Introduction to Learner Corpora: Construction and Explorations in Chinese and Related Languages

Howard Chen, Keiko Mochizuki, and Hongyin Tao

1 Learner Corpora

As corpus construction and corpus linguistics evolve rapidly to become an indispensable methodology in linguistics and related field over the past several decades (McEnery & Wilson, 1996), research in learner corpus (LC) has gained considerable momentum as an area at the interface of corpus linguistics and applied linguistics (Granger et al., 2015). A learner corpus is a computerized collection of (inter)language samples produced by learners of a second[1]/foreign language, hence the term "computer learner corpus" in the early phase of learner corpus development (Granger, 1998; Leech, 1998). The catalyst of learner corpus research development is undoubtedly the International Corpus of Learner English (ICLE), which took off in Belgium in the 1990s (Granger, 1998) and which now has seen its third iteration (Granger et al., 2020). In the field of Chinese as a second or foreign language, LC is generally known as 學習者語料庫 "learner corpus" or as 中介語語料庫 "interlanguage corpus" (Zhang & Tao, 2018). LC is one of the few areas in corpus linguistics where, in our estimation, Chinese language studies have kept pace with international field development. For example, from 1993 to 1995, the research team at the Beijing

[1] The term "second language" is used to cover all non-first language acquisition.

H. Chen (✉)
Department of English, National Taiwan Normal University, 162, HePing East Road, Section 1, 106 Taipei, Taiwan
e-mail: hjchenntnu@gmail.com

K. Mochizuki
Graduate School of Global Studies, Tokyo University of Foreign Studies, Tokyo 183-8534, Japan
e-mail: mkeiko@tufs.ac.jp

H. Tao
Department of Asian Languages and Cultures, University of California, Los Angeles, CA 90095, USA

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_1

Language and Culture University constructed the first CSL learner corpus, the L2 Chinese Interlanguage Corpus ((汉语中介语语料库系统, Chu et al., 1995), with essays composed by learners from multiple countries on the HSK test. This influential corpus, too, is now undergoing major development for a second edition (Zhang, this volume). Elsewhere in Chinese-speaking communities, LC development has also seen impressive achievements. For example, the Spoken Learner Corpus (華語為第二語口語語料庫), developed at the National Taiwan Normal University and sponsored by the ROC government, is one of the few collections of spoken learner data (Chang, 2016). However, efforts in the Chinese learner corpus have gone far beyond the confine of the greater China region. Learner corpora of both heritage (Ming & Tao, 2008) and non-heritage learners have been developed in the US (Jin, Zhang, and Tao, this volume) and Japan (Sano et al., this volume), among others. (For a recent review of LC in Chinese, see Zhang & Tao, 2018.)

From the very beginning, researchers have identified some of the key areas that make LC both unique and challenging (Granger, 1998). First of all, learner language or interlanguage is diverse as it is heavily influenced by the learner's first language and the proficiency level of the learner in the target language. This poses challenges that are not common in native speaker language corpus design and collection. Second, in terms of processing, learner language is characterized by errors and irregularities, which present issues in processing and annotation that are again very different from dealing with first language corpora, where errors are usually not the focus. Third, learner language analysis requires special care and treatment. In the area of statistical analysis, for example, what needs to be counted and in what ranges of texts in learner performance data one is to count the data can have important implications for the validity of the analysis (Gries, 2015, 2021). Lastly, learner language research is inherently contrastive and comparative, due to the crosslinguistic nature of adult second language learning. This gives rise to the so-called Contrastive Interlanguage Analysis methodology (Granger, 2015b).

At the same time, learner corpora have proven to be useful at multiple levels. In addition to identifying learner error tokens, types, and their related frequency information (Leech, 2011), as well as interlanguage developmental paths, LC can be used to reveal features of the first language in comparison with the second language. LC can also be used to address pedagogical needs in terms of reference and instructional materials design (Granger, 2015a) and classroom activities. More recent applications of LC have been extended to natural language processing (NLP), where computational systems are designed for automated scoring and automatic error detection and correction (Granger et al., 2015:3).

This edited volume[2] attempts to take stock of some of the major undertakings of the Chinese learner corpus and reflects the state of the art in corpus and related approaches

[2] The project grew out of two conferences and the related research projects: the 6th international Workshop on Advanced Learning Sciences hosted by the University of Pittsburgh, in 2018 (http://www.iwals2018.pitt.edu/) and the International Symposium on Diverse Approaches to Second Language Acquisition: Learner Corpora, Evaluation and Brain Sciences (http://www.tufs.ac.jp/ts/personal/mkeiko/project/) at Tokyo University of Foreign Studies (TUFS) in 2019. The TUFS project was supported by JSPS KAKENHI Grant JP17H02357 "Research on cross-referential

to Chinese as a second language (CSL). CSL as a field has flourished in the past few decades due to the increasingly important role of the Chinese language on the world stage; yet studies of CSL based on learner corpora have been less well developed due to the limited availability of sharable data as well as the underdevelopment in the theoretical front. This volume aims to represent the latest research in this area by (1) assembling a large group of active researchers from multiple international research communities (US, China, Hong Kong, Macau, Japan, Taiwan, and France); (2) discussing the latest resources and technologies in Chinese learner corpus and corpus building; (3) basing CSL studies on data from learners of Chinese with a wide range of first language backgrounds (English, Japanese, Korean, Thai, Vietnamese, and French); and (4) integrating corpus methods with a wide range of related methods in allied fields—language acquisition, usage-based linguistics, psycholinguistics, and neurolinguistics, among others.

2 This Collection

The volume is divided into three broad categories: (1) Chinese learner corpus construction; (2) explorations in learner corpora in Chinese and related languages; (3) typological and comparative approaches to L2 Chinese and related languages. A summary of the papers in each section is provided below.

Part I: Learner Corpus Construction and Processing

This part contains four papers. In the first, Zhang Baolin, the main architect of the influential Learner Corpus of Chinese (LCC) developed at the Beijing Language and Culture University, describes the designing principles of the updated and expanded version of LCC. Three main issues are addressed in the paper: (1) network security for open access and stable operations; (2) improved functionality meeting the needs of a wide range of users; (3) a user-friendly web interface for ease of use for all types of users (specialists and students alike).

Weiping Wu draws our attention to pragmatic issues in the construction and exploitations of LC, especially in terms of contexts of learners' understanding and acquisition and application of pragmatic knowledge in oral communication, which have rarely been dealt with in the Chinese LC field. He draws on data from the Language Acquisition Corpus constructed with oral productions by CSL learners of various language and cultural backgrounds.

Ting-Yu Yang, Hui-Mei Yang, Wei-Jei Lee, Chen-Yu Liu, and Howard Hao-Jan Chen introduce the construction of the Chinese Learner Written Corpus at the National Taiwan Normal University (NTNU), an error-tagged two-million-word learner corpus with 119 error tags. Via retrieval and detailed analysis of the various

learners' corpora of English, Chinese and Japanese though international educational collaboration at secondary and tertiary levels".

error tags, they uncovered that the top 12 types of common error in the learner corpus accounted for more than 50% of the total number of errors.

Part II: Explorations in Learner Corpora in Chinese and Related Languages

The majority of the papers are in the second category, which explores features of Chinese interlanguages based on LC and from various perspectives.

In their paper titled Cross-referentiality of Multilingual Error Learner Corpora of Chinese, English and Japanese for Second Language Acquisition of Chinese Grammar, Hiroshi Sano, Yeong-il Yi, Chia-Hou Wu, Go Inoue, YaMing Shen, Noboru Oyanagi, and Keiko Mochizuki present a Japanese L1 learner corpus of Chinese, https://corpus.icjs.jp/ with discussions on methods of collecting, (error) tagging, and annotating of learner data. The effects of L1 on learners' acquisition of Chinese grammatical items such as auxiliaries, resultative complements, and determiners are presented.

Hong Gang Jin, Jie Zhang, and Hongyin Tao present a comparative corpus-based study on L1 and L2 verb complement constructions of manner and states (VCM/S), which is a rather unique formulaic sequence in Chinese. This chapter makes use of the existing L1 and L2 Chinese corpus data and finds that there are marked quantitative and qualitative differences between L1 and L2 VCM/S production at both construction and component levels.

In an investigation of the acquisition of relative clauses (RCs) in Chinese based on the Test of Chinese as a Foreign Language (TOCFL) Learner Corpus, Liping Chang shows that, regardless of learners' language background, object-extracted RCs (ORCs) are easier for Chinese L2 learners to acquire than subject-extracted RCs (SRCs), and she proposes that word order plays a key role in learning Chinese RCs.

Jia-Fei Hong, Hsin-Tzu Jen, and Yao-Ting Sung examine data in the Chinese Written Corpus to uncover Chinese L2 learners' error patterns in writing. The data was generated by learners from varied language backgrounds and proficiency levels. Their analysis of the data shows that misformation is the most common structural error type, which is attributed to learners' difficulty in the use of adverbs.

Ting-Yu Yang, Hui-Mei Yang, Wei-Jei Lee, Chen-Yu Liu, and Howard Hao-Jan Chen examine the phenomenon of omission of the adverb 都 *dou* "all, completely" in the Chinese Learner Written Corpus of NTNU. Their analysis shows that omission of *dou* often occurs when it serves as a scope adverb to quantify noun phrases that include elements such as 每 *mei*, 所有的 *suoyoude*, 任何 *renhe*, 隨時 *suishi*, and 到處 *daochu*, while omissions occur less often when 都 *dou* serves as a modal particle or a time adverb.

In their paper titled Acquisition of the Chinese Indefinite Determiner "One + Classifier" and English Articles in Two-way Learner Corpora, Zhang Zheng, Laurence Newbery-Payton, and Sho Fukuda reveal a number of patterns. First, English L1 learners of Chinese overuse the "one + classifier" structure for indefinite reference, analogous to English indefinite articles, whereas Japanese L1 learners show underuse of this structure, despite Chinese and Japanese both being regarded as "classifier languages". Second, data from the TUFS Learners' Corpus of English reveals that

Chinese L1 learners use the definite article in a more native-like way than Japanese L1 learners. Finally, Chinese L1 learners of Japanese use the "one + classifier" structure more frequently than native speakers.

Keiko Mochizuki and Yasuhiro Shirai explore the acquisition of telic forms in Chinese and Japanese grammar based on TUFS co-referential learner corpora of Chinese and Japanese. They show that Japanese learners have difficulties acquiring resultative compound verbs expressing telicity and the atelic auxiliary verb "huì". On the other hand, Chinese learners have difficulties acquiring aspectual compound verbs (e.g. inchoative -kakeru/kakaru, -dasu, and perfective -ageru/-agaru), and overuse of resultative intransitive verbs in transitive/intransitive pairs in Japanese. They contend that these difficulties in learning telicity are due to a typological difference in cognition: Chinese is a "bounded-cognition prominent" type language while Japanese is an "unbounded-cognition prominent" type language. They also explore effective pedagogy based on learner's native languages.

In her paper on the (non-)acquisition of the Chinese Definiteness Effect: A usage-based account, Ludovica Lena investigates the acquisition by French L1 learners of Chinese the Definiteness Effect (DE) that characterizes Chinese existential-presentational construction (EPC). This paper is based on elicited oral productions of 15 French advanced learners of L2 Chinese. Both the use and non-use of EPCs are analyzed and the patterns are accounted for in terms of referent-introducing and subsequent tracking contexts.

Part III: Typological and Comparative Approaches to L2 Chinese and Related Languages

Three papers fall in this category. In an investigation of modal verbs such as 会 huì, 要 yào, and 能 néng, Zhang Zheng, Sho Fukuda, Laurence Newbery-Payton, Tomohito Ishida, and YaMing Shen reveal that the influence of L1 on L2 is observed in the written production of Japanese and English learners of Chinese. Further evidence for the influence of L1 on L2 modal verb use is observed in the written production of Chinese and Japanese learners of English. Taken together, the data provides evidence showing typical difficulties for L1 speakers of Chinese, English, and Japanese learning each other's languages.

Kumiko Sakoda, a leader of the International Corpus of Japanese as a Second Language (I-JAS) project (https://chunagon.ninjal.ac.jp/static/ijas/about.html), collects data from a total of 1,050 research subjects from 12 different native languages: English, Chinese, Korean, German, French, Vietnamese, Russian, Spanish, Indonesian, Hungarian, Turkish, and Thai. The paper is mainly concerned with request expressions. One of the main findings is that "suspended (or incomplete) clauses", such as "I have a favor to ask you, but …" which are frequently used by Japanese native speakers, are rarely used by those learners. Second, "the confirmation expressions" ("is it OK?") were observed more frequently among Chinese speakers compared with the other speakers, and it can be considered a negative transfer from learners' native language, Chinese.

Finally, Haining Cui, Hyeonjeong Jeong, Yoshihiro Mochizuki, and Keiko Mochizuki focus on how learner's native language typology affects the second language acquisition of word order and morphology with evidence from Chinese learner corpora by L1 English and Japanese and brain science. First, word order typology, SVO or SOV affects the acquisition of Chinese "Verb + Complement" resultative compound verbs, since word structure reflects syntactic word order. English L1 learners are easy to acquire the Chinese resultative compound verbs, because English has the same SVO and similar SOVC resultative construction. On the other hand, for Japanese L1 learners of Chinese, even advanced learners find it quite difficult to acquire the resultative compound verbs, since the Japanese word order is SOV, there is no SOVC resultative construction. This word order typology affects the acquisition of lexical aspects by Japanese and Korean L1 learners.

We hope that this collection will spur further interest in learner corpora. We also hope that it can help the field-building efforts in LC, especially in terms of the three themes addressed in this volume: corpus construction and data processing; explorations in patterns of learner interlanguage; and crosslinguistic and interdisciplinary investigations.

References

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, Chengzhi, Chen, X., Zhang, W., Zhang, W., Wei, P., Zhang, W., & 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.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Grnnger (Ed.), *Learner English on Computer. Edited by Sylviane Granger*, Longman London and New York, pp. 3–18.

Granger, S. (2015a). The contribution of learner corpora to reference and instructional materials design. *In Sylviane Granger, in Granger, Gilquin and Meunier Eds., 2015*, 485–510.

Granger, S. (2015b). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research, 1*(1), 7–24.

Granger, S., Gilquin, G., & Meunier, F. (2015). *Introduction: Learner corpus research—past, present and future* (pp. 1–5). Cambridge University Press.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020). The International Corpus of Learner English. Version 3. Louvainla-Neuve: Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877.

Gries, S. Th. (2015). Statistics for learner corpus research. S. Th. Gries, in Granger, Gilquin and Meunier 2015, 159–181.

Gries, S. T. (2021). A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics, 9*(2), 1–33. https://doi.org/10.32714/ricl.09.02.02

Leech, G. (1998). *Preface to learner English on computer*. Edited by Sylviane Granger, Longman London and New York, xiv–xx.

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), A Taste for Corpora: In honour of Sylviane Granger. 7–31. John Benjamins Publishing Company Amsterdam/Philadelphia.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press.

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). National Foreign Language Resource Center, University of Hawai'i.

Zhang, J., & Tao, H. (2018). Corpus-based research in Chinese as a second language. In C. Ke (Ed.), *The Routledge Handbook of Chinese Second Language Acquisition* (pp. 48–62). Routledge.