hasko, V., & Meunier, F. (2013). Capturing second language development through learner corpus analysis. *The Modern Language Journal*, *97*(s1), 1–101.

Huang, W. (2012). Zixing tezheng dui hanzi wenhuaquan zhonggaoji shuiping xuexizhe shuxie hanzi de yingxiang [Effects of formal features on the intermediate and advanced learners from the Chinese character culture zone in writing characters]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *26*(1), 106–114.

Huang, W. (2013). Muyu yinsu dui zhonggaoji shuiping xuexizhe shuxie hanzi de yingxiang – Yi Ri, Han, Ou'mei xuesheng weili [L1 Effects on intermediate- and advanced learners' acquisition of Chinese characters: The cases of Japanese, Korean, European and American learners]. In X. L. Cui & B. L. Zhang (Eds.), Di'er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji [*Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora*] (pp. 289–297). Beijing: Beijing Language and Culture University Press.

Jing-Schmidt, Z. (2011). Gaonianji Hanyu xizuo zhong lingzhidai shiyong de kuawenhua beijing bijiao [Zero anaphora in higher level Chinese writings across learner backgrounds]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *25*(2), 258–267.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 47–66). Amsterdam & Philadelphia: John Benjamins.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London & New York: Longman.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). London: Longman.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208.

Lu, X. (2014). *Computational methods for corpus annotation and analysis*. Dordrecht: Springer.

Ming, T., & Tao, H. (2008). Developing a Chinese heritage language corpus: issues and a preliminary report. In A. W. He & Y. Xiao (Eds.), *Chinese as a heritage language: Fostering rooted world citizenry* (pp. 167–187). Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam & Philadelphia: John Benjamins.

Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam & Philadelphia: John Benjamins.

Norris, J. M., & Ortega, L. (2007). The future of research synthesis in applied linguistics: Beyond art or science. *TESOL Quarterly*, *41*(4), 805–815.

Qu, M. (2013). Jiyu "HSK Dongtai Zuowen Yuliaoku" de "gei" zi ju xide yanjiu [Acquisition of the gei sentences based on the HSK Dynamic Composition Corpus]. In X. L. Cui & B. L. Zhang (Eds.), Di'er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji [*Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora*] (pp. 201–211). Beijing: Beijing Language and Culture University Press.

Ren, H. (2010). Guanyu zhongjieyu yuliaoku jianshe de jidian sikao – yi "HSK Dongtai Zuowen Yuliaoku" wei li [Towards the construction of the Chinese interlanguage corpus – Using the HSK Dynamic Composition Corpus as an example]. *Yuyan Jiaoxue Yu Yanjiu* [*Language Teaching and Research*], *21*(6), 8–15.

Scott, M. (2015). *WordSmith Tools 6.0*. Liverpool: Lexical analysis software.

Segalowitz, N., & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173–199.

Shi, J. (1998). Waiguo liuxuesheng 22 lei xiandai Hanyu jushi de xide shunxu yanjiu [Foreign students' acquisition order of 22 modern Chinese sentence structures]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *46*(4), 77–98.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tao, H. (2015). Profiling the Mandarin spoken vocabulary based on corpora. In S. Wang & C. Sun (Eds.), *The Oxford handbook of Chinese linguistics* (pp. 336–347). Oxford: Oxford University Press.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese: Core vocabulary for learners*. New York: Routledge.

Xiao, X., & Huang, Z. (2013). Hanguo xuesheng zhongjieyu geju changjuzi dingyu fuzadu fazhan yanjiu [Korean students' development of the complexity of relative clauses]. In X. L. Cui & B. L. Zhang (Eds.), *Di'er jie Hanyu zhongjieyu yuliaoku jianshe yu yingyong xueshu taolunhui lunwen xuanji* [*Selected papers from the 2nd International Conference on the Construction and Applications of Chinese Learner Corpora*] (pp. 234–241). Beijing: Beijing Language and Culture University Press.

Yang, D. (2003a). Yingyu muyu xuexizhe quxiang buyu de xide shunxu [Sequence of acquiring the directional complements by English-speaking learners of Chinese]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *17*(2), 52–65.

Yang, D. (2003b). Chaoxianyu muyu xuexizhe quxiang buyu xide qingkuang fenxi [Sequence of acquiring the directional complements by Korean-speaking learners of Chinese]. *Jinan Daxue Huawen Xueyuan Xuebao* [*Journal of the College of Chinese Language and Culture of Jinan University*], *45*(4), 20–31.

Yang, D. (2004). Riyu muyu xuexizhe quxiang buyu xide qingkuang fenxi [Sequence of acquiring the directional complements by Japanese-speaking learners of Chinese]. *Jinan Daxue Huawen Xueyuan Xuebao* [*Journal of the College of Chinese Language and Culture of Jinan University*], *46*(3), 23–35.

Yang, Y., Li, S., Guo, Y., & Tian, Q. (2006). Jianli Hanyu xuexizhe kouyu yuliaoku de jiben shexiang [Tentative ideas on constructing Chinese learner spoken corpus]. *Hanyu Xuexi* [*Chinese Language Learning*], *26*(3), 58–64.

Zhang, B. [Bo]. (2008). *Jiyu zhongjieyu yuliaoku de Hanyu cihui zhuanti yanjiu* [*An interlanguage corpus-based study on Chinese vocabulary*]. Beijing: Beijing University Press.

Zhang, B. L. (2010a). Huibi yu fanhua – Jiyu "HSK Dongtai Zuowen Yuliaoku" de "ba" zi ju xide kaocha [Avoidance and overgeneralization – An investigation of acquisition of the ba-sentence based on the HSK Dynamic Composition Corpus]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *24*(2), 263–278.

Zhang, B. L. (2013). Guanyu tongyongxing Hanyu zhongjieyu yuliaoku biaozhu moshi de zai renshi [Re-considering the models of annotation of all-purpose Chinese interlanguage corpus]. *Shijie Hanyu Jiaoxue* [*Chinese Teaching in the World*], *27*(1), 128–140.

Zhang, B. L. (2016). Zai tan Hanyu zhongjieyu yuliaoku de jianshe biaozhun [Standardization of Chinese interlanguage corpora revisited]. *Yuliaoku Yuyanxue* [*Corpus Linguistics*], *3*(1), 21–30.

Zhang, B. L., & Cui, X. (2013). "Quanqiu Hanyu Zhongjieyu Yuliaoku Jianshe he Yanjiu" de sheji li'nian [Design concepts of "the construction and research of the interlanguage corpus of Chinese from global learners"]. *Yuyan Jiaoxue Yu Yanjiu* [*Language Teaching and Linguistic Studies*], *24*(5), 27–34.

Zhang, J. (2011). Acquisition of the Chinese resultative verb complements by learners of Chinese as a foreign language: a learner corpus approach. Ph.D. dissertation, The Pennsylvania State University.

Zhang, J. (2014). A learner corpus study of L2 lexical development of Chinese resultative verb compounds. *Journal of the Chinese Language Teachers Association*, *49*(3), 1–24.

Zhang, J., & Lu, X. (2013). Variability in Chinese as a foreign language learners' development of the Chinese numeral classifier system. *The Modern Language Journal*, *97*(s1), 46–60.

Zhang, R. (2013). Sange Hanyu zhongjieyu yuliaoku ruogan wenti de bijiao yanjiu [A comparison study on some problems in three Chinese interlanguage corpora]. *Yuyan Wenzi Yingyong* [*Applied Linguistics*], *21*(3), 133–140.