Chinese Language Learning Sciences

Howard Hao-Jan Chen
Keiko Mochizuki
Hongyin Tao *Editors*

Learner Corpora: Construction and Explorations in Chinese and Related Languages

Springer

Chinese Language Learning Sciences

This book series investigates several critical issues embedded in fundamental, technical, and applied research in the field of Chinese as second language (CSL) learning and teaching, including learning mechanism in the brain, technology application for teaching, learning and assessment. The book series discusses these issues from the perspectives of science (evidence-based approach) and technology. The studies in the book series use the methods from the fields of linguistics (such as corpus linguistics and computational linguistics), psychological and behavioural sciences (such as experimental design and statistical analyses), informational technology (such as information retrieval and natural language processing) and brain sciences (such as neuroimaging and neurolinguistics). The book series generally covers three main interdisciplinary themes: (1) fundamental investigation of Chinese as a first or second language acquisition, (2) development in Chinese language learning technology, and (3) applied research on Chinese language education.

More specifically, the book series involves seven research topics:

- language transfer mechanism in Chinese as a second language
- factors of Chinese as a second language acquisition in childhood
- cultural influence on Chinese acquisition
- information technology, corpus
- teaching material design
- teaching strategies and teacher training
- learning models
- assessment methods

All proposals will be sent out for external double-blind review. Review reports will be shared with proposers and their revisions will be further taken into consideration. The completed manuscript will be reviewed by the Series Editors and editorial advisors to ensure the quality of the book and also seek external review in order to ensure quality before formal publication.

Please contact Melody Zhang (e-mail: melodymiao.zhang@springer.com) for submitting book proposals for this series.

Howard Hao-Jan Chen · Keiko Mochizuki ·
Hongyin Tao
Editors

Learner Corpora: Construction and Explorations in Chinese and Related Languages

Springer

Editors
Howard Hao-Jan Chen
Department of English
National Taiwan Normal University
Taipei, Taiwan

Hongyin Tao
Department of Asian Languages
and Cultures
University of California
Los Angeles, CA, USA

Keiko Mochizuki
Graduate School of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan

ISSN 2520-1719 ISSN 2520-1727 (electronic)
Chinese Language Learning Sciences
ISBN 978-981-19-5730-7 ISBN 978-981-19-5731-4 (eBook)
https://doi.org/10.1007/978-981-19-5731-4

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Contents

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.

Learner Corpus Construction and Processing

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

Baolin Zhang

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

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

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

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

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

1 Introduction

1.1 The Development of Corpus Construction and Its Applied Research

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

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

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

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

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

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

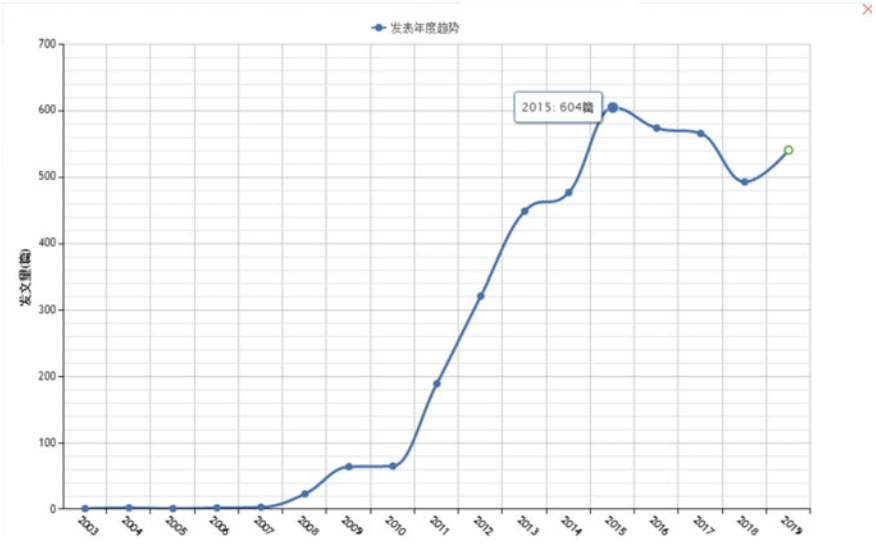


Fig. 1 Annual number of research papers based on HSK

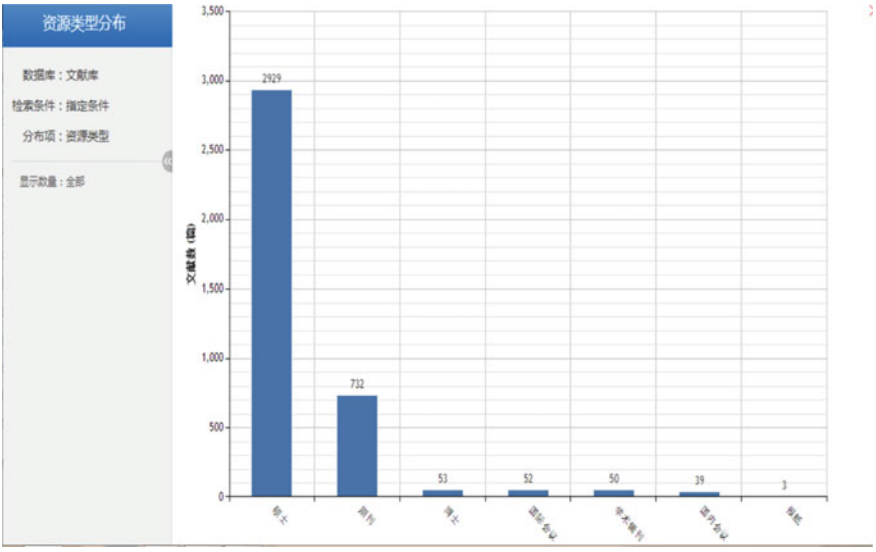


Fig. 2 Distribution of research resource types based on HSK

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

1.2 Ontological Study of Chinese Interlanguage Corpus Construction

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

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

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

Construction standards are another aspect of corpus construction that requires attention. Research in this area is a response to the lack of unified standards in the

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

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

1.3 Existing Problems

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

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

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

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

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

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

1.4 Solutions

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

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

2 Design Principles

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

2.1 Aim

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

2.2 Principles

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

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

The second consideration that came into play when redesigning the HSK corpus was the need to ensure that the functions of the corpus were computationally powerful enough to meet user demands. The HSK corpus version 1.0 and version 1.1 are products of the 1.0 generation, which were built during the initial period of Chinese interlanguage corpora construction (cf. Zhang, 2019). These two versions embody

defining characteristics of that time in that they are simple and large scale, not fully functional, and have difficulty satisfying users' demands in many aspects. For example, in these earlier versions, it is possible to search for the usage of the separable word "合" *He*, but not for the usage of "离" *Li*. One can search for some sentence constructions with marker words such as the "把" *Ba* construction and the "比" *Bi* construction, but not for sentences with the "是……的" *Shi……De* construction or the "连……也/都……" *Lian……Ye/Dou……* construction, because these sentence constructions require two search terms.

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

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

3 Functional Design

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

3.1 Search

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

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

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

3.1.1 General Search of Strings

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

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

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

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

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

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

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

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

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

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

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

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

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

3.1.2 Sentence and Text Search

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

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

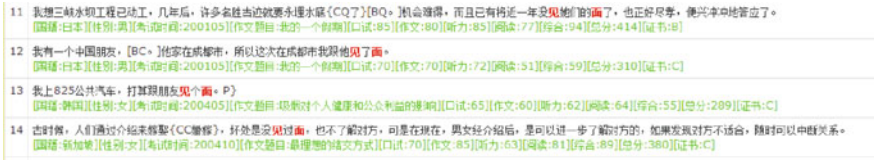


Fig. 8 Results produced searching for the separable words “见面”

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

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

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

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

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

Fig. 10 Results produced searching for the error text

3.1.3 Advanced Search

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

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

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

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

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

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

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

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

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

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

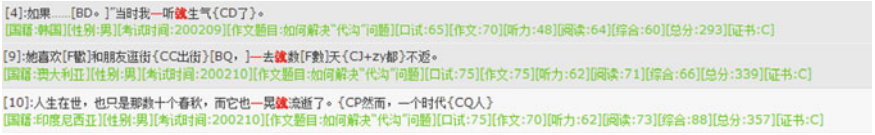


Fig. 14 Results produced searching for the “一……就……” construction

Fig. 15 Search results generated for complex sentence retrieval

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

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

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

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

Details can be seen as follows.

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

3.2 Corpus Presentation

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

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

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

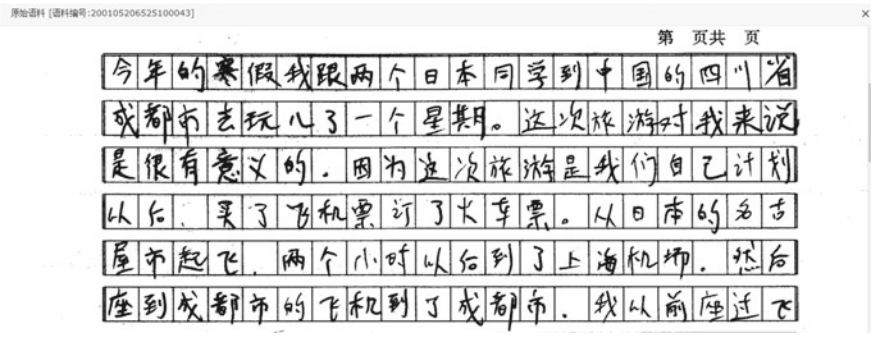


Fig. 18 An original composition in the corpus

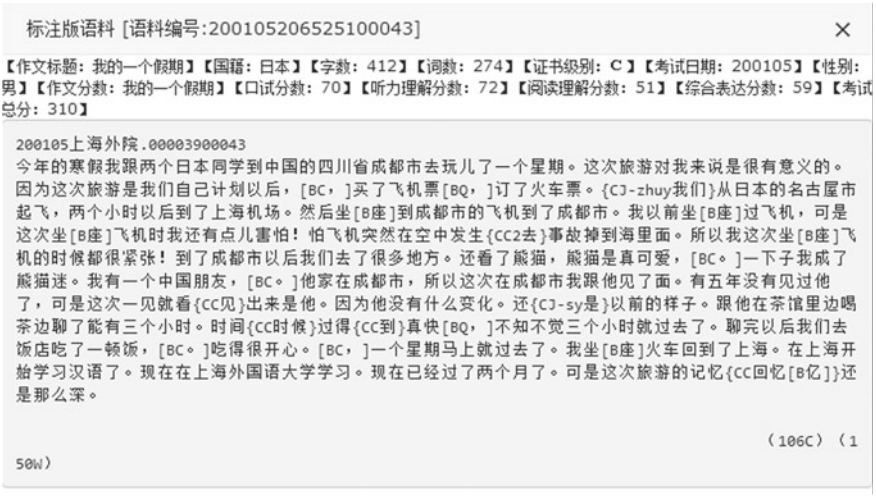


Fig. 19 Annotated full text in the corpus

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

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

3.3 Statistical Analysis

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

Examples of statistical charts are listed as follows.

Fig. 21 An overview of the corpus

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

Fig. 23 Characters information sorted by year

3.4 Others Functions

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

(1) Crowdsourcing maintenance

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

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

(2) Feedback

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

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

(3) Personal workspace

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

Fig. 24 Sample feedback message

(4) Automatic language data download

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

(5) Increase and accumulation of the relevant resources

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

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

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

Fig. 25 Statistics example graph

4 Conclusions

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

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

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

Some Pragmatic Issues in Learner Corpus: A CSL Perspective

Weiping Wu

Abstract Unlike elements in language structure (phonology, semantics, and syntax), factors related to language use are much more difficult to handle and are often neglected, or simply ignored, in the construction of a corpus. For example, how to design the tasks and prompts while gathering oral samples so that pragmatic factors become an integrated part of the data collected? Instead of treating pragmatic issues as some "extra elements" to be identified after the samples are collected while building the corpus, the author presents a systematic approach in which pragmatic factors are treated as part of the design before the construction of the corpus. Both theoretical framework and specific steps taken in the implementation are discussed in this paper in the context of understanding and using pragmatic knowledge in oral communication. All examples used are from the Language Acquisition Corpus constructed with oral productions by CSL learners of various language and cultural backgrounds.

Keywords Learner corpus · Corpus construction · Pragmatic factors · CSL learning

1 Introduction

This paper explores the role of pragmatic factors in the construction of corpus, focusing on how a systematic approach can be applied in obtaining oral productions of CSL learners from different language and cultural backgrounds and how to organize the data obtained in the corpus. Because of the availability of pragmatic clues associated with the data, such a corpus can then provide opportunities for studies related to L2 production beyond the structure of the language.

To provide a larger context in which we discuss this CSL learner corpus, let's take a closer look at this area of linguistic research. In terms of language, the majority

W. Wu (✉)
Center for Linguistics and Applied Linguistics (CLAL), Guangdong University of Foreign Studies (GDUFS), Guangzhou, China
e-mail: 202070007@oamail.gdufs.edu.cn

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_3

of the corpora and related studies so far are predominantly centered around English, including some of the most widely used online corpora, such as the Global Web-based English, the Corpus of Contemporary American English, Corpus of Historical American English, the TV Corpus, the Movie Corpus, and the British National Corpus. (cf. https://www.english-corpora.org/). When it comes to learner corpus, studies reported in the Journal of Learner Corpus Research, among others, can provide us with a glimpse of what is going on in this field, especially the special issue in which the editors (Brezina & Flowerdew, 2019) put together some of the impressive studies related to the Trinity Lancaster Corpus.

Studies related to the Chinese language, on the other hand, are still few and far between compared with what has been achieved in English, even though rapid progress can be seen in recent years. Among some of the popular ones are various corpora as listed online (cf. https://www.cncorpus.org), those maintained by the Academic Sonica in Taiwan (cf. http://www.sinica.edu.tw/SinicaCorpus/), as well as others that seem to focus on specific areas of language use, like the MLC (by the Chinese University of Communication, cf. http://ling.cuc.edu.cn/RawPub/). Due to the availability of data from large-scale proficiency tests in the past decades (e.g., the HSK, which is a proficiency test taken by hundreds and thousands of CSL learners from all over the world who want to enter Chinese language programs in universities in China), learner corpus for CSL has been developing very quickly (Chen & Tao, 2019; Tao, 2017; Tao et al., 2020; Zhang & Tao, 2018; Zhang et al., 2019). Other learner corpora similar in nature include the BCC Corpus by Beijing Language and Culture University and the CCL Corpus by Beijing University. The Language Acquisition Corpus focusing on Spoken Chinese (a.k.a. LAC/SC) to be discussed below is unique because of the availability of information related to pragmatic factors for each oral sample in the corpus.

In reference to the corpora mentioned above, the CSL learner corpus based on oral production is still at the very initial stage of its development path compared with those based on data from ESL. The LAC/SC now provides direct access to the original sound files for each of the oral productions as well as a clean version of the written transcription in Chinese characters. It is hoped that such a model with built-in pragmatic factors can contribute to narrowing the gap between ESL and CSL studies based on corpora.

The discussion below will be divided into four parts, each of which is briefly described here to provide an overall picture. The next part will explain two concepts behind the construction of the LAC/SC, one being the distinction between language structure (LS) and language use (LU), the other, whether the final goal of all learning activities is "appropriate culturally" or just "correct structurally". The third part of this paper describes the structure of the LAC/SC, including what we mean by pragmatic factors and how they are identified and dealt with in the process of corpus construction. Problems met, and possible solutions applied, are discussed in the fourth part, covering data eliciting procedure, task design, and measures taken to guarantee adequate comprehension of tasks by L2 learners. The final part of this paper offers some concluding remarks, representing our current understanding of creating and implementing the pragmatic framework in building a CSL spoken corpus.

2 Two Fundamental Concepts Behind the CSL Learner Corpus

Understanding the concepts behind the two distinctions discussed here is key to understanding the logic and reasoning behind the construction of the LAC/SC. In the distinction we make between language structure (LS) versus language use (LU) in L2 teaching and learning, we propose that LU be viewed as a system of systems, consisting of three key components: Interlocutors, Setting, and the Timing of the communication event (or Purpose if clues for timing are not available). Each of these components can be further divided into sub-categories. We argue that such a view is comparable to the way we view LS, which is also a system of systems consisting of phonology (Sound), semantics (Words), and syntax (Grammar). For the convenience of discussion, we will use the following abbreviations and equations to represent these two systems:

$$\mathrm{LU} = \mathrm{I} + \mathrm{S} + \mathrm{T/P}$$

$$\mathrm{LS} = \mathrm{Pho} + \mathrm{Sem} + \mathrm{Syn}$$

In the second distinction we make between being "structurally correct" versus "culturally appropriate" in reference to the final goal of language learning, we believe that all CSL learners' production for communication purposes should be the latter and not the former. That means one step is missing between the final goal and most of our current curricula and teaching practices, most of which seem to stop when students "understand" what is being taught and their productions are correct in terms of LS.

It is not surprising to find such a reality in the CSL field because, in various subfields under the general heading of language teaching and learning, the focus of attention has been overwhelmingly on the structure of language (Chao, 1968; Lado, 1957; Wang, 2010; Wu, 1993). In recent years, we started to hear calls for attention to pragmatic ability in discussions related to CSL teaching, second language acquisition, and pedagogy (Ran, 2004; Rose & Kasper, 2001; Wang, 2006; Wu, 2006, 2016). Common sense would tell us that people call for attention means there is a lack of attention. As pointed out by Li in a recent interview (Li, 2021), most of the attention in linguistic studies in China was on the research on language structure and, by comparison, neglecting the real situations in actual language use. Teaching materials preparation with various vocabulary lists, grammar points, and sentence pattern lists can serve as typical examples of such focused attention. For many years, teaching activities within any language learning program tend to center around the explanation of grammar points, which also indicates that the point of attention is on language structure. Various tests in the teaching and learning process, such as the common practice of a "quiz" on grammar after each lesson, as well as many proficiency tests (Clark & Li, 1986; Ke, 1994; Li, 1997; Liu, 2008; Xiong et al.,

2002) that are not supposed to be closely related with any particular curriculum, are often designed with three key components of the LS: pronunciation, vocabulary, and grammar.

Corpus construction in recent years, similarly, follows the same general direction, with tagging of grammatical categories and errors based on deviation from standard pronunciation and grammar rules. This is certainly understandable because research on language structure has been long and many. Moreover, the basic structures of any language are always the starting point of a learning program. How can CSL learners use Chinese if they don't know the pronunciation of a word, what it means, and how to use relevant grammatical rules to put word strings together when they speak?

Once we are out of the classroom and out of the school, once L2 learners get into real communication in real life with real people for meaningful exchange of ideas, however, problems arise. When we come face to face with scenarios in our daily life, we realize what we need to have meaningful and smooth communication is way beyond the knowledge of language structure. We have to consider and remember, unlike native speakers who usually do that without thinking, who we are and to whom we are talking, where we are, and why we are talking at that particular moment. These are the basic elements of communication. Proper understanding and application of such elements will contribute to the communicative ability of the language user. Careful analysis of any communicative event tells us that issues related to these can be grouped into categories, which can be related to LS but are not part of the LS. It reiterates the points we made above, and somewhere else (Wu, 2006, 2008a, b, 2019, 2020), that Pragmatic Factors are what native speakers can intuitively make use of when they talk, but L2 learners cannot due to the lack of such intuition. So telling CSL learners what these factors are is a duty that teachers cannot avoid.

Now let's return to the final goal for all L2 learners, being appropriate culturally versus being correct structurally. Obviously, the former must include the latter but not vice versa. To use a metaphor here, where is the finishing line in the school language program if we treat all L2 learners as athletes participating in the marathon? Although no one would openly deny that all our language teaching activities should aim at the application of knowledge for real communication, and not just "finishing the teaching tasks" as required by the curriculum, it is also hard for any of us to deny that the reality in most cases is still "doing the teaching job" as required by the curriculum, which is unfortunately still largely if not totally based on structure. To go the extra mile from being correct to being appropriate requires too much extra efforts and too much resources.

As a result, it is not uncommon to see CSL learners at the higher end of the proficiency level produce utterances that are correct in terms of pronunciation and grammar, but culturally not appropriate in real communication with real people, thus failing the very purpose of communication. There are many examples from the data collected for the LAC/SC to illustrate this. On the discourse level, the absence of a formal greeting to show respect in a formal setting, for instance, is a case in point. As cited in a research based on the corpus (Fan, 2018), out of 15 oral productions

by advanced CSL learners in a formal setting, 10 of them (2/3) did not start properly when they made a speech as a representative on behalf of a delegation, here is one of them:

e.g.1 (Note: First utterance of the speech, absence of any greeting)
我們-也-知道-我們-來-到-這邊-會-麻煩-你們-啊 ……
Women-ye-zhidao-women-lai-dao-zhebian-hui-mafan-nimen-ah
We-also-know-we-come-particle-here-will-bother-you-particle……
We also knew that we would bother you when we came here

(LAC/SC sample id: Kw0129-SS008)

This is of course a very polite way of saying things, but certainly, it should not be the first utterance when you start talking! More examples of similar nature and relevant discussions along this line can be found in the research reported in a Ph.D. dissertation based on the LAC/SC (Fan, 2018).

3 Pragmatic Framework and Its Application in Corpus Construction

How to implement pragmatic factors in the construction of the corpus? Earlier, we have identified the three essential categories (I, S, and T/P) under LU, which jointly contribute to the appropriateness of oral production by CSL learners. We recognize that not every communicative event has obvious clues to these categories. For example, clues for the timing of the communication, or timing as a factor, are sometimes missing if it is not crucial in that particular communication event. In such a case, P (purpose) can often be used alternatively to fill in the gap. Findings from sociolinguistic research tell us that appropriateness in communication by native speakers is not by chance, but by the speaker's thorough understanding of these essential factors in communication and the social rules, most of which are oblivious to L2 learners. As teachers, we need to tell our students, like what we tell them in their learning process about the structure of the language, and let them know what these elements are. One way to do this, as we did in the construction of the LAC/SC, is to make explicit relevant information about all the three categories, which in most cases tend to be "understood or inferred" by native speakers.

Like grammar rules governing sentence formation and word selection in LS, there are rules governing the choices we make in LU, including pronunciation, vocabulary, and grammar. To make available clues for these three categories for CSL learners will therefore help them make the right choice like native speakers. The use of the polite form "nin" in Chinese instead of "ni", for example, is governed by the LU rule related to the I category. It is a two-way distinction in which the polite form "nin" is used when the speaker knows he is "talking up". We use L→H in the Pragmatic Framework to indicate such a relationship, which can be further clarified as follows:

Relationship among Interlocutors:

We can use L to stand for Low, and H for High, as an indication of status. For CSL learners, we can introduce three types of relationships, among friends (L→L or H→H, in which both parties are equal), from subordinate to superior (L→H), or vice versa (H→L), in which the speaker should be aware if he is talking up or talking down, and choose the polite form in the former case. In the Chinese culture, there are some common examples in the L→H relationship:

Age: to someone of your parent's or grandparent's age.

Social status: to someone with a higher social status, such as your boss, your teacher, someone with a higher rank in the official or social hierarchy, etc.

In the S category, we can introduce a three-way distinction: informal, formal, and ceremonial or ritual. To borrow the example from another study (Feng, 2018), the choice of words in each situation may differ even if the meaning to be expressed is the same, as indicated below:

Informal: use "pian", as in "pianren (騙人)"

Formal: use "qipian", as in "qipian laoshi ren (欺騙老實人)"

Ceremonial: use "qi", as in "chengbuqiwo (誠不欺我)"

In the T/P category, the situation is a bit subtle in comparison to the other two categories, and harder for CSL learners to grasp. We can call it a two-way distinction because, in contrast to "the right moment", there is the "wrong moment". Even if you observe rules on interlocutors and settings, what you say may still be inappropriate if you choose the wrong moment to talk. If you want to propose a toast at the dinner table, for example, you will have to wait for your turn. Doing it too early or too late may render your toast inappropriate in the Chinese culture, no matter how polite you may be. This is perhaps most difficult for CSL learners because it is not something we can spell out for them as we do in the I and S categories. Being able to do this requires the knowledge and skill that even native speakers are not sure of from time to time. This is an area that is waiting to be explored and, before we can identify and find a way to explain and label what we "feel", the best way of doing things at the present is to raise a flag in the mind of all CSL learners, with the hope that such a flag will help them understand what may go wrong in their communication.

In each of the three categories given above, there are of course many more layers in each of the sub-categories. The Setting category may have varieties in different situations, such as semi-formal between informal and formal. To make it easy for CSL learners who participated in the data-collecting process, however, we limited the variations and just draw their attention to the existence of pragmatic factors known as I, S, and T/P.

To practice what we preach, we made every effort to include the I, S, and T/P clues while designing the tasks, which cover a wide variety of content areas with calculated degrees of difficulties. Responses to these tasks were obtained from participants and rated with confirmation to LU rules in mind, in addition to factors covered under

language structure. For instance, two very similar response samples from the same task to "express thanks at the farewell party" would be rated differently simply because one of them has no greetings at the beginning, even all other aspects (ideas expressed, complexity of vocabulary used and grammatical structure employed, etc.) are very similar.

Most of the oral productions were collected using the testing format. That means learners provided their responses either while taking the exam for real or in the situation in which a test is simulated. As mentioned above, pragmatic factors were used as the key criteria in the assessment of the proficiency level. Inclusion of pragmatic factors is actually a common practice in many well-recognized oral proficiency assessment tools, such as the Oral Proficiency Interview (OPI) by the American Council on the Teaching of Foreign Languages (ACTFL), or the Simulated Oral Proficiency Interview (SOPI) by the Center for Applied Linguistics (CAL), which pioneered large-scale oral proficiency assessment in the early 1980s.

Now let's take a closer look at some specific tasks used in the data-collecting process for the LAC/SC and see how information related to I, S, and T/P was included. Pasted below is an example of a language task used as part of the oral proficiency test. It is a task to elicit the spoken production from CSL learners with English L1 background, aiming at CSL learners whose proficiency level is expected to be at the Superior level according to the ACTFL proficiency guidelines (https://www.actfl.org/resources/actfl-proficiency-guidelines-2012). Please note that all the pragmatic factors related to LU are underlined.

e.g.2 Task description in English (for learners whose L1 is English)

You are at a farewell party given by your host organization in China for a group of teachers from your school, of which you are the leader. After the host makes a speech thanking you for the job you have done, you are invited to say a few words of thanks on behalf of all the teachers. During your one semester teaching in China, the host organization has been very helpful in many ways, making arrangements for accommodation, providing opportunities for teacher-student communication, doing the best they could to facilitate your teaching, and so on. Now think about what you want to say in this formal situation. After your Chinese host's introduction, respond on behalf of your group, expressing your appreciation for the hospitality of your host organization, acknowledging any inconvenience your group may have caused, and offering to reciprocate their hospitality. As in a formal speech, end your talk with a toast.

Prompt in Standard Chinese (Mandarin):

各位來賓, 現在我們請貴方的代表給我們講話。

Referring to the underlined parts above, which provide the pragmatic factors that are also part of the assessment criteria for the oral productions elicited by this type of task, we can now fill in the content of what I, S, and P stand for:

Interlocutors (clues for the I category):

Who you are: leader of a teacher delegation.
To whom you are speaking: the host and his/her team.

Setting (clues for the S category):

Formal situation (farewell party)
In public

Purposes of communication, with reference to content (clues for the T/P category):

Wait for your turn and
Speak on behalf of your team

This is what we mean by making explicit clues for LU, so that the speaker will use such a framework in their oral production, with appropriateness as one of the aims. In addition to the specific description telling you that this is a formal situation where former words and certain ritual in public speech are expected, clues of a formal situation are also given in the Chinese prompt (such as "gewei" and "guifang"), which serves as the final reminder before the speaker starts talking. We all understand that any prompt can elicit a response in such a situation, even if you simply say "now start talking". Providing contextual clues in language assessment, nevertheless, has now become the hallmark of the communicative approach in testing, also called Stage III in the history of assessment (Li, 1997; Douglas, 2000, Wu, 2008b). Test items or prompts in the Stage I period (coincided with a focus on LS in language teaching) will not bother to provide any contextual clues, such as quoted below:

e.g.3 Sample of prompt for oral test in Stage I assessment

"Talk about the most unforgettable person in your life"

Giving a speech in public in formal settings is designed for CLS learners at the advanced level. For those with a lower proficiency level, the task below will be more appropriate for eliciting their oral production.

e.g.4 Task description in English (for learners whose L1 is English)

You have a friend from Xiamen who likes reading a lot. Please tell him what types of book you like and what books are worth reading. Please think about it and answer after listening to the question in Mandarin.

Prompt in Standard Chinese (Mandarin):

你呢?你喜歡看什麼書呢?

We can see here that, similarly, pragmatic factors are also provided and can be summarized using the same categories above:

Interlocutors:

Who you are: a friend of somebody you talk to
To whom you are speaking: a friend to whom you speak

Setting:

informal situation

Purposes of communication, with reference to content:

Small talk on personal hobbies

Comparing the two task descriptions above, we start to see how a systematic approach in providing pragmatic factors in LU is implemented from the very beginning in the process of corpus construction before oral productions were collected. It was expected that the conditions set in each of these tasks under each of the three categories would produce data more appropriate for studies focusing on LU. If more and more research in this direction could be carried out from time to time, it would eventually contribute to a better understanding of LU as a system similar to LS.

For the past decades, studies on each of the subsystems under LS have produced a multitudinous amount of literature, most of which are somewhat related to theories in structuralism as a school in linguistics. Language teaching and learning as a field has benefited tremendously from such studies. Although impossible to quantify, we can speculate based on common sense that, if only one percent of the efforts for linguistic research from now on could focus on LU and its subsystems, we would soon understand much more about how language teaching and learning can benefit from sociolinguistic research.

The second step taken was to deal with the challenge that all task descriptions should be clearly understood by the speaker. Using L1 of the CSL learner in the testing formal is a practice in SOPI. Such a format has been recognized as one of the solutions to cover all participants, including those at the lower end of the proficiency spectrum. Over the years, this approach has attracted many controversies. For learners with advanced proficiency levels, it would be more efficient to use L2 in the whole process, involving no translation and thus avoiding possible misunderstanding of the requirement regarding pragmatic factors. Referring to the different stages in the history of assessment as an academic field, we can see that using L1 is determined by the development of testing theories and practices. Most of the assessment tools now are, or claim to be, communicative in nature. The aim is therefore to assess the ability of the learner in using language for communication, focusing on LU, rather than taking stocks of their knowledge of the language, focusing on LS. In order to achieve this goal, a simple request such as “circle the right answer” will not work.

The complexity of test instructions focusing on LU, and the description of the tasks to be performed, requires much more in terms of the grammatical structure, the difficulty level of the words involved, and the discourse features related to cultural beliefs and practices. Learners at the lower end of the proficiency level will not be able to understand the task descriptions with all the contextual clues in the language they are learning. Looking back, we can see the use of L1 became the alternative in

L2 tests at the beginning, and later the norm of large-scale assessment, especially in assessment tools that cover the whole spectrum of proficiency level (from novice to distinguished according to the ACTFL proficiency guidelines, or A1 to C2 according to the CEFR).

With the two steps taken, we found in the deep briefing after the data-collecting that the requirement of the tasks was well understood by the participants. At least they were aware that the speaker was "talking up" or "talking down", the setting was formal or casual. Whether or not they could adjust their oral production to match such pragmatic factors, however, would then depend on their own proficiency level.

4 Structure of a CSL Learner Corpus with Built-In Pragmatic Factors

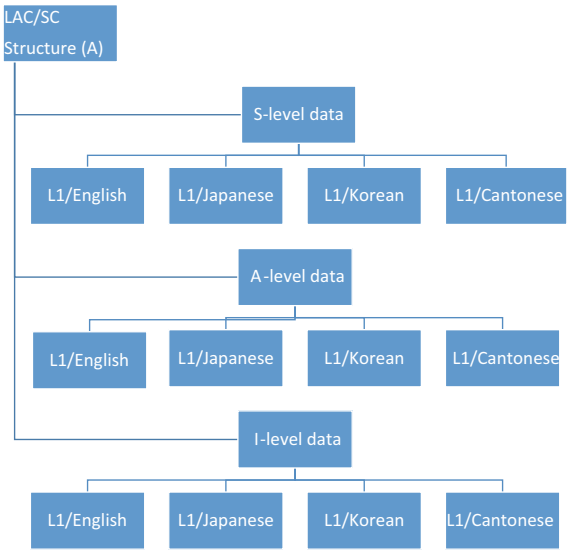
Once all the CSL leaner contributions were collected, two trained assessors would cross-rate each and every sample to determine the proficiency level. If there was a major difference in the ratings, a third assessor would be called in to have the final rating. All rated samples are then put under the LAC/SC structure and transcribed.

Bearing in mind the two distinctions mentioned previously (LS vs. LU, correct vs. appropriate), we can go on to provide an overall picture of the LAC/SC with reference to these distinctions. Since most of the CSL learners at that time were from Japan, Korea, and English-speaking countries, data collection was conveniently grouped according to the L1 of each group. For comparison, similar data were also collected from local students whose L1 is Cantonese. One of the reasons for adding this group was to see the differences and similarities in pragmatic ability between this group, which is a subset of the Chinese language and culture, and the other three groups, which were not part of the Chinese language and culture family. Such an addition turned out to be very helpful later on because it provides one more dimension in L2 acquisition studies: the comparison of the in-group versus the out-group in their understanding of Chinese cultural concepts and practices, as demonstrated by the degree of appropriateness in their oral productions. It also shed light on why pragmatic factors are important and should not be neglected in the corpus construction process.

To fully explain the organization of the corpus, we can trace the path of our construction process from the very beginning, when we were still trying to decide on the top layer of the corpus. There were two possibilities, as illustrated by the two figures below. In structure A, the top layer is the proficiency level of the oral production data, while learners with different L1 backgrounds are grouped beneath, under each proficiency level.

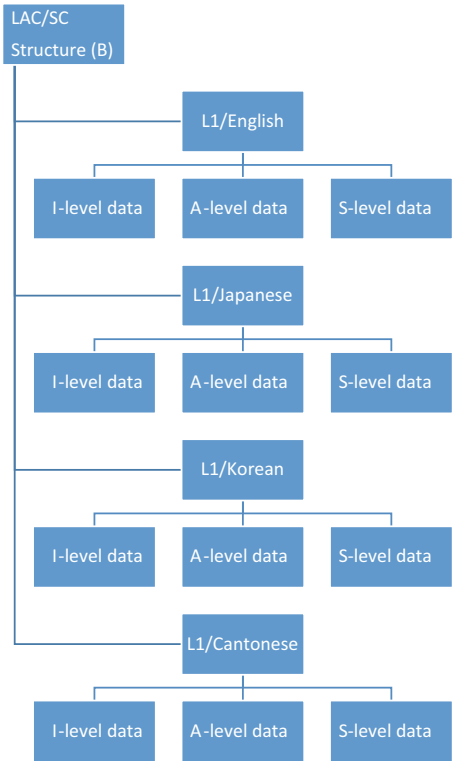
Structure A: CSL Learner Corpus based on the proficiency level

(Note: Proficiency levels used here in the corpus are based on the ACTFL Proficiency Guidelines, where "I" stands for Intermediate, "A" for Advanced, and "S", Superior).



As shown in Structure A, there are three levels in the chart, each covering all oral production data from different L1 groups at a specific level. These three levels are fixed and cannot be changed and is therefore a closed system. Under each of them, however, it is an open system that allows block building. In this chart, we have data from CSL learners whose L1 is English, Japanese, Korean, and Cantonese, respectively, but we can see from the structure that, should we have data from other L1 learners (e.g. Russian and Thai according to the plan), we can easily add more blocks so that, instead of 4 L1 groups, we will have 5 or 6, or more as we continue. With reference to the systematic implementation of pragmatic factors, which tend to have the same features at the same proficiency level, this structure was very attractive because it can also make things easier for pragmatic annotation and tagging down the road. This allows the necessary flexibility for an ongoing research project of a similar nature.

Structure B, on the other hand, used the learner group as the fixed layer, commanding all the oral production data at different proficiency levels from the same L1 group.



The advantage of such an arrangement is ready access to any study focusing on one L1 learner group and to any comparison study within the same L1 group, regardless of the proficiency level in their oral production. From the perspective of language acquisition studies, such a structure offers conveniences in following the development of acquisition within a particular L1 group. Once the collection of data started, however, it was discovered that getting a sizeable group of learners at the same proficiency level within any L1 group was more difficult than expected, especially those at the higher end of the proficiency level. That means space allocated to that particular L1 group at a particular level would remain unfilled for an unknown period of time. Such uncertainty is hard to tolerate in most research projects and, moreover, it may also lead to inconveniences in any attempt to do research within the same proficiency level because of the lack of data.

Given the fact that the systematic implementation of pragmatic factors depends on a sizeable population of advanced learners, and with consideration of the difficulties involved in obtaining enough oral production of CSL learners at the high end of the proficiency, Structure A was finally adopted and all data collected were grouped accordingly. With the most recent update, the LAC/SC now has the following data:

CSL learners with English L1 (90 + samples):

transcribed and checked, in 3 proficiency levels;

CSL learners with Japanese L1 (45 + samples):
transcribed, in 3 proficiency levels;

CSL learners with Korean L1 (45 + samples):
transcribed, in 3 proficiency levels;

CSL learners with Cantonese L1 (600 + samples):
with 90 + of them transcribed and checked in 3 proficiency levels.

By design, each of the speaking samples has approximately 11 min of speaking time, covering 12 different content areas in various settings under three categories: informal, formal, and ceremonial or ritual.

As discussed earlier, three pragmatic factors were built in at the very beginning of the data-collecting process, when the tasks for participants were created. A pragmatic frame that includes information about the Interlocutors, the Setting, and the Timing (or Purpose) of the speech sample obtained and used in the corpus is available at two levels: the task level for all speakers, as well as the individual speaker level for all his/her tasks. The advantage of the availability of such information is obvious: it is now possible to conduct studies related to the appropriateness of oral production, either at the task level across L1 groups to find the similarities and differences, or at the speaker level to find the unique features associated with a certain L1 group.

Due to resource limitations, tagging and annotation for LAC/SC are still waiting to be completed. Studies have been done, nevertheless, based on clean copies of the transcription with sound files, including studies of prosodic features based on sound files, and sociolinguistic research focusing on advanced CSL learners' oral production based on the transcription and the sound files. Compared with other common corpus-based research focusing on phonological, semantic, and syntactic studies, one outstanding feature of the LAC/SC is the possibility to do research on the pragmatic ability of CLS learners. It is expected that, once the tagging of pragmatic features is completed and made searchable (e.g. presence/absence of greeting at beginning of a speech in a formal setting would be very useful for studies at a discourse level), more research can be done focusing on the pragmatic ability of the CSL learner at different proficiency levels; the salient features related to the appropriateness of a particular L1 group or proficiency level; and their understanding and use of words, phrases, and grammatical devices to show modesty as the native speakers tend to do, among others.

5 Concluding Remarks

Like all ongoing research, it is impossible, nor is it responsible, to draw any conclusion at this stage because LU as a system is still new to many, and too many questions remain unanswered. Based on the experience in the problem-finding and solution-seeking process while building the LAC/SC so far, however, it is reasonable to point out the following in reference to the pragmatic issues discussed in this paper.

1. In the construction of a learner corpus, data related to language structure (LS) is a given but information related to language use (LU) should be included. Moreover, it should be an integrated part of the corpus, starting from the very beginning as part of the overall design, not just an add-on later on. Framing the task of eliciting oral responses with real communication settings is a positive step in the right direction.
2. While LS is a system of systems consisting of phonology, semantics, and syntax (LS = Pho + Sem + Syn), LU can also be treated as a system of systems consisting of interlocutors (I), setting (S), and timing, or purpose of the communicative event (T/P) if clues for timing is not available or not important (LU = I + S + T/P). Employment of both systems, and not just one of them, could serve as a solid foundation in the construction of the CSL learner corpus, or any corpus for that matter.
3. Each of the subsystems in LU, like those in LS, can be further divided into categories and sub-categories, (e.g. the three categories under Setting: informal, formal, and ceremonial or ritual). Each of the sub-categories certainly has layers that allow further division for research purposes, such as semi-formal between informal and formal.
4. Compared with studies in LS, research is badly needed for LU as a system with reference to development in sociolinguistics. We must admit that only very little is understood about LU at this stage and there are many more factors in this system than what we have discussed here. Findings from more research in this direction, however, are expected to eventually contribute to the goal of culturally appropriate productions from CSL learners.
5. For tagging and annotation of the data in the LAC/SC down the road, both information for LS and LU should be included, starting with those related to I, S, and T/P at this stage. It would be impossible to study the appropriateness of language use if no information about LU is available.

Looking back and looking around, we must reiterate that studies on LS have been long and many, while those on LU are still sporadic by comparison. That means it is natural for us to see many more questions and challenges for any research focusing on LU. However tentative the concluding remarks above may seem to be, and no matter how big a hole we can see in various aspects of the LAC/SC as reported here, or similar projects reported somewhere else, what we have discussed in this paper helps us see that the study of pragmatic factors in corpus construction can contribute to the development of our field, even if some of us feel that our discussion has led to many questions and offered few answers.

References

Publications

Brezina, V., & Flowerdew, L. (Eds.) (2019). *Learner corpus research: new perspectives and applications*. Bloomsbury Academic.

Chao, Y. R. (1968). *A grammar of spoken Chinese*. University of California Press.

Chen, H., & Tao, H. (2019). Academic Chinese: from corpora to language teaching. In X. Lu, & B. Chen (Eds.), *Computational and corpus linguistic approaches to Chinese language teaching and learning* (pp. 57–79). Berlin & Singapore: Springer.

Clark, J. L. D., & Li, Y. C. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages*. Center for Applied Linguistics.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.

Fan, L. (2018). Pragmatic competence of advanced CSL learners in spoken Chinese: A comparison of native speakers of English and of Cantonese. Ph.D. Dissertation in Linguistics, The Hong Kong Polytechnic University, Hong Kong.

Feng, S. (2018). A sketch of Chinese *Yuti* Grammar. Beijing Language and Culture University Press.

Ke, C. (1994). An empirical investigation of the relationship between a simulated oral proficiency interview and the ACTFL oral proficiency interview. *Selecta*, *15*, 6–10

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press.

Li, X. (1997). The science and art of language assessment. Hunan Education Press.

Li, Y. (2021). Responsibilities and concerns of linguistics. *Journal of North China University (Social Science Edition)*.

Liu, J. (2008). Research on pragmatic ability: Current status, problem and revelations. *Foreign Language Research*, *4*.

Ran, Y. (2004). New achievement in the interdisciplinary research on pragmatics and second language acquisition: Review on pragmatic development in a second language. *Foreign Language Teaching and Research*, *2*.

Rose, K., & Kasper, G. (Eds.). (2001). *Pragmatics in language teaching*. Cambridge University Press.

Tao, H. (2017). Spoken Chinese corpora: Construction and sample applications in research and language pedagogy. *Bulletin of the Chinese Linguistic Society of Japan., 2017*(264), 25–43.

Tao, H., Jin, H., & Zhang, J. (2020). A corpus-based investigation of manner/state complement constructions in Mandarin Chinese. *Sinica Venetiana, 6*, 1–40. https://doi.org/10.30687/978-88-6969-406-6/001

Wang, H. (2006). Pragmatics in foreign language teaching and learning: Reflections on the teaching of Chinese in China. In W. Chan et al. (eds.). *Foreign Language teaching in Asia and beyond: Current perspectives and future directions*. Center for Language Studies, National University of Singapore.

Wang, C. (2010). How we learn a foreign language. Foreign Language Teaching and Research Press.

Wu, W. (1993). *Towards a theory of teaching chinese as a second language*. Springfield, VA: ERIC Document Reproduction Service. ED 366 216.

Wu, W. (2006). Pragmatic points in teaching Chinese: A practical approach. *Chinese Teaching in the World, 1*, 91–96.

Wu, W. (2008a). Pragmatic framework and its role in language learning: With special reference to Chinese. In W. Chan et al (eds.), *Processes and process-orientation in foreign language teaching and learning*. Germany: De Gruyter Mouton. (Reprinted 2011).

Wu, W. (2008b). Teaching Chinese as a foreign language: theory and practice in proficiency test with a pragmatic approach. *Journal of Chinese Language Teaching, Beijing University Press*, *4*.

Wu, W. (2016). Chinese language pedagogy. In S. Chan, J. Minett, & F. Li (Eds.), *The Routledge Encyclopedia of the Chinese Language* (pp. 137–151). Routledge.
Wu, W. (2019). Language structure vs language use in TMP: Focusing on pragmatic ability of learners. *TCSOL Quarterly*, (1).
Wu, W. (2020). Implementation of the pragmatic framework in teaching: a systematic approach for CSL. *TCSOL Quarterly*, (2).
Xiong, D., et al. (2002) Research on large-scale recorded oral assessment for College English. *Foreign Language Teaching and Research*, *4*.
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.
Zhang, B., et al. (2019). *Research on tagging of Chinese interlanguage corpus*. Beijing University Press.

(on-line)

ACTFL https://www.actfl.org/resources/actfl-proficiency-guidelines-2012.
CEFR https://www.commoneuropeanframework.org.
English Corpora https://www.english-corpora.org/.
Learner Corpus Association https://www.learnercorpusassociation.org/.
Learner corpora around a world https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.
Media Language Corpus http://ling.cuc.edu.cn/RawPub/.
On-line Corpora www.cncorpus.org.
Sinica Corpus http://www.sinica.edu.tw/SinicaCorpus/.

A Preliminary Study on Chinese Learners' Written Errors Based on an Error-Tagged Learner Corpus

Ting-Yu Yang, Hui-Mei Yang, Wei-Jei Lee, Chen-Yu Liu, and Howard Hao-Jan Chen

Abstract With the development of technology, the need for compiling computer-based learner corpora has gradually gained more attention from language teachers and researchers. A learner corpus can reflect learners' authentic use of a target language, which provides useful information for language teachers, researchers, and textbook editors. Limitations of retrieving errors in learner corpora, however, still exist. For example, it is difficult to retrieve omission errors if a corpus is not error-tagged beforehand. To offer researchers an error-tagged learner corpus of Chinese, this study manually error-tagged the two-million-word Chinese Learner Written Corpus of National Taiwan Normal University. A preliminary analysis of errors tagged in the learner corpus shows a total of 48,266 errors distributed to 119 tags. These 48,266 errors are mostly distributed to the incorrect selection of words or the missing of necessary word-level components, and the misuse of nouns, action verbs, adverbs, and structural particles is especially common. Among the 119 tags, the top 12 common error tags (i.e., occurring more than 1,000 times) accounted for more than 50% of the total errors, and incorrect selections of nouns and action verbs together constituted more than 27% of the total errors. These 12 common error types, especially the wrong choice of nouns and action verbs, should thus be regarded to be particularly difficult for second language (L2) learners of Chinese to acquire. Analysis of the top 12 common errors also reveals that learners' misuse of verbs, adverbs, and structural particles were somewhat varied (i.e., involving different types of target modification, such as missing, redundant, and incorrect selection), whereas their misuse of nouns mostly resulted from an incorrect selection. A comparison between the top 10 common error types in this study with those in Lee et al. (2016) reveals that, regardless of some discrepancies in ranking, 90% of the top 10 error tags overlapped in the two studies, suggesting that these error types are indeed difficult for L2 Chinese

T.-Y. Yang (✉) · H.-M. Yang · W.-J. Lee · H. H.-J. Chen
Department of English, National Taiwan Normal University, 162, Section 1, Heping East Road, Taipei City 106, Taiwan
e-mail: christiney37@gmail.com

C.-Y. Liu
English Language Center, Ming Chuan University, 5, De Ming Road, Taoyuan City 333, Taiwan

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_4

learners to acquire and should be investigated further. Based on the findings yielded in this study, suggestions for further research on L2 Chinese learners' errors are provided.

Keywords Chinese teaching · Learner corpus · Error-tagging · Error analysis

1 Introduction

1.1 The Development of Learner Corpus

The concept of error analysis was firstly introduced by Corder (1967), who pointed out the significance of analyzing language learners' erroneous output to understand the linguistic features and developmental process of their interlanguage. Since then, analyses of language learners' errors have been one of the main research areas in the field of second/foreign language (L2) learning (Pan & Liu, 2006). Early studies on language learners' errors were mostly based on language teachers' reports on learners' erroneous sentences observed in their teaching, which often included a limited number of language learners' errors. The problem with small-sized samples stems from the fact that no statistical analysis can be performed to formulate rules of learners' interlanguage (Corder, 1967; Nemser, 1974; Selinker, 1972). Thus, the limited number of errors identified in early studies makes it difficult for researchers to systematically establish the causes of learners' errors and to obtain more generalizable results to point out their linguistic features.

The importance of collecting and analyzing a large quantity of learner errors to gain more generalizable results urges the establishment of a learner corpus. Learner corpora are electronic collections of authentic linguistic output by L2 learners. They consist of data larger than the types (e.g., output from elicitation tasks) commonly used in second language acquisition (Granger, 2003), and therefore afford researchers the confidence to report significant recurrent patterns or errors produced by L2 learners (McEnery et al., 2019). In addition, the electronic format of learner corpora allows researchers to extract target language structures from a large number of data for further analysis with a wide range of software tools, saving researchers more time and effort in the manipulation of the data (Granger, 2003).

With the wide application of learner corpora in research and the compilation of teaching/learning materials, more and more research institutes and publishers are involved in the building of learner corpora. The first learner corpus, Longman Learners' Corpus, was compiled by Longman Publishing Group in the late 1980s, which contains 10 million words of English learners' essays and exam scripts worldwide. In 1990, Sylviane Granger started building International Corpus of Learner English, and she continues to expand its size to more than 5.5 million words, which consists of learners' written data from 25 first language (L1) backgrounds. Since the 1990s, the number of learner corpora has been rapidly increasing. According to a survey by Centre for English Corpus Linguistics of Louvain-La-Neuve (2020),

there are more than 180 learner corpora around the world, consisting learners' written/spoken data from more than 20 target languages. Currently, more than half of the corpora target the output of English learners, and around 25 of them contain more than 1 million words.

The growing trend of teaching/learning Chinese as a Second/Foreign Language (CSL/CFL) also encourages the development of Chinese learner corpora. To the best of our knowledge, the biggest learner corpora of learners' Chinese is Jinan Chinese Learner Corpus, a 6-million-character corpus containing exam scripts and assignments by learners from over 50 different L1 backgrounds (Wang et al., 2015). The second largest corpus is the 4.24-million-character HSK Dynamic Composition Corpus, which covers more than 11,000 compositions by exam takers of Hanyu Shuiping Kaoshi (HSK). The third largest corpus is the Continuity Corpus of Chinese Interlanguage of Character-error System, a 2-million-character corpus consisting of learners' sentence-makings and essays (Zhang, 2013). While the three corpora deal with simplified Chinese, attempts have also been made to build learner corpora of traditional Chinese. For example, Chinese Learner Written Corpus of National Taiwan Normal University collects more than 2 million characters of writings in traditional Chinese by learners from more than 60 different L1 backgrounds. Another corpus dealing with traditional Chinese is the 1.5-million-character TOCFL Learner Corpus, which collects 4,567 exam scripts from the Test of Chinese as a Foreign Language (TOCFL).

With these resources, researchers have employed these learner corpora to investigate Chinese learners' interlanguage and yielded some insightful results. For example, Zhang (2010) examined Chinese learners' use of 把 *bǎ*-sentences from HSK Dynamic Composition Corpus and discovered that the learners' avoidance of 把 *bǎ*-sentences was not as obvious as indicated in previous studies. Also based on HSK Dynamic Composition Corpus, Wang (2010) investigated Russian CSL learners' erroneous use of the particle 了 *le* and reported that missing 了 *le* was the most frequent error in these learners' writing. Hu's (2012) investigation of CSL learners' use of the adverb 都 *dou* revealed that low-level learners tended to misuse 都 *dou* significantly more often than both intermediate-level and advanced-level learners. In addition to the use of HKS Dynamic Composition Corpus, studies based on Chinese Learner Written Corpus were also conducted to examine learners' interlanguage. Wang et al. (2013) investigated Chinese learners' uses of two sets of synonymous verbs: 幫 *bang*, 幫忙 *bang-man*, 幫助 *bang-zhu*, and 變 *bian*, 變得 *bian-de*, and 變成 *bian-cheng*, and findings of their study showed that learners often wrongly replace 幫忙/幫助 *bang-man*/*bang-zhu* with 幫 *bang* and 變得/變成 *bian-de*/*bian-cheng* with 變 *bian*. Lin et al. (2014) examined the use of directional complement 起來 *qilai* based on Chinese Learner Written Corpus, and they discovered that the learners had great difficulty in using the stative meaning of 起來 *qilai*, which was mostly attributable to misformation.

Construction of these existing Chinese learner corpora provides a considerable amount of learner output for researchers to explore CSL/CFL learners' interlanguage with quantitative statistics; however, some error types, such as omission errors, might

not be easily retrieved by the direct use of these corpora. To better resolve this problem, further processing of learner data with error-tagging is suggested.

1.2 The Development of Error-Tagged Learner Corpus

Learner corpus researchers (e.g., Díaz-Negrillo & Domínguez, 2006; Jia, 2007; Tono, 2003) have been advocating the importance of annotating learners' grammatical errors to provide useful information for the development of L2 research and/or teaching (Brook & Hirst, 2012; Granger, 2015; Swanson & Charniak, 2013; Wang & Seneff, 2007). Error-tagged learner corpora, however, are relatively scant. With the help of computer programs, most of the current learner corpora are annotated with part-of-speech (POS) tags, which allow users to carry out meaningful searches of target linguistic features (e.g., nouns, verbs, and adjectives) rather than a single word form (McEnery et al., 2019). Nevertheless, annotation of learners' errors requires more time and effort since tagging learners' grammatical errors heavily relies on human judgment and can only be done manually (Lüdeling & Hirschmann, 2015). Thus, only few current learner corpora are error-tagged.

To the best of our knowledge, two of the largest error-tagged learner corpora are Cambridge Learner Corpus and Longman Learners' Corpus. Cambridge Learner Corpus, currently the largest error-tagged learner corpus, contains annotations of 30 million words, the error-tagging system of which was devised by Cambridge University Press. This error-tagged corpus has become one of the major resources for publishers to compile English teaching/learning materials and dictionaries (Nicholls, 2003). Longman Learners' Corpus, built by Longman Publishing Group, is composed of 10 million words with error-tagging and also serves as a useful reference for the publisher to compile dictionaries. The dictionary, Longman Dictionary of Common Errors, is in fact compiled based on the learner corpus. Other error-tagged learner corpora of English include the 1-million-word Chinese Learner English Corpus, the 2.5-million-word HKUST Corpus of Learner English, and the 700,000-word Japanese EFL Learner Corpus.

In addition to learner English, efforts have also been made to the annotation of learner Chinese. The HSK Dynamic Composition Corpus is currently considered the most comprehensive error-tagged learner corpus of simplified Chinese. In the corpus, errors are manually annotated and distributed into four major categories, namely character-level errors (11 cases), word-level errors (5 cases), sentence-level errors (28 cases), and discourse errors (1 case). Based on the error-tagged data, investigations on CSL/CFL learners' interlanguage have been conducted to reveal learners' overall error distribution (e.g., Hsu, 2011) or errors in specific linguistic forms (e.g., Han, 2016; Jin, 2011; Li, 2013; Zang, 2014). Other error-tagged learner corpora of simplified Chinese include the Jinan Chinese Learner Corpus and the Continuity Corpus of Chinese Interlanguage of Character-error System.

Regarding the construction of error-tagged learner corpus in traditional Chinese, the TOCFL Learner Corpus is one of the corpora that contains around 1 million

characters of manually annotated errors produced by learners of traditional Chinese, in which 2,837 out of 4,567 learner essays that are graded at least 3 are error-tagged. Errors in the corpus are also distributed into four major categories, which are somewhat different from the error category of the HSK Dynamic Composition Corpus. In the TOCFL Learner Corpus, a total of 36 error types are categorized into word-level errors (16 cases), grammatical function-level errors (11 cases), sentence pattern-level errors (7 cases), and mixture errors (2 cases). Based on the corpus, Lee et al. (2016) analyzed 33,835 grammatical errors in the 2,837 essays and reported the top 10 error types that account for 47% of the total errors as follows: incorrect selection of action verb (n = 3,809, 11.26%), incorrect selection of noun (n = 2,167, 6.40%), missing adverb (n = 1,755, 5.17%), missing aspectual particle (n = 1,602, 4.73%), missing auxiliary (n = 1,357, 4.01%), incorrect selection of adverb (n = 1,168, 3.45%), missing structural particle (n = 1,165, 3.44%), missing action verb (n = 1,040, 3.07%), redundant aspectual particle (n = 1,003, 2.96%), and incorrect selection of stative verb (n = 780, 2.31%). While an incorrect selection of action verb is the most common error type, half of the 33,835 errors are attributed to missing word-level linguistic components.

Although efforts have been made to construct error-tagged learner corpora of Chinese, most of the current corpora, however, are based on simplified Chinese. While Lee et al. (2016) have contributed to the building of error-tagged learner corpora of traditional Chinese, the size of which is comparatively smaller than the HSK Dynamic Composition Corpus and the Jinan Chinese Learner Corpus. To provide CSL/CFL researchers with more resources for the study of learners' inter-language around the world, the current study aims to annotate data in the Chinese Learner Written Corpus of National Taiwan Normal University and to reveal the common error types made by CSL learners in Taiwan, results of which could offer researchers useful insights for further research on CSL learners' common errors. In the next two sections, we will firstly describe how we annotated errors in Chinese Learner Written Corpus, and present common error types identified in the corpus.

2 Method

2.1 The Learner Corpus

In this study, the Chinese Learner Written Corpus (http://kitty.2y.idv.tw/~hjchen/cwrite-mtc/main.cgi) was chosen as the target corpus for error-tagging. The corpus contains 4,288 essays (totally 2.14 million characters) written by CSL learners from 64 different countries at the Mandarin Training Center of National Taiwan Normal University during 2010–2012. All of the essays were take-home assignments hand-written by CSL learners and later manually typed as electronic files by the corpus builders. Genres of the essays include general epistle (e.g., a letter to your parents/siblings/friends), narrative (e.g., an unforgettable trip), argumentation

(e.g., a comparison between what you have in your home country and what we have in Taiwan), and application (e.g., your autobiography), written by learners across five proficiency levels (i.e., A2, B1, B2, C1, and C2 refer to the Common European Framework of Reference for Languages) and graded from 2 to 9.

2.2 Tagging of the Learner Corpus

2.2.1 Error Domain and Category

To annotate errors in the Chinese Learner Written Corpus, we adopted the hierarchical tag sets of grammatical errors established by Chang (2017), an error classification system that combines both target modification taxonomy (TMT) and linguistic category classification (LCC). The TMT system is "based on the ways in which the learner's erroneous version is different from the presumed target version" (James, 1998, p.106), while the LCC system is carried out "in terms of where the error is located in the overall system of the target language based on the linguistic item which is affected by the error" (James, 1998, p.105). In the error classification system by Chang, an error is tagged simultaneously with a capital letter denoting target modification based on TMT and subsequent lowercase letters denoting the linguistic category of the error based on LCC. There are four error types of target modification, namely missing (M), redundant (R), incorrect selection (S), and word ordering error (W). As for linguistic category, there are totally 36 error types distributed into word-level error, grammatical function-level error, sentence pattern-level error, and mixture error (see Table 1). The advantage of using such a mixed error classification system is that the annotator can effectively assign an error to a specific tag without referring to the tagset each time. Once the annotator specifies how an erroneous surface structure deviates from the target language based on the four main types (i.e., M, R, S, and W), the annotator will only need to identify the problematic linguistic item of that error.

2.2.2 Error Marking Tool

In this study, we employed a software developed by a programming team led by Prof. Yuen-Hsien Tseng at NTNU to annotate errors in the learner corpus, the interface of which is shown in Fig. 1. The left column shows the text files of the learner corpus, and the other two columns present the running text of each selected file. Annotators can mark errors in a chosen text in the central column, and errors will be highlighted in red with error tags. The right column then presents the text corrected by annotators, and corrections will be highlighted in blue.

Table 1 Tags of errors in linguistic category (adopted from Chang, 2017)

Linguistic category	
Word-level (16 cases)	Action verb (v), auxiliary (aux), stative verb (vs), noun (n), pronoun (pron), conjunction (conj), preposition (p), numeral (num), demonstrative (det), measure word (cl), sentential particle (sp), aspectual particle (asp), adverb (adv), structural particle (de), question word (que), plural suffix (plural)
Grammatical function-level (11 cases)	Subject (sub), object (obj), noun phrase (np), verb phrase (vp), preposition phrase (pp), modifier (mod), time expression (time), place expression (loc), transitivity (tran), separable structure (vo), [numeral/determiner + measure] phrase (dm)
Sentence pattern-level (7 cases)	Complex noun clause (rel), 把 *ba*-sentence (ba), 被 *bei*-sentence (bei), 讓 *rang*-sentence (rang), 是 *shi*-sentence (shi), 有 *you*-sentence (you), other patterns (pattern)
Mixture (2 cases)	Formation (form), ambiguity of syntactic or meaning (sentence)

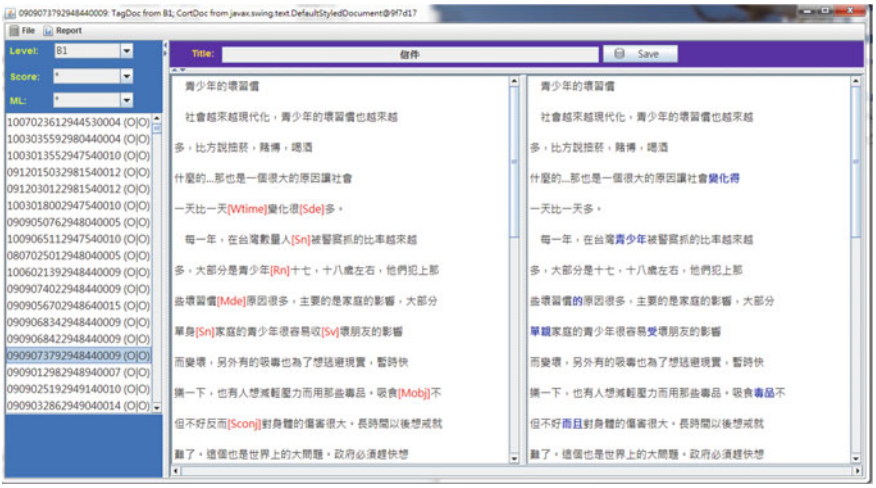


Fig. 1 The interface of the error marking software

2.2.3 Principles of Error Marking

To ensure the consistency of the two human annotators' error identification and marking, the annotators would have to follow the annotation guidelines developed in this study. First, corrections of errors were made with two premises. The first premise was that annotators' corrections should not alter what learners intended to express. In addition, annotators should use words/phrases in accordance with learners' language proficiency. Secondly, annotators would firstly determine the target modification of an error (i.e., M, R, S, W) and then assign the erroneous element to the linguistic category.

3 Results and Discussion

3.1 Number and Distribution of All the Annotated Errors

In the learner corpus, 48,266 errors were identified and annotated by the annotators, which were distributed into 119 error tags. The numbers and percentages of the 119 error tags are presented in Table 2. Many of the errors belonged to incorrect selection and missing linguistic components, which respectively took up 39.86 and 36.24% of the total errors. As for the linguistic category, 80.7% of the total errors belonged to the word-level, while errors at the other three levels took up less than 20%. In addition, incorrect selection of word-level linguistic components, missing word-level linguistic components, and redundant word-level linguistic components totally accounted for 77.42% of the total errors, whereas word ordering errors of word-level linguistic components took up only around 3%. Further examinations of errors at the word-level revealed that nouns, action verbs, and adverbs were the top three commonly misused linguistic components, all of which accounted for more than 13% of the total errors, and the fourth commonly misused linguistic components, structural particle (de), amounted to around 9% of the total errors. These four commonly misused components amounted to around 50% of the total errors.

In sum, the distribution of the 48,266 errors revealed that CSL learners have greater difficulties in choosing the right words or making correct sentences with necessary word-level components. These deficiencies were especially serious in their use of nouns, action verbs, adverbs, and structural particles. Since half of the total errors were in the four word classes, more investigations on words in these word classes should be further conducted to better understand how and why CSL learners misuse these components in their writing.

3.2 The Most Frequent Error Tags in the Learner Corpus

To further understand the common error types in the learner corpus, error tags with more than 1,000 counts were identified for further discussion. Figure 2 illustrates the distribution of the top 12 error tags with more than 1,000 counts, which accounted for more than 50% of the total errors. The most common errors were attributed to the incorrect selection of nouns (Sn) and action verbs (Sv), the summation of which constituted 20% of the total errors; the other 10 error types, on the other hand, represented around 30%. Table 3 presents example sentences extracted from the learner corpus for the 12 error tags.

While the incorrect selection of nouns was the most frequent errors identified in the learner corpus, it was also the only one out of the 12 error types that related to the misuse of nouns. Among the top 12 error types, three resulted from the misuse of verbs (i.e., Sv, Svs, and Mv), three resulted from the misuse of adverbs (i.e., Madv, Sadv, and Radv), two resulted from the misuse of structural particles (i.e., Mde and Rde),

Table 2 Distribution of the 48,266 errors among the 119 error tags

		M		R		S		W		Subtotal	
		n	%	n	%	n	%	n	%	n	%
Word-level	v	1,474	3.05	830	1.72	4,181	8.66	0	0	**6,485**	**13.44**
	vs	449	0.93	204	0.42	1,926	3.99	126	0.26	**2,705**	**5.6**
	n	463	0.96	577	1.2	5,556	11.51	330	0.68	**6,926**	**14.35**
	aux	1,741	3.61	337	0.7	508	1.05	88	0.18	**2,674**	**5.54**
	pron	69	0.14	140	0.29	251	0.52	0	0	**460**	**0.95**
	conj	940	1.95	527	1.09	1,021	2.12	59	0.12	**2,547**	**5.28**
	p	911	1.89	535	1.11	385	0.8	0	0	**1,831**	**3.79**
	num	84	0.17	36	0.07	49	0.1	0	0	**169**	**0.35**
	det	320	0.66	224	0.46	339	0.7	51	0.11	**934**	**1.94**
	cl	377	0.78	95	0.2	495	1.03	6	0.01	**973**	**2.02**
	sp	31	0.06	31	0.06	16	0.03	0	0	**78**	**0.16**
	asp	1,202	2.49	810	1.68	107	0.22	84	0.17	**2,203**	**4.56**
	adv	2,923	6.06	1,239	2.57	1,618	3.35	632	1.31	**6,412**	**13.28**
	de	2,478	5.13	1,129	2.34	603	1.25	189	0.39	**4,399**	**9.11**
	que	34	0.07	26	0.05	86	0.18	13	0.03	**159**	**0.33**
	Subtotal	**13,496**	**27.95**	**6,740**	**13.96**	**17,141**	**35.51**	**1,578**	**3.26**	**38,955**	**80.7**
Grammatical function-level	sub	347	0.72	127	0.26	0	0	201	0.42	**675**	**1.4**
	obj	293	0.61	12	0.02	0	0	219	0.45	**524**	**1.09**
	np	368	0.76	253	0.52	177	0.37	176	0.36	**974**	**2.02**
	vp	25	0.05	41	0.08	84	0.17	104	0.22	**254**	**0.53**

(continued)

Table 2 (continued)

		M		R		S		W		Subtotal	
		n	%	n	%	n	%	n	%	n	%
	pp	19	0.04	40	0.08	9	0.02	243	0.5	**311**	**0.64**
	mod	33	0.07	67	0.14	0	0	24	0.05	**124**	**0.26**
	time	176	0.36	57	0.12	245	0.51	103	0.21	**581**	**1.2**
	loc	311	0.64	112	0.23	207	0.43	93	0.19	**723**	**1.5**
	tran	0	0	0	0	0	0	20	0.04	**20**	**0.04**
	dm	72	0.15	123	0.25	14	0.03	56	0.12	**265**	**0.55**
	vo	65	0.13	0	0	0	0	31	0.06	**96**	**0.2**
	plural	0	0	15	0.03	0	0	0	0	**15**	**0.03**
	Subtotal	**1,709**	**3.53**	**847**	**1.73**	**736**	**1.53**	**1,270**	**2.62**	**4,562**	**9.46**
Sentence pattern-level	rel	0	0	0	0	14	0.03	3	0.01	**17**	**0.04**
	ba	61	0.13	44	0.09	14	0.03	2	0	**121**	**0.25**
	bei	71	0.15	42	0.09	18	0.04	3	0.01	**134**	**0.28**
	rang	175	0.36	51	0.11	97	0.2	0	0	**323**	**0.67**
	shi	886	1.84	448	0.93	92	0.19	42	0.09	**1,468**	**3.04**
	you	504	1.04	385	0.8	88	0.18	29	0.06	**1,006**	**2.08**
	pattern	99	0.21	47	0.1	49	0.1	0	0	**195**	**0.4**
	Subtotal	**1,796**	**3.73**	**1,017**	**2.12**	**372**	**0.77**	**79**	**0.17**	**3,264**	**6.76**
Mixture	form	495	1.03	0	0	377	0.78	0	0	**872**	**1.81**
	sentence	0	0	0	0	613	1.27	0	0	**613**	**1.27**
	Subtotal	**495**	**1.03**	**0**	**0**	**990**	**2.05**	**0**	**0**	**1,485**	**3.08**

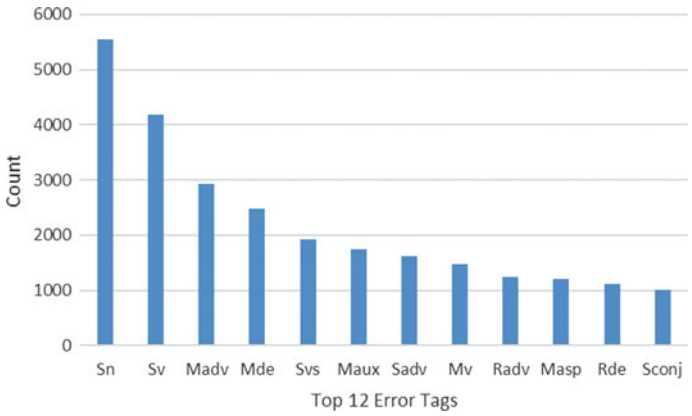


Fig. 2 Distribution of the top 12 error tags

and the others related to the misuse of different word-classes. Based on the findings, it was obvious that the learners were prone to misuse verbs, adverbs, and structural particles in various ways. On the contrary, their misuse of nouns was mostly attributed to incorrect selections, suggesting that learners' incorrect use of nouns might result from their confusion of nouns with similar meanings or forms. Hence, research on CSL/CFL learners' misuse of nouns is suggested to specifically investigate learners' difficulties in acquiring and differentiating synonymous nouns. As for the misuse of verbs and adverbs, researchers are suggested to examine CSL/CFL learners' use of specific verbs/adverbs and uncover the causes of their misuse(s).

3.3 Comparison of Results in This Study and the Previous

In addition to presenting the common error types in our learner corpus, we also compared findings yielded in our study with those in Lee et al. (2016). The reasons for drawing such a comparison are that both the two studies used the same error annotation system and investigated CSL learners' written production in traditional Chinese. Comparisons between the two studies might help us to identify the common errors produced by SL/FL learners of traditional Chinese. Table 4 presents the comparisons of the top 10 error tags in the two studies.

As shown in Table 4, nine out of the top 10 error tags in this study also appeared in Lee et al. (2016). The top 3 error tags in the two studies were an incorrect selection of nouns (Sn), incorrect selection of action verbs (Sv), and missing adverbs (Madv), though the top two were in reversed orders. From top 4 to top 10, however, rankings in the two studies were somewhat different. Discrepancies in the rankings of missing auxiliary (Maux), incorrect selection of adverbs (Sadv), and missing action verbs (Mv) in the two studies were small. Missing action verbs ranked eighth in both studies.

Table 3 Example sentences and suggested corrections of the top 12 error tags

Rank	Tag	Example sentence
1	Sn	*(a) 雖然青年人[Sn]吸毒已成為當前很多國家的社會問題。 *Suiran qingnianren* [Sn] *xidu yi chengwei dangqian henduo guojia de shehui wenti.* (b) 雖然年輕人吸毒已成為當前很多國家的社會問題。 *Suiran nianqing ren xidu yi chengwei dangqian henduo guojia de shehui wenti.* 'Although youngsters' use of drugs has currently become a social problem in many countries.'
2	Sv	*(a) 老師教得很好,常使[Sv]我們複習,以便我們都會用新學到的生詞、句型等。 *Laoshi jiao de hen hao, chang shi* [Sv] *women fùxi, yibian women duhui yong xin xue dao de shengci, ju xing deng.* (b) 老師教得很好,常幫我們複習,讓我們都會用新學到的生詞、句型等。 *Laoshi jiao de hen hao, chang bang women fuxi, rang women duhui yong xin xue dao de shengci, ju xing deng.* 'The teacher teaches very well, who often helps us review things we learned so that we can use the newly acquired words, sentence patterns, etc.'
3	Madv	*(a) 每次選擇的時候,有好悶的感覺 *Mei ci xuanze de shihou, [Madv] you hao men de ganjue.* (b) 每次選擇的時候,都有好悶的感覺。 *Mei ci xuanze de shihou, dou you hao men de ganjue.* 'I feel so stuffy every time when I have to make choice.'
4	Mde	*(a) 我們唱歌要比誰唱[Mde]最好。 *Women changge yao bi shui chang* [Mde] *zui hao.* (b) 我們唱歌要比誰唱得最好。 *Women changge yao bi shui chang de zui hao.* 'We sing to compete for the best singer.'
5	Svs	*(a) 在美國,家庭主婦越來越少,職業婦女越來越豐富[Svs]。 *Zai meiguo, jiating zhufu yue lai yue shao, zhiye funu yue lai yue fengfu* [Svs]. (b) 在美國,家庭主婦越來越少,職業婦女越來越多。 *Zai meiguo, jiating zhufu yue lai yue shao, zhiye funu yue lai yue duo.* 'There are less housewives yet more professional women in the United States.'
6	Maux	*(a) 他不但[Maux]說兩個語言而且會跳舞! *Ta budan* [Maux] *shuo liang geyuyan erqie hui tiaowu.* (b) 他不但會說兩個語言而且會跳舞! *Ta budan hui shuo liang ge yuyan erqie hui tiaowu.* 'He can not only speak two languages but also dance.'
7	Sadv	*(a) 那時候,冬天好[Sadv]到了。每天的風景與變化對當時的我來說, 都很美麗。 *Na shihou, dongtian hao* [Sadv] *daole. Meitian de fengjing yu bianhua dui dangshi de wo lai shuo, dou hen meili.* (b) 那時候,冬天剛好到了。每天的風景與變化對當時的我來說,都很美麗。 *Na shihou, dongtian ganghao daole. Meitian de fengjing yu bianhua dui dangshi de wo lai shuo, dou hen meili.* 'At that time, winter had just arrived. The everyday changing scenery was very beautiful to me at that time.'

(continued)

Table 3 (continued)

Rank	Tag	Example sentence
8	Mv	*(a) 台灣的文化跟美國[Mv]起來完全不一樣。最大的差別是宗教的影響。 *Taiwan de wenhua gen meiguo* [Mv] *qilai wanquan bu yiyang. Zuida de chabie shi zongjiao de yingxiang.* (b) 台灣的文化跟美國比起來完全不一樣。最大的差別是宗教的影響。 *Taiwan de wenhua gen meiguo bi qilai wanquan bu yiyang. Zuida de chabie shi zongjiao de yingxiang.* 'The culture of Taiwan is completely different from that of the United States. The biggest difference is the influence of religion.'
9	Radv	*(a) 不同的利益團體對於環保與經濟發展的價觀非常不同,而且非常[Radv]互不信任。 *Butong de liyi tuanti duiyu huanbao yu jingji fazhan de jia guan feichang butong, erqie feichang* [Radv] *hu bu xinren.* (b). 不同的利益團體對於環保與經濟發展的價觀非常不同,而且互不信任。 *Butong de liyi tuanti duiyu huanbao yu jingji fazhan de jia guan feichang butong, erqie hu bu xinren.* 'Different interest groups have very different views on environmental protection and economic development, and they do not trust each other.'
10	Masp	*(a) 我是從日本來的。不過我想很多日本同學們介紹[Masp]日本。 *Wo shi cong riben lai de. Buguo wo xiang henduo riben tongxuemen jieshao* [Masp] *riben.* (b) 我是從日本來的。不過我想很多日本同學們介紹過日本。 *Wo shi cong riben lai de. Buguo wo xiang henduo riben tongxuemen jieshaoguo riben.* 'I am from Japan. But I think many Japanese classmates have introduced Japan.'
11	Rde	*(a) 所以我每天不但要很早地[Rde]起來,還要乖乖地聽旅館裡的人的話。 *Suoyi wo meitian budan yao hen zao de* [Rde] *qilai, hai yao guaiguai de ting luguan li de ren dehua.* (b) 所以我每天不但要很早起來,還要乖乖地聽旅館裡的人的話。 *Suoyi wo meitian budan yao hen zao qilai, hai yao guaiguai de ting luguan li de ren dehua.* 'Hence, I need to not only get up very early every day but also listen to staff in the hotel.'
12	Sconj	*(a) 只是[Sconj]這樣,他才能在挽救他的家庭告一段落後,進入人生並追求心裡上的啓示。 *Zhishi* [Sconj] *zheyang, ta caineng zai wanjiu ta de jiating gao yiduanluo hou, jinru rensheng bing zhuiqiu xinli shang de qishi.* (b) 只有這樣,他才能在挽救告一段落後,進入人生並追求心裡上的啓示。 *Zhiyou* [Sconj] *zheyang, ta caineng zai wanjiu ta de jiating gao yiduanluo hou, jinru rensheng bing zhuiqiu xinli shang de qishi.* 'Only in this way can he enter life and pursue the quest of spiritual enlightenment after saving his family.'

Note For each error tag, both erroneous sentence and suggested correction are provided. Erroneous sentences are labeled with * (a), and suggested corrections are labeled with (b)

Rankings of missing auxiliary and incorrect selection of adverbs were both one place higher in Lee, Chang, and Tseng. On the contrary, missing structural particles (Mde), incorrect selection of stative verbs (Ssv), and missing aspectual particles (Masp) ranked quite differently in this study and in Lee, Chang, and Tseng. Missing structural particles and incorrect selection of stative verbs ranked respectively the

Table 4 Comparisons between the top 10 error tags in this study and those in Lee, Chang, and Tseng (2016)

	This study			Lee et al. (2016)		
Rank	Tag	n	%	Tag	n	%
1	Sn	5556	11.51	Sv	3809	11.26
2	Sv	4181	8.66	Sn	2167	6.40
3	Madv	2923	6.06	Madv	1755	5.19
4	Mde	2478	5.13	Masp	1602	4.73
5	Svs	1926	3.99	Maux	1357	4.01
6	Maux	1741	3.61	Sadv	1168	3.45
7	Sadv	1618	3.35	Mde	1165	3.44
8	Mv	1474	3.05	Mv	1040	3.07
9	Radv*	1239	2.57	Rasp*	1003	2.96
10	Masp	1202	2.49	Svs	780	2.31

Note Tags with asterisks (*) are overlapped items in the two studies

fourth and the fifth in this study, yet they only ranked the seventh and tenth in Lee, Chang, and Tseng. In contrast, missing aspectual particles, the tenth common error tags in our study, ranked fourth in Lee, Chang, and Tseng. In addition, errors of redundant adverbs (Radv) ranked ninth in our study, whereas it was not included in the top 10 common error tags in Lee, Chang, and Tseng. The ninth common error tag in their study was redundant aspectual particles (Rasp), while this error type ranked the seventeenth out of the total errors in our study.

Findings of the comparison revealed that 90% of the top 10 error tags in our study overlapped with those in Lee et al. (2016), suggesting that these error types are indeed common in CSL learners' written production and should be further investigated in future research. Regardless of the 90% coverage of the top 10 error tags, rankings of the overlapped items in both studies were sometimes different, such as missing structural particles, incorrect selection of stative verbs, and missing aspectual particles. In addition, errors of redundant adverbs were also not listed in the top 10 error types in Lee, Chang, and Tseng. For these discrepancies, two possible explanations are provided. The first explanation lies in the different contexts where data in the two annotated corpora were gathered. Data in our study consisted of learners' writing assignments, while those in Lee et al. (2016) consisted of exam scripts. Exam scripts might better reflect learners' language proficiency in a way that no consultation of resources was allowed within the context of examination (Yang, 2003); nevertheless, the pressure learners experienced during the test might somewhat negatively influence their actual language use and thus cast doubt on the authenticity of the learner data. As a result, the contextual difference between the two sets of data might contribute to the different rankings of the top 10 error tags in the two studies.

Another explanation for the discrepancies lies in the proficiency levels of the learners in the two corpora. Our learner corpus contains writing assignments

produced by learners from the basic level to the advanced level. The learner corpus in Lee et al. (2016), however, comprises exam scripts that scored at least 3 or higher, which represent learners at the intermediate to advanced levels. The different range of language proficiency of the two datasets might be the cause of the ranking differences of some error types in the two studies. Some error types ranking higher in our study but lower in Lee, Chang, and Tseng might be errors that are more often made by learners at the lower level (e.g., missing structural particles and incorrect selection of stative verbs). On the other hand, errors ranked higher in their study yet lower in ours might be difficult for even higher-level learners to acquire. However, since the cross-level comparison was not the focus of the current study, the potential influence of proficiency levels on the two studies' different findings could not be confirmed. More research should be done to examine the distribution of each error tags at different proficiency levels.

4 Conclusion and Suggestions for Future Research

This study was set out to annotate errors in the Chinese Learner Written Corpus of National Taiwan Normal University and to present an overview of CSL learners' common error types. Manual annotation of the corpus yielded 48,266 errors distributed into 119 error tags, and more than 75% of the total errors belonged to incorrect selection or missing linguistic components. Among the four linguistic categories, around 80% of the total errors were caused by the misuse of word-level linguistic components, and about 50% of the total errors resulted from the misuse of nouns, action verbs, adverbs, and structural particles. Among these four commonly misused word classes, noun-based errors were mostly made by incorrect selection, whereas verb-based (including action verbs and stative verbs), adverb-based, and structural particle-based errors were committed in more diverse ways (i.e., incorrect selection, missing, and redundancy). Comparisons of the top 10 error tags in the current study and the previous one revealed that nine out of the top 10 error tags overlapped in the two studies, while rankings of the nine error tags in one study were somewhat different from the other. Regardless of the ranking difference, the 90% overlapping rate of the top 10 error tags in the two corpora suggests that these errors are indeed commonly misused items in CSL learners' writing and should be further investigated in future research.

Based on the findings yielded in this study, suggestions for future research are offered. First, CSL learners' use of nouns, verbs, adverbs, and structural particles should be extensively investigated, since these four word classes took up more than 50% of the total errors. Investigations on learners' use of these components might better reveal learners' difficulties in acquiring them and further provide useful information for effective material writing and teaching. In addition, since noun-based errors were mostly attributed to incorrect selection of other nouns, further examinations of CSL learners' perceptive and productive knowledge of synonymous nouns are also recommended to uncover how and why CSL learners made such type of

errors. In addition to targeting the four word classes, research on the common error types made by CSL learners at different proficiency levels is also suggested. Comparisons between findings in our study and those in the previous one have indicated that language proficiency might play a role in CSL learners' production of different error types; cross-level comparisons of common errors made by learners at different proficiency levels are hence recommended to discover whether longitudinal changes occur in CSL learners' making of errors.

References

Brooke, J., & Hirst, G. (2012). Measuring interlanguage: Native language identification with l1-influence metrics. In *Proceedings, 8th ELRA Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul.

Centre for English Corpus Linguistics. (2020). Learner Corpora around the World. Louvain-la-Neuve: Universite catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Chang, L.-P. (2017). TOCFL 學習者語料庫的偏誤標記 [The error annotation of TOCFL Learner Corpus]. In H.-J. Chen (Ed.), *Corpus and teaching Chinese as a second language* (pp. 159–196). Taipei: Taiwan Higher Education Press Co.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics, 5*, 161–170.

Diaz-Negrillo, A., & Dominguez, J. F. (2006). Error tagging systems for learner corpora. *Revista Española De Lingüistica Aplicada, 19*, 83–102.

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480.

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

Han, L. (2016). 基于HSK 动态作文语料库的连词"而且"偏误分析 [An error analysis of the conjunction *erqie* based on the HSK Dynamic Composition Corpus.] *Journal of Lanzhou Institute of Education*, *32*(3), 18–19.

Hu, R.-J. (2012). 试析留学生范围副词"都"的偏误 [An error analysis of *dou* used by foreign learners of Chinese]. *Foreign Language and Literature, 1*, 72–73.

Hsu, H.-M. (2011). 基于HSK动态语料库研究留学生作文中的偏误现象 [*A research on the errors of foreign students' composition: Based on the HSK Dynamic Corpus*]. Master's thesis: Liaoning Normal University.

James, C. (1998). Errors in language learning and use: Exploring error analysis. *London: Addison Wesley Longman.*

Jia, X.-H. (2007). 日本人汉语学习者语料库的建立与语法偏误分类法 [Establishment and grammar errors categories of learner corpus in Chinese learner of Japanese]. *Research of Applied Linguistic, 1*, 12–16.

Jin, L.-J. (2011). 韩国留学生使用介词"在"的偏误分析 [The error analysis of Korean student's use of the preposition *zai*]. *Journal of the Graduates at Sun Yat-Sen University (social Sciences), 32*(4), 6–11.

Lee, L. H., Chang, L. P., & Tseng, Y. H. (2016). Developing learner corpus annotation for Chinese grammatical errors. In *Proceedings of the 20th International conference on Asian language processing* (pp. 254–257). Tainan, Taiwan.

Li, J. (2013). 基于HSK 动态作文语料库的泰国学生"有" 字句的习得考察 [A study on the second language acquisition of *you*-sentence of Thai students based on HSK dynamic composition Corpus]. *Overseas Chinese Education*, (4), 388–397.

Lin, Y.-T., Chen, H.-J., & Wang, C.-C. (2014). 從學習者語料庫探究趨向補語[起來]之偏誤情形及教學建議 [A learner corpus-based study on Chinese directional complement *Qilai*]. *Journal of Chinese Language Teaching, 11*(4), 73–109.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge University Press.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics, 39*, 74–92.

Nemser, W. (1974). Approximative systems of foreign language learners. *Error analysis: perspectives on second language acquisition*, 55–63.

Nicholls, D. (2003). The Cambridge learner corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference* (vol. 16, pp. 572–581).

Pan, P., & Liu, L. (2006). 学习者语料库与外语教学研究 [Learner corpus and foreign language teaching research]. *Journal of Beijing International Studies University, 4*, 53–55.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching, 10*(1), 209–232.

Swanson, B., & Charniak, E. (2013). Extracting the native language signal for second language acquisition. In *Proceedings of the 2013 Conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 85–94).

Tono, Y. (2003). Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference* (pp. 800–809). Lancaster: University Centre for Computer Corpus Research on Language.

Wang, H.-C. (2010). 俄罗斯留学生使用"了"的偏误分析 [An error analysis on *le* used by Russian students]. *Chinese Language Learning, 3*, 99–104.

Wang, Y.-T., Chen, H.-J., & Pan, I.-T. (2013). 基於中介語語料庫之近義動詞混用情形調查與分析-以[幫],[幫助],[幫忙]及[變],[變得],[變成] 為例 [Investigation and analysis of Chinese synonymous verbs based on the Chinese learner corpus: Example of "bang", "bang-zhu", "bang-mang" and "bian", "bian-de", "bian-cheng"]. *Journal of Chinese Language Teaching, 10*(3), 41–64.

Wang, C., & Seneff, S. (2007). Automatic assessment of student translations for foreign language tutoring. In *Human language technologies 2007: The conference of the north American chapter of the association for computational linguistics; proceedings of the main conference* (pp. 468–475).

Wang, M., Malmasi, S., & Huang, M. (2015). The Jinan Chinese learner corpus. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 118–123).

Yang, H.-Z. (2003). 中国学习者英语语料库 *[Chinese Learner English Corpus]*. Shanghai Foreign Language Education Press.

Zang, W.-T. (2014). 对外汉语教学中带有标记词的强调句研究 [*Research on the emphatic patter with token words in teaching Chinese as a foreign language*]. Master's thesis, Senyan Normal University.

Zhang, B.-L. (2010). 回避与泛化——基于"HSK 动态作文语料库"的"把" 字句习得考察 [Avoidance and overgeneralization —an investigation of acquisition of the *ba*-Sentence based on the HSK Dynamic Composition Corpus]. *Chinese Teaching in the World, 2*, 263–278.

Zhang, R.-P. (2013). 三个汉语中介语语料库若干问题的比较研究 [A comparison study on some problems in three Chinese interlanguage corpora]. *Institute of Applied Linguistic, 3*, 133–140.

Explorations in Learner Corpora in Chinese and Related Languages

Cross-Referentiality of Multilingual Error Learner Corpora of Chinese, English and Japanese for Second Language Acquisition of Chinese Grammar

Hiroshi Sano, Yeong-il Yi, ChiaHou Wu, Go Inoue, YaMing Shen, Noboru Oyanagi, and Keiko Mochizuki

Abstract This paper presents an empirical study on the difficulties in learning Chinese as a second language based on learners' corpora written by native Japanese speakers at CEFR-based A2, B1 and B2 levels. The first part of this paper will discuss the procedures for how to collect learners' corpora, proofread, establish an error tag system and annotate errors. Next, we will focus on how the linguistic typology of a learner's L1 affects the acquisition of Chinese grammar. We will focus on three grammatical categories, (1) epistemic modality (realis/irrealis), e.g. an irrealis auxiliary Hui (會), (2) determiner phrase (DP), a determiner "One (一) + Classifier + Noun Phrase". Our findings are that even advanced Japanese L1 learners at CEFR B2 level tend to lack the irrealis auxliary Hui (會), the resultative complements and the determiner "One (一) + Classifier + Noun Phrase". On the other hand, English L1 Chinese learner corpus displays an overuse of "One (一) + Classifier", even in an

H. Sano · Y. Shen · N. Oyanagi · K. Mochizuki (✉)
Tokyo University of Foreign Studies, 3-11-1 Asahicho, Fuchu, Tokyo, Japan
e-mail: mkeiko@tufs.ac.jp

H. Sano
e-mail: sano@tufs.ac.jp

Y. Shen
e-mail: yamingshen2003@yahoo.co.jp

N. Oyanagi
e-mail: univ-oyanagi@nihon5ch.net

Y. Yi
Graduate School of Humanities, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
e-mail: nlp201@outlook.com

C. Wu
Software Development Department, IDEARUX Co., 2F., No, Sec. 6, Roosevelt Rd., Wenshan, Taipei City 116, Taiwan
e-mail: clothoray@gmail.com

G. Inoue
New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, UAE
e-mail: gogogo@kfactoryino.com

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_5

atelic context like a negative construction or a conditional construction where a “One (一) + Classifier” should not occur. This striking contrast between Japanese L1 and English L1 learners are due to the learner’s L1 typology. In Japanese, the tense system “Past -TA/Nonpast-RU” is grammatically obligatory and more prominent than the epistemic modality “realis/irrealis” and aspect “ perfective/nonperfective” system. In addition, the Japanese Noun Phrase has no determiner “a/an, the”, “this/that/my/your/~’s”. On the other hand, English L1 learners tend to treat the “One (一) + Classifier” as an article although it does not appear in an atelic event structure.

Keywords Online Dictionary of Misused Chinese based on Learners’ Corpora · Learner’s L1 typology · The irrealis auxliary Hui (會) · The resultative complements · The determiner “One (一) + Classifier + Noun Phrase” · Annotation system

Abbreviations

ACC	Accusative
ASP	Aspect
BA	*bǎ* Construction (direct object marker)
CAU	Causative suffix
CL	Classifier
DAT	Dative
DES	Desiderative form
DUR	Durative aspect
EXP	Experiential aspect
GEN	Genitive
INS	Instrumental
ITS	Intransitive suffix
LOC	Location
NEG	Negative
NML	Nominalizer
NOM	Nominative
NONPAST	Nonpast
PAST	Past
PFV	Perfective aspect
PP	Pragmatic particle
PSS	Passive suffix
POL	Polite suffix
Q	Question
QT	Quotative particle
SE	Sentence extender
SFP	Sentence final particle
SFX	Suffix

TOP Topic marking particle
TRS Transitive suffix

1 Cross-Referential Learners' Corpora of English, Chinese and Japanese

First, we introduce the benefit of using Cross-Referential Learners' Corpora of English, Chinese and Japanese to reveal various types of L1 transfer. For example, L1 Japanese and L1 English learners of Chinese show a difference in the acquisition of the Chinese determiner "One + Classifier + Noun Phrase" which is obligatory in a telic sentence. L1 English learners acquire "One + Classifier + Noun Phrase" correctly while L1 Japanese advanced learners show difficulty in acquiring the determiner usage of "One + Classifier + Noun Phrase". We suggest that the linguistic cognitive typology in L1 affects the second language acquisition of the aspectual system in the target language: both English and Chinese are "Bounded" oriented languages as Li and Thompson (1981), Tai (1984), Tai (2003) while Japanese is an "Unbounded" type language in terms of aspectual boundedness as discussed in Ikegami (1991), Furukawa (2001), Mochizuki (2004), Mochizuki (2007), Mochizuki (2009), Shen & Mochizuki (1997), Shen (2009) and Newbery-Payton & Mochizuki (2020).

This paper explores, through the use of learner corpora of Chinese by L1 Japanese and L1 English, how differences in learners' native languages affect second language acquisition.

(1) Tokyo University of Foreign Studies "Learners' Error Corpora of Japanese/English/Chinese Searching Platform" https://corpus.icjs.jp/

These three error tagged learner corpora were developed through collaboration with universities outside of Japan. The corpus in (1) is comprised of the sections listed below in (2), (3) and (4). Some of the essays are taken from different university classes and are therefore uncontrolled in terms of content. The majority of essays, however, are translations of a text titled "Memories of Study Abroad in Shanghai". The Japanese, Chinese and English versions of the text are provided in the appendix. The corpus also records learner data, including learners' native language, length of study and any language proficiency test scores.

(2) Tokyo University of Foreign Studies International Center of Japan Studies "The Learners Language Corpus of Japanese and Online Error Dictionary" https://corpus.icjs.jp/corpus_ja/

This learner corpus was developed in collaboration with the University of Leeds, Peking University, Shanghai International Studies University and Akita International University. It contains Japanese essays written by learners at these institutions whose native language is English or Chinese. A total of 129 essays were corrected and

error tagged by the third author of this paper. The majority of the data comes from students studying Japanese as an additional foreign language at Peking University and Japanese majors at Shanghai International Studies University. Learners translated the Chinese text into Japanese. Use of a dictionary was permitted.[1]

(3) Tokyo University of Foreign Studies Learners' Error Corpora of English Searching Platform https://corpus.icjs.jp/corpus_eng/index.php

(3) is a corpus of English learners whose native language is Japanese or Chinese. 284 corrected and error tagged essays are included, together with learner information. The data from Japanese-native speakers include homework tasks written by first year English majors at the Tokyo University of Foreign Studies, as well as translations of the "Memories of Study Abroad in Shanghai" text. 24 native speakers of English from America, the UK, Australia and Singapore corrected the essays.

Standardized methods of correction and tagging were decided at weekly meetings. The data from Chinese native speakers were collected through collaboration with Shanghai International Studies University and National Taiwan Normal University.

(4) Tokyo University of Foreign Studies and National Taiwan Normal University (henceforth, The TUFS_NTNU) Learner Error Corpus Learners' Error Corpora of Chinese Searching Platform https://corpus.icjs.jp/corpus_ch/index.php

The Chinese learners' corpus consists of 369 essays by Chinese majors at the Tokyo University of Foreign Studies. The essays are corrected, error tagged, and include learner information. Data from a wide range of learners are included, from 2nd year students at low to intermediate proficiency levels to 4th and 5th year students at Cefr B2 level with one-year study abroad experience. The data includes homework tasks as well as translations of the Chinese version of the "Memories of Study Abroad in Shanghai" task as the appendix of this chapter shows. Both types of tasks allowed the use of a dictionary. A total of 25 native speakers of Chinese (university staff and graduate students) corrected, error tagged and checked essays at weekly meetings under the guidance of the second author of the current paper.

In addition, while the data cannot be made public because learners' consent has not been obtained, data obtained from National Taiwan Normal University has also been corrected, error tagged and used for research purposes. The data consists of essays written by native speakers of English as part of the Test of Chinese as a Foreign Language (TOCFL), a Chinese proficiency test used in Taiwan.

2 Cross-Referentiality of Multilingual Learner Corpora

The Japanese, English and Chinese learner corpora described above allow comparison of 6 combinations of native language/language of study as shown in Fig. 1.

[1] During the course of the project, it became clear that while use of a dictionary greatly influences vocabulary production, it does not have a large influence on the production of grammatical forms.

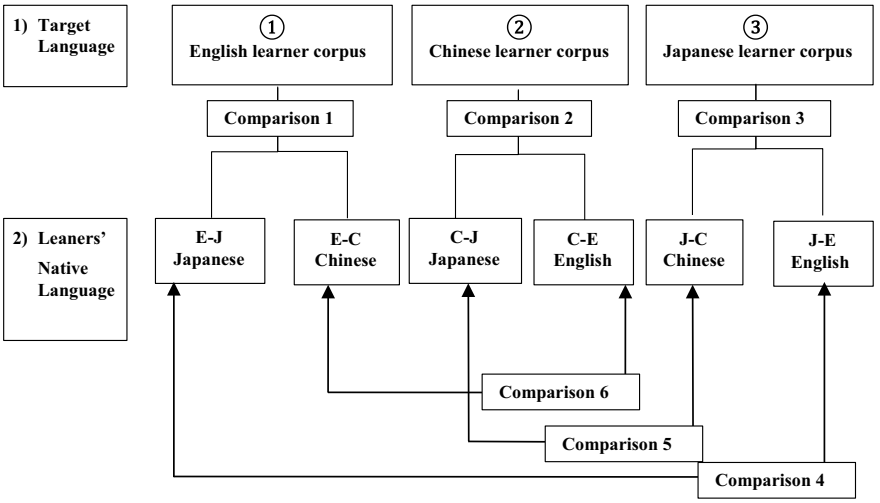


Fig. 1 Patterns of comparison between combinations of language of study and native language

This is useful for second language acquisition research that considers how patterns of acquisition of grammatical forms in each language differ depending on learners' native language.

To give one example, in order to examine what features of Japanese affect the acquisition of Chinese by Japanese-native speakers, the relevant pair is (2) in Fig. 1. We explore this comparison between Japanese and English-native speakers of Chinese later in Sect. 4.

3 Procedures

3.1 The Learner Error Corpora of Chinese

The characteristics of the data set of the Learner Corpus of Chinese at Tokyo University of Foreign Studies are as follows as we showed in Mochizuki et al. (2015) (Table 1).

These compositions are proofread by native speakers of Chinese with Ph.D students in linguistics/language education with sufficient experience in teaching Chinese at university level. Proofread compositions clearly indicate errors and corrections so that the errors can be identified within the respective sentences.

The TUFS_NTNU Learner Error Corpus includes learner's information as shown in Table 2.

Table 1 TUFS learner corpus of Chinese collected in May 2013–August 2014

Academic year	Level chinese major students	Number of essays	Approximate number of words	Number of students
2013	Advanced (4th year)	95	45,500	35
	Intermediate (2nd/3rd year)	132	51,200	58
2014	Advanced (4th year)	21	12,500	23
	Intermediate (2nd/3rd year)	34	25,100	69
Total		282	134,300	185

Table 2 Example of learner's profile

1	Learner's ID	Th_Ch_001
2	Name	Tokyo Taro
3	Major	Chinese
4	Year	3
5	Gender	male
6	Age	21
7	Nationality	Japan
8	Residential History	Canada 4–9; Japan 0–4,9–21
9	Native Language	Japanese
10	Language of Education	Japanese, English
11	Length of Chinese study	3 years and 2 months
12	Institution	Tokyo University of Foreign Studies
13	Study Abroad Experience Institution / Period	Mandarin Center, National Taiwan Normal University, August1-31. 2014
14	Speaking with my family	Japanese
15	Speaking with friends	Japanese
16	Language used in Elementary School	5–9 English, 9–12 Japanese
17	Language used in Junior High School	Japanese, English
18	Language used in Senior High School	Japanese, English
19	Test of Chinese as a Foreign Language (TOCFL)	Band B (2014)
20	Hanyu Shuiping Kaoshi (HSK)	Band 5 (2012)
21	English TOEFL(iBT)	108 (2013)
22	TOEIC	955 (2012)
23	IELTS (academic)	8.0 (2013)

The TUFS_NTNU Learner Error Corpus has four key features:

(1) compositions are written by experienced learners majoring in Chinese in Japan,
(2) compositions go through an appropriate proofreading process conducted by university teachers,
(3) errors and corresponding corrections are recorded,
(4) the detailed profiles of the learners are also recorded.

3.2 Error Tag Categories

There are two tag categories for misuse: Error and Modify. "The Error tag" indicates grammatical errors while "the Modify tag" indicates inappropriate use of expressions ('expression' tag), punctuation and Chinese characters as shown in Fig. 2.

The Error tag consists of the following four subcategories: Replace, Delete, Insert and Move. The Replace tag indicates the need to replace an Error with another correct expression. The Delete tag indicates that deleting an Error will lead to a correct expression. The Insert tag indicates that inserting new expressions will lead to a correct expression. The Move tag indicates a word order Error.

The Modify tag consists of the following three subcategories: Expression, Punctuation and Chinese character. The Expression tag indicates that it is preferable to use another expression or that the misuse cannot be categorized as any one specific error. The Punctuation tag indicates the need for correction in view of the style of writing. The Chinese character tag indicates the misuse of a Chinese character. As

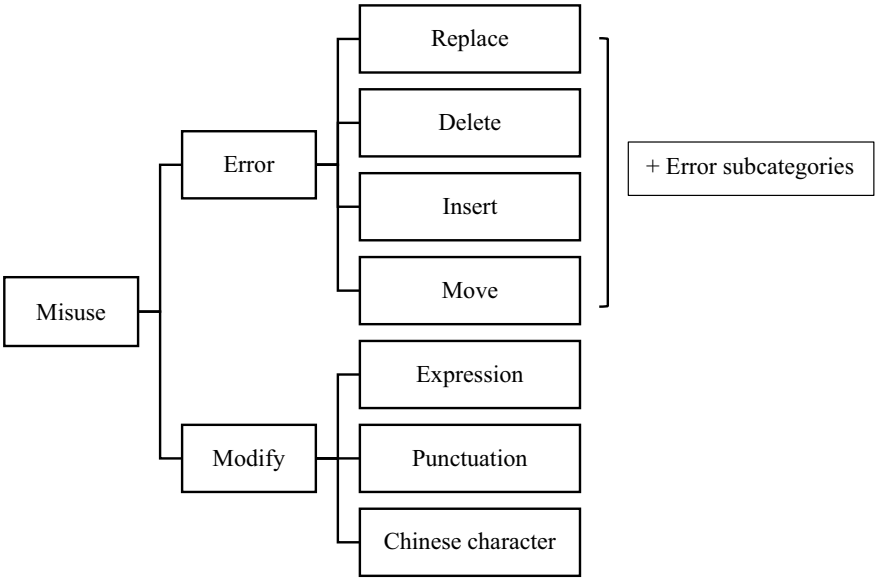


Fig. 2 Misuse tag system: classification and in-text marking of syntactical, lexical, stylistic, rhetorical and notational misuses

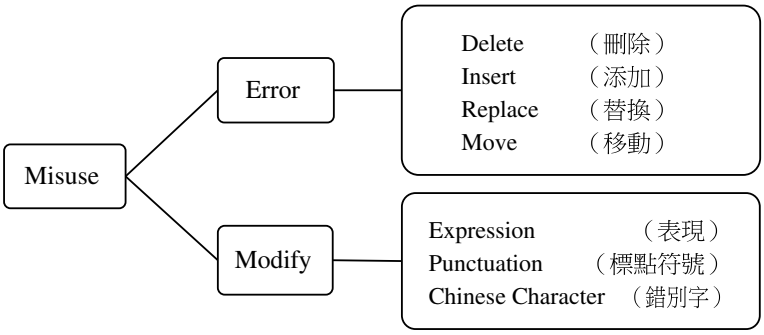


Fig. 3 Tag system: tag list in chinese based on subcategories of misuse

subcategories of the Error tag, we have designed the 74 tags as shown in (1) referring to the grammatical system in *Modern Chinese Grammar* 张斌 (Zhang Bin) and 齐沪杨 (Qi Huyang) et al., 2002: 273–467) (Fig. 3 and Table 3).

3.3 Method of Proofreading and Annotation

We use the 'TUFS_TNR_Chinese Writing Correction 2014' and 'TUFS_TNR_Chinese Error Corpus Tagger 2014' (2014) tools developed by 于康 (Yu Kang) and 田中良 (Ryo Tanaka) for proofreading and annotation. The procedures are as follows. First, compositions written by learners in a WORD file are converted to text files. Next, Errors and the corresponding corrections are added to the composition texts using the 'TUFS_TNR_Chinese Writing Correction 2014' system. The following Fig. 4 is an example of proofreading using 'TUFS_TNR_Chinese Writing Correction 2014'.

The 'TUFS_TNR_Chinese Writing Correction 2014' system displayed in Fig. 4 has two windows: the left window displays the composition text and the right window displays corrections. Each correction in the right window and its corresponding Error expression in the left window are marked up in the same color for better visibility.

For annotation, 'TUFS_TNR_Chinese Writing Correction 2014' and 'TUFS_TNR_Chinese Error Corpus Tagger 2014' (2014) enable free creation of tags and the displaying of a tag list underneath the composition text as shown in Fig. 5.

The first step in annotating a composition is to designate the region of each misused expression in the composition text. The second step is to choose one of 'Replace 替換, Delete 刪除, Insert 添加, Move 移動, Expression 表現, Punctuation 標點符號, Chinese Character 錯別字' and click on the appropriate button. This procedure enables annotations to be made automatically. The third step is to choose one of the Error subcategories, e.g. 'Resultative Complement 結果補語'. This click-annotation system greatly reduces the burden of annotation. 'TUFS_TNR_Chinese Writing Correction 2014' also has the function to convert annotated data into XML data (Fig. 6).

Table 3 Subcategories of error

	大分類categories	小分類 subcategories
1	名詞 nouns	時間名詞 time nouns, 處所名詞 place nouns, 方位詞 locative nouns
2	數詞 numerals	
3	量詞 classifiers	
4	動詞 verbs	狀態動詞 stative verbs, 動作動詞 action verbs, 存現動詞 existential verbs, 關係動詞 copular verbs, 能願動詞 auxiliary verbs, *趨向動詞* directional verbs, 使令動詞 causative verbs
		及物動詞 transitive verbs, 不及物動詞 intransitive verbs, 雙賓動詞 double object verbs
		重疊動詞 verb reduplication
5	形容詞 adjectives	
6	副詞 adverbs	程度副詞 degree adverbs, 範圍副詞 scope adverbs, 時間副詞time adverbs, 情態副詞 modal adverbs, 否定副詞 negative adverbs, 語氣副詞 tone adverbs, 關聯副詞 correlative adverbs
7	代詞 pronouns	人稱代詞 personal pronouns, 指示代詞 demonstrative pronouns, 疑問代詞 interrogative pronouns
8	連詞 conjunctions	
9	介詞 prepositions	
10	助詞 particles	結構助詞 structural particles, 時態助詞 aspectual particles, 時制助詞 tense particles, 比況助詞 comparative particles, 表數助詞 quantitative particles, 列舉助詞 relational particles, 語氣助詞 modal particles, 其他助詞 others
11	短語 phrases	量詞短語 classifier phrases, 方位短語 locative phrases, 介詞短語 prepositional phrases, **"的"**字短語 "de" phrases
12	主語 subjects	
13	賓語 objects	雙賓語 double object
14	補語 complements	結果補語 result complements, 趨向補語 direction complements, 可能補語potential complements, 程度補語 degree complements, 情態補語 state complements, 數量補語 quantity complements, 介詞短語補語 location complements
15	疑問句 questions	是非問句 yes–no questions, 特指問句 wh- questions, 選擇問句 disjunctive questions, 正反問句 A-not-A questions

(continued)

Table 3 (continued)

	大分類categories	小分類 subcategories
16	句式 construction	主謂謂語句 SV construction, **“把”**字句 “ba” construction, **“被”**字句 “bei” construction, 連動句 serial-verb construction, 強調句emphatic construction, 兼語句 pivotal construction, 使役句 causative construction, 存現句existential construction, 比較句 comparative construction, **“連”**字句 “lian” construction
17	複句 complex sentences	並列複句 coordinate relation: 承接複句 progressive relation, 遞進複句 successive relation, 選擇複句 alternative relation, 注解複句
		偏正複句 subordinaterelation: 因果複句 causativerelation, 條件複句 conditionalrelation, 轉折複句 concessionrelation,讓步複句 hypotheticalrelation, 目的複句 purposiverelation

Fig. 4 Proofreading system

4 Cross-Linguistic Analysis of Errors

We will discuss two significant Error types in two learners’ corpora by comparing the TUFS-NTNU corpus written by Japanese-native speakers at TUFS with the TOCFL learners’ corpus of Chinese written by English-native speakers (henceforth, TOCFL corpus). (張莉萍 Chang Li-Ping, 2013) (Table 4).

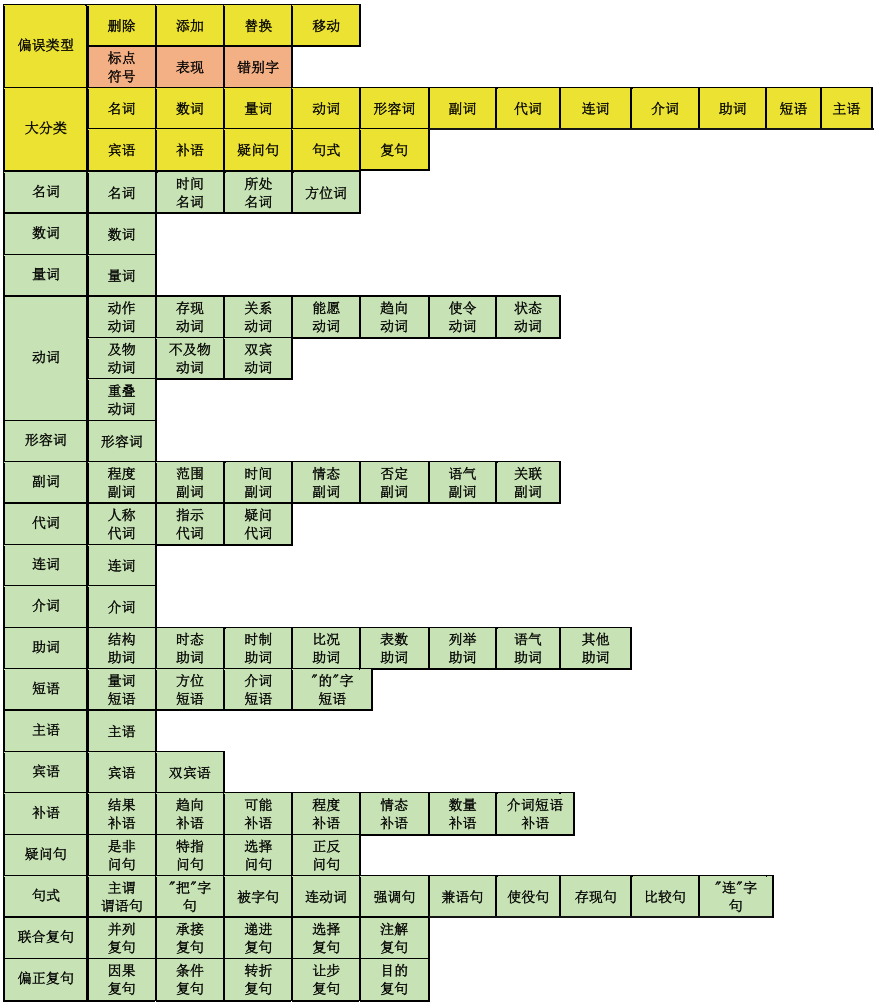


Fig. 5 Annotation system: tag buttons

4.1 Acquisition of Classifier Phrase (量詞短語) "One (一) + Classifier (量詞)"

Mochizuki et al. (2015) discusses one of the most significant error categories observable in the TUFS-NTNU Corpus is the lack of "One (一) + Classifier (量詞)" while the TOCFL (trial version) Corpus displays an overuse of "One (一) + Classifier (量詞)".張莉萍 Chang Li-Ping (2014:68) also indicates the same contrast between English-Native learners and Japanese-Native learners.

Table 5 compares the frequency of "One (一) + Classifier (量詞) '-ge 個'" in the TUFS-NTNU Corpus and the TOCFL (trial version) Corpus.

Fig. 6 Digitization framework (XML Data)

Table 4 The TOCFL English-Native Learners' Corpus of Chinese

TOCFL (CEFR)	Number of compositions	Number of chinese characters	Number of students
基礎 (A2)	223	119,971	223
進階 (B1)	344	31,852	344

Table 5 shows an interesting contrast in the frequency of “一個” between The TOCFL English- Native Learners' Corpus and The TUFS-NTNU Japanese-Native Learners' Corpus. The TOCFL English-Native Learners' Corpus displays a higher frequency than The TUFS-NTNU Japanese- Native Learners' Corpus. Upon conducting a chi squared test, a significant difference between the data sets was discovered (0.1%, $x^2 = 150.03$, $p = 0.000$).

Table 5 The Frequency of "One (一) + Classifier '-ge 個'"

	CEFR Level	Number of Chinese characters	Occurrence of "一個"
The TOCFL Trial version English-Native Learners' Corpus	B1	119,971	586 tokens
	A2	31,852	159 tokens
	Total	151,823	745 tokens
			1,490 Chinese characters
The TUFS-NTNU Japanese-Native Learners' Corpus	A2-B2	134,094	385 tokens
			770 Chinese characters

4.2 Lack of "One (一) + Classifier": Japanese Learners

Let us examine the lack of "One (一) + Classifier(量詞)" in The TUFS-NTNU Japanese-Native Learners' Corpus. The following examples (5) to (12) show that each sentence lacks the bracketed "One (一) + Classifier" in The TUFS-NTNU Japanese-Native Learners' Corpus. There are almost no examples of overuse of "One (一) + Classifier" in The TUFS-NTNU Japanese-Native Learners' Corpus.

(5) Copula "是 Shi" Construction:

'Topic(Old Information) + "是 Shi"+ Comment(New Information)'

a. 我 認爲 這 是 **(一種)** 有益 的 愛好。
Wǒ rènwéi zhè shì (yì zhǒng) yǒuyì de àihào.
I think this be 1 CL useful NOM interest

b. 但是， 生 孩子 是 **(一件)** 不 簡單 的 事。
Dànshì, shēng háizi shì (yī jiàn) bù jiǎndān de shì.
but give birth to son(or daughter) be 1 CL NEG easy NOM thing

c. 從前 我 去 過 京都， 京都 是 **(一個)** 很 美麗 的 地方。
Cóngqián wǒ qù guo Jīngdū, Jīngdū shì (yí ge) hěn měilì de difang.
Before I go EXP Kyoto Kyoto be 1 CL very beautiful NOM place

d. 但是， 找 到 好 工作 並 不 是 **(一件)** 好 的 事情。
Dànshì, zhǎo-dào hǎo gōngzuò bìng bú shì (yī jiàn) hǎo de shìqíng.
but look for- find good work and NEG be 1 CL good NOM thing

e. 小學生 跟 他們 交流 是 **(一個)** 好 機會，
Xiǎoxuéshēng gēn tāmen jiāoliú shì (yí ge) hǎo jīhuì,
primary school students with they exchange be 1 CL good opportunity

但是， 孩子們 聽 得 懂 他們 的 課 嗎？
Dànshì, háizimen tīng de dǒng tāmen de kè ma?
but son(or daughter) listen to DE understand they GEN class Q

f. 現在， 環境 問題 是 世界 的 **(一個)** 很 大 的 課題。
Xiànzài, huánjìng wèntí shì shìjiè de (yí ge) hěn dà de kètí.
now environment issue be world GEN 1 CL very big NOM problem

(6) Existential "You 有" Construction

a. 東大和 有 **(一個)** 很 大 的 公園---
Dōngdàhé yǒu (yí ge) hěn dà de gōngyuán ---
Higashiyamato have 1 CL very big NOM park

東大和南公園, 附近 也 有 **(一條)** 小 河。
Dōngdàhénángōngyuán, fùjìn yě yǒu (yī tiáo) xiǎo hé.
Higashiyamato minami Park nearby also have 1 CL small river

b. 這 台 電視 還有 **(一個)** 功能, 那 就是 聽 音樂。
Zhè tái diànshì háiyǒu (yí ge) gōngnéng, nà jiùshì tīng yīnyuè.
this CL TV also have 1 CL function that exactly listen to music

(7) Perfective Construction with "-le 了"

這幾年, 一位 很 有名 的 漫畫家 畫 了
Zhè jǐ nián, yī wèi hěn yǒumíng de mànhuà jiā huà le
recent years 1 CL very famous NOM comic artist draw PFV

(一部) 跟 帶廣 的 挽曳賽馬 有關 的 漫畫。
(yí bù) gēn Dàiguǎng de wǎn yè sàimǎ yǒuguān de mànhuà.
1 CL with Obihiro GEN Banee-keiba relate to NOM comic

(8) "Give" Construction and "Become" Construction

a. 比如， 開始 工作 掙 錢 以後，
Bǐrú, kāishǐ gōngzuò zhèng qián yǐhòu,
for example start work earn money after

我 想 送 給 父母 **(一份)** 禮物， 例如， 海外旅行。
wǒ xiǎng song gěi fùmǔ (yī fèn) lǐwù, lìrú, hǎiwài lǚxíng.
I want to give for father and mother 1 CL present for example oversea travel

我 也 在 留心 這 些 事情，
Wǒ yě zài liúxīn zhèxiē shìqíng,
I also DUR attentive this CL thing

希望 能 成為 **(一個)** 很 好 的 領導!!
xīwàng néng chéngwéi (yí ge) hěn hǎo de lǐngdǎo!!
hope to can become 1 CL very good NOM leader

(9) Presentative Construction

最近 他 在 車站 附近 開 了 **(一家)** 中餐館。
Zuìjìn tā zài chēzhàn fùjìn kāi le (yì jiā) zhōngcānguǎn.
resently he at station nearby open PFV 1 CL Chinese food restaurant

(10) Resultative/Directional Verb Compound

其他 同學 也 舉出 **(一些)** 有意思 的 食物，
Qítā tóngxué yě jǔ-chū (yìxiē) yǒu yìsi de shíwù,
other schoolmate also give-take out 1 CL interesting NOM food

比如 納豆， 豆漿 等等。
bǐrú nàdòu, dòujiāng děngdeng.
for example nattoo soy milk and so on

(11) "Modifier +的 DE+ Noun"

a. 原來 有 很 多 溫泉 的 日本 的 **(一個)** 特色
Yuánlái yǒu hěn duō wēnquán de Rìběn de (yí ge) tèsè
originally have very many hot spring GEN Japan GEN 1 CL characteristic

就是 飯店 旅館業 很 發達。
jiùshì fàndiàn lǚguǎnyè hěn fādá.
quite right hotel business very developed

b. 在 我 的 印象 裡 很 深 的 **(一件)** 事
Zài wǒ de yìnxiàng li hěn shēn de (yí jiàn) shì
at I GEN impression inside very deep NOM 1 CL thing

是 小學 5年級 的 時候 媽媽 幫助 我 練習 跑步。
shì xiǎoxué 5 niánjí de shíhou māma bāngzhù wǒ liànxí pǎobù.
be primary school 5th grade NOM time mother help I practice run

c. 去年 網絡 上 的 **(一篇)** 文章
Qùnián wǎngluò shang de (yì piān) wénzhāng
last year internet on GEN 1 CL essay

“中國 女性 和 日本 女性 的 一生” 引人註目。
“Zhōngguó nǚxìng hé Rìběn nǚxìng de yìshēng” yǐnrén zhù mù.
Chinese woman and Japanese woman GEN all one’s life gather attention

(12) ‘Source’ with New Information:

那個 名字 來源 于 **(一條)** 從 南 到 北 延伸 的 坡道。
Nàge míngzi láiyuán yú (yì tiáo) cóng nán dào běi yánshēn de pōdào.
that CL name originate for 1 CL from south to north extend NOM slope

The reason why it is difficult for Japanese learners of Chinese to learn the principle of “One (一) + Classifier” is because Japanese grammar is insensitive to ‘Boundedness’ (有界性) which controls the occurrence of “One (一) + Classifier”.

Shen (1995)’s “bounded/unbounded” theory can explain why “One (一) + Classifier” is necessary in the following constructions:

(13)

a. Perfective Construction with “-le 了”
b. GOAL in “Give” Construction
c. “Become” Construction,
d. Presentative Construction
e. Resultative/Directional Verb Compound

since all cases in (13) have “telicity”, the subcategory of “bounded” concept in the temporal structure.

“One (一) + Classifier” also often appears after “是 Shi” / “You 有” constructions. Both constructions have the following informational structure:

(14)

“是 Shi”/“You 有” Construction	Topic	“是 Shi”/“You 有”	“一 + Classifier” NP
(1) Informational structure	Old information		New information
(2) Boundedness			Bounded

It is supposed that the NP with new information is a bounded entity because the NP with new information is a focus in terms of cognition.

Shen (沈家煊) (1995) discusses the interaction between “One (一) + Classifier” and the concept of ‘bounded’ and ‘unbounded’ events. Shen (1995) indicates that a “One (一) + Classifier” is necessary before a ‘bounded’ Noun Phrase(NP) in ‘Telic’ events as follows:

(15) indirect Object in a Move Construction:

a. 盛 碗 里 两 条 鱼。
Chéng wǎn li liǎng tiáo yú.
fill bowl in 2 CL fish

b. *盛 碗 里 鱼。
Chéng wǎn li yú.
fill bowl in fish

(16) Resultative Object (結果賓語)

a. 蚊子 叮 了 小王 两 个 大包。
Wénzi dīng le Xiǎo Wáng liǎng ge dà bāo.
mosquito sting PFV Xiao Wang 2 CL big boil

b.*蚊子 叮 了 小王 大 包。
Wénzi dīng le Xiǎo Wáng dà bāo.
mosquito sting PFV Xiao Wang big boil

(17) Resultative Complement (結果補語)

a. 打 破 两 块 玻璃。
Dǎ pò liǎng kuài bōli.
beat-break 2 CL glass

b*打 破 玻璃。
Dǎ pò bōli.
beat-break glass

(18) Directional Complement(趨向補語)

a. 飞 进来 一 个 苍蝇。
Fēi jìnlái yí ge cāngyíng.
fly- enter-come 1 CL fly

b. *飞 进来 苍蝇。
Fēi jìnlái cāngyíng.
fly- enter-come fly

(19) Verb+ “-le 了” construction

a. 吃 了 一 个 苹果。
Chī le yí ge píngguǒ.
eat PFV 1 CL apple

b. *吃 了 苹果。
Chī le píngguǒ
eat PFV apple

4.3 Overuse of "One (一) + Classifier (量詞)" by English-Native Learners

We find the reverse phenomenon in The TOCFL English-Native Learners' Corpus: the overuse of "One (一) + Classifier". The following examples (20) to (27) show that the bracketed "One (一) + Classifier" should be deleted.

(20) Conditional:

有 什麼 問題 就 跟 我 打 **(一通)** 電話 吧！
Yǒu shénme wèntí jiù gēn wǒ dǎ (yì tōng) diànhuà ba!
have any questions then with I call 1 CL phone SFP

(21) Plan:

我們 游 完 泳 我 **計畫** 我們 去 電影院 看 **(一部)** 電影。
Wǒmen yóu-wán-yǒng wǒ jìhuà wǒmen qù diànyǐngyuàn kàn (yí bù) diànyǐng.
we swim-finish-swim I plan we go cinema see 1 CL movie

(22) Potential:

a. 我們 也 可以 去「西門町」 看 電影， 打 撞球，
Wǒmen yě kěyǐ qù「Xīmén Dīng」kàn diànyǐng, dǎ zhuàngqiú,
we also can go Ximen town see movie play billiards

或 去 **(一個)** 茶店 談天 說笑。
huò qù (yí ge) chá diàn tántiān shuōxiào
or go 1 CL teahouse chat and laugh

b. 你 看 我 已經 可以 用 中文 寫 **(一封)** 信，…
Nǐ kàn wǒ yǐjīng kěyǐ yòng Zhōngwén xiě (yī fēng) xìn,
you look I already can use Chinese write 1 CL letter

(23) Future Activity:

我 記得 你 說 過 你 喜歡 丟 飛盤，
Wǒ jìde nǐ shuō guo nǐ xǐhuan diū fēipán,
I remember you say EXP you like to throw flying disc

所以 我 會 把 **(一張)** 飛盤 帶來。
suǒyǐ wǒ huì bǎ (yì zhāng) fēipán dàilái.
Therefore I will BA 1 CL flying disc bring-come

(24) Topic Noun in "是 Shi" construction:

我 媽媽 上 上 個 週末 來 台灣 看 我。
Wǒ māma shàngshang ge zhōumò lái Táiwān kàn wǒ.
I mother the week before last come Taiwan visit I

我們 去 的 **(一個)** 地方 是 花蓮。
Wǒmen qù de (yí ge) dìfang shì Huālián
we go NOM 1 CL place be HuaLian

(25) "When" Clause: Old Information

你 開 **(一個)** 慶祝會 的 時候
Nǐ kāi (yí ge) qìngzhù huì de shíhou
you hold 1 CL celebration NOM time

我 不 能 參加 是 因為 我 在 外國 工作。
wǒ bù néng cānjiā shì yīnwèi wǒ zài wàiguó gōngzuò.
I NEG can attend be because I in foreign country work

(26) Negation:

a. 我 在 台北 沒有 發生 **(一個)** 大 問題，…….
Wǒ zài Táiběi méiyǒu fāshēng (yí ge) dà wèntí
I in Taipei NEG occur 1 CL big problem

b. 他們 有 一個 農場， 我 去 他們 的 家 以前，
Tāmen yǒu yí ge nóngchǎng, wǒ qù tāmen de jiā yǐqián,
they have 1 CL farm I go they GEN house before

還 沒 去 **(一個)** 農場…
hái méi qù (yí ge) nóngchǎng
yet NEG go 1 CL farm

(27) Missed Action:

今天 他 不但 忘 了 帶 手機， 也 忘 了 帶 **(一瓶)** 水。
Jīntiān tā búdàn wàng le dài shǒujī, yě wàng le dài (yī píng) shuǐ.
Today he not only forget PFV bring mobile phone also forget PFV bring 1 CL water

The interlanguage of Chinese created by English-native speakers displays the following incorrect overgeneralization:

(28) Overgeneralization by English-native learners of Chinese a/an NP = "一 + Classifier" NP.

Shen (1995)'s "bounded/unbounded" theory can also explain why "One (一) + Classifier" cannot appear in (29) to (31): all cases express atelic events and an entity in an atelic event should be unbounded. Shen (1995) indicates that a "One (一) + Classifier" cannot appear in the following atelic structures.

(29) Verb Reduplication(动词重叠式):

a. (*) 今天 要 谈谈 两 个 问题。
Jīntiān yào tán-tan liǎng ge wèntí.
today will discuss-discuss(=discuss a little) 2 CL problem

b. * 星期天 在 家 洗洗 一 件 衣服。
Xīngqītiān zài jiā xǐ-xi yí jiàn yīfu.
Sunday at house wash-wash (=wash a little) 1 CL clothes

(30) Durative Aspect Marker "-Zhe 着"

a. Progressive Aspect: *他 正 吃 着 三 碗 饭。
Tā zhèng chī zhe sān wǎn fàn
he DUR eat DUR 3 CL cooked rice

b. Resultative State: *山 上 架 着 两 门 炮。
Shān shang jià zhe liǎng mén pào
mountain on put up DUR 2 CL cannon

(31) Negation:

a. *今天 不 谈 两 个 问题。
Jīntiān bù tán liǎng ge wèntí
today NEG discuss 2 CL problem

b.*这个月 不 演 三 场 电影。
Zhègeyuè bù yǎn sān chǎng diànyǐng
this month NEG perform 3 CL movie

4.4 Comparative Analysis of Error Types by Japanese Learners and English-Native Learners

The contrast between the lack of "One (一) + Classifier" in The TUFS-NTNU Japanese-Native Learners' Corpus and the overuse of "One (一) + Classifier" in The TOCFL English-Native Learners' Corpus suggests a difference in Noun Phrase Structures in Chinese, English, and Japanese.

Japanese syntax has no 'functional category', therefore there is no syntactic node (i.e. 'determiner') to accommodate a constituent like "a/an, the" while English has 'determiner' as Fukui (1995) , Huang. Li and Simpson (2014) propose. This syntactic

difference between English and Japanese causes the contrast between the lack and the overuse of "One (一) + Classifier" in Japanese-native learners and English- native learners.

In addition, Ikegami (1981, 1991, 2007), Kageyama (1997, 2002, 2021a, b) and Kageyama and Jacobsen (2016) suggest that Japanese is an "unboundedness-oriented" "less-individualization" type language in terms of having no grammatical category of number, ellipsis of subject/object, and no determiner node in Noun Phrase like "a/the/this/my". This "unboundedness-oriented" and "less-individualization" feature is reflected in second language acquisition of Chinese and English by Japanese learners. Since Japanese grammar has no syntactic strategy to individualize an entity/event, it is very difficult to acquire both the principle of "One (一) + Classifier" NP which appears in a bounded/individualized noun, and the usage of the articles "a/an, the" in English. According to "TUFS_ NTNU Learners' Corpus of English", the most frequent Error category in the Japanese-native learner's corpus is articles "a/an, the" as shown in "TUFS_ NTNU Learners' Corpus of English": https://corpus.icjs.jp/corpus_eng/index.php.

On the other hand, English is a "boundedness-oriented" "high-individualization" type language in terms of having an obligatory grammatical category of number, determiner node, and an obligatory subject/object. The reason why the English-native TOCFL corpus displays an overuse of "One (一) + Classifier" is because the principle of individualizing a noun is different between English and Chinese. Chinese cannot individualize a noun in an atelic unbounded event like a future event, a potential, a negation, a missed action or a conditional. On the other hand, in English, each noun is itself classified according to its property: countable or uncountable. The principle of individualization is not controlled by "Bounded/Unbounded" cognition.

5 Conclusion

This paper introduced an empirical study on the difficulties in learning "One (一) + Classifier (量詞)" in Chinese based on learners' corpora written by English-native learners and Japanese-native learners at CEFR-based A2 and B1 levels. The interesting contrast between the TOCFL English-native learner's corpus and the TUFS-NTNU Japanese learners' corpus displays the contrastive "overuse" versus "the lack of <一 + Classifier>".

The overuse of "One (一) + Classifier" in the English-native TOCFL corpus suggests the overgeneralization by English-native learners of Chinese that "a/an NP" is equivalent to "One (一) + Classifier" NP. We also assume that the lack of "One (一) + Classifier" in the TUFS-NTNU Japanese learners' corpus suggests the lack of individualization in terms of cognition in Japanese. The different features of the three languages are summarized (Table 6).

This comparative research into cross-linguistic learners' corpora suggests that it is indispensable to explore the pedagogy of Chinese based on learners' native language to develop more efficient and advanced learning science.

Table 6 Different Features in Number, Classifier and degree of Individualization

	(1) Number (Singular/Plural)	(2) Classifier	(3) Degree of Individualization
English	Obligatory	No classifier	High
Chinese	None except for 我們 (wǒ men) 這些 (zhè xiē)	Rich system	Middle "一 + Classifier" occurs in a "bounded" cognition
Japanese	None except for Watashi-**tachi**(we), kore-**ra** (these)	Not as rich a system as in Chinese	Low No article No determiner in syntax

Acknowledgements This study was supported by JSPS KAKENHI(Grant Number KAKEN JP25284101 "Construction of a Japanese-English-Chinese Online Error Corpus and Development of English, Japanese and Chinese Language Pedagogy taking into Account Learners' Native Languages" and also supported by National Taiwan Normal University, Peking University,Shanghai International Studies University and International Center for Japanese Studies,Tokyo University of Foregin Studies.

Appendix

Translation Task in Cross-Referential Multilingual Error Learner Corpora of Chinese and English

The Japanese, Chinese and English versions of the translation task are provided below for reference. Translations were conducted by Keiko Mochizuki together with Ms. Caroline E. Kano and YaMing Shen, all three understand Chinese, English and Japanese.

A reviewer queries whether the Chinese and English versions are truly equivalent to the Japanese original. We acknowledge the inherent difficulty of providing an exact translation, and recognize that this is one limitation of our methodology. We would like to stress, however, that the English text was used primarily as a reference for the authors during analysis, and that errors were identified based on their grammaticality, rather than on their degree of adherence to the English text.

１）

a. 私は、 20 代 から 30 代 にかけて、 北京、 上海、 ロンドン、
Watashi-wa, nijyuu-dai kara sannjyuu-dai nikakete, Pekin, Shannhai, Rondon,
I-Top twenties from thirties to Beijing Shanghai London

台湾 に 留学した こと が あります。
Taiwan ni ryuugakushita koto ga arimasu.
Taiwan DAT study abroad-PAST NMLZ NOM have-POL-NONPAST

b. When I was in my twenties and early thirties, I myself had the opportunity of studying in Beijing, Shanghai, London and Taiwan.

c. 在 我 二三十岁 的 时候 也 曾经 到 北京、 上海、
Zài wǒ èr-sānshi-suì de shíhou yě céngjīng dào Běijīng, Shànghǎi,
when I 20 30 years old NOM time also once to Beijing Shanghai

伦敦 以及 台湾 留学 过。
Lúndūn yǐjí Táiwān liúxué guo.
London and Taiwan study abroad EXP

２）

a. 留学時代 の 思い出 として、 今、 懐かしく 思い出す の は、
Ryuugaku-jidai no omoide toshite, ima, natukashiku omoidasu no-wa,
study abroad GEN memories of still now fondly remember-NOTPAST SE-TOP

先生方 の お宅 に 招かれ、 おもてなし を
sennsei-gata no otaku ni manekare, omotenashi o
professors GEN homes DAT invite-PASS-NOTPAST warm hospitality ACC

受けた 思い出 です。
uketa omoide desu.
received-PAST memories POL-NOTPAST

b. Of all my memories of studying abroad, what I still now remember most fondly, are the occasions when I was invited to the homes of my professors, and the warm hospitality I received.

c. 如今 每 当 我 回 想 起 当时 的 留学 生活 时，
Rújīn měi dāng wǒ huí xiǎng qi dāngshí de liúxué shēnghuó shí,
now every when I recall-remember at the time NOM life of studying abroad time

总是 会 想起 每回 到 老师 家 里 做客时 的 情景。
zǒngshì huì xiǎngqǐ měi huí dào lǎoshī jiā-li zuòkè-shí de qíngjǐng.
always can remember every time arrive teacher house-inside being a guest NOM scene

３）
a. まず、 最初に、 上海 留学中 の
Mazu, saisho ni, Shannhai ryuugakuchuu no
in this connection first Shanghai studying GEN

思い出 について お話 します。
omoide nituite ohanashi shimasu.
memories about talk do-POL-NONPAST

b. In this connection, I would first like to talk about my memories of studying in Shanghai.

c. 首先， 就 让 我 谈谈 在 上海 留学时
Shǒuxiān jiù ràng wǒ tántán zài Shànghǎi liúxué-shí
first of all then let I talk-talk(talk a little) in Shanghai study abroad

的 一段 回忆。
de yíduàn huíyì.
NOM a period of memory

４）

a. 東京 外国語 大学 で 中国 語学 の

Tookyo-gaikokugo-daigaku de chuugoku-gogaku no
Tokyo University of Foreign Studies LOC Chinese linguistics GEN

修士 号 を 得た 私 は、 中国 政府 公費 留学生
shuushi-goo o eta watashi-wa, Chuugoku-seifu-kouhi-ryuugakusei
M.A ACC receive-PAST I-TOP Chinese government-sponsored exchange student

として、 1986 年 から 1988 年 にかけて、
toshite, 1986-nenn kara 1988-nenn nikakete,
as 1986 from 1988 to

復旦大学(Fudan University) に 留学 しました。
Fukutann-daigaku ni ryuugaku shimashita.
Fudan University DAT study-PST do-POL-PAST

b. After receiving my M.A. in Chinese from Tokyo University of Foreign Studies, I went as a Chinese government-sponsored exchange student to Fudan University, where I studied from 1986 to 1988.

c. 从 1986 年 到 1988 年， 在 修完 东京外国语 大学 的
Cóng 1986 nián dào 1988 nián, zài xiūwán Dōngjīng-wàiguóyǔ dàxué de
from 1986 year to 1998 year when finish Tokyo University of Foreign Studies NOM

硕士 课程 之后，我 以 中国 政府 公费 留学生
shuòshì-kèchéng zhīhòu, wǒ yǐ Zhōngguó-zhèngfǔ gōngfèi liúxuéshēng
master degree after I as China government-sponsored public expense exchange student

的 身份 到 上海 复旦大学 留学 了 两年。
de shēnfèn dào Shànghǎi FùdànDàxué liúxué le liǎng nián.
NOM identity to Shanghai Fudan University study abroad PFV 2years

５）

a. 指導 教授 は、 著名 な 中国語 学者 であった
Shidou-kyoujyu-wa, chomei na Chuugokugo-gakusha deatta
academic supervisor-TOP eminent linguist of Chinese linguistics be-PAST

胡裕樹 教授 でした。
HuYushu kyoujyu deshita.
Hu Yushu professor POL-PAST

b. My academic supervisor was the eminent Sinologist, Professor Hu Yushu.

c. 我 的 指导教授 是 著名 的 汉语 语言学家 胡裕树 教授。
Wǒ de zhǐdǎojiàoshòu shì zhùmíng de Hànyǔ yǔyánxuéjiā Hú Yùshù jiàoshòu.
my GEN academic supervisor be famous NOM Chinese linguist Hu Yushu professor

６）

a. その頃は、 復旦大学 の 先生方 に は、
Sonokoro-wa, Fukutan-daigaku no sennseigata ni wa,
those days-TOP Fudan University GEN professors DAT TOP

研究室 が なく、 論文 指導 は、
kennkyuushitu ga naku, ronnbunn-shidou-wa,
own room NOM have-NEG-NONPAST supervision-TOP

大学 に 隣接する 宿舎 に 住んでいらっしゃる
daigaku ni rinsetsusuru shukusha ni sunndeirasshararu
university building DAT adjoining university lodgings DAT live-POL-NONPAST

ご自宅 の 書斎 兼 寝室 で 行われました。
gojitaku no shosai kenn shinnshitu de okonaw-are-mashita.
private GEN study cum bedroom DAT conduct-PASS-POL-PAST

b. In those days, professors at Fudan University did not have their own room, and supervision of students' theses would be conducted in their private bedroom-cum-study in the university lodgings adjoining the university building, where they lived.

c. 当时 复旦大学 的 老师们 并没有 个人的 研究室，
Dāngshí FùdànDàxué de lǎoshī-men bìng-méiyǒu gèrén-de yánjiūshì,
at that time Fudan University GEN teachers NEG-have individual-NOM laboratory

每次 的 论文 指导课 都是 在 紧邻 大学 的 老师宿舍
měicì de lùnwén zhǐdǎo-kè dōu-shì zài jǐnlín dàxué de lǎoshī-sùshè
every time GEN thesis guidance both-be in close to university NOM teacher dormitory

里 的 书房 兼 寝室里 进行 的。
li de shūfáng jiān qǐnshì-li jìnxíng de.
in NOM study cum bedroom-in doing NOM

７）
a. 先生方 の ご自宅 に は 電話 も なく、
Sennseigata no gojitaku ni wa denwa mo naku,
professors GEN lodgings DET TOP telephone also NEG-NONPAST

突然 訪ねていく こと が 多かった のですが、
totuzen tazuneteiku koto ga ookatta no-desu ga,
unexpectedly call on-NOTPAST NMLZ NOM often-PST SE-POL-NOTPAST

突然 お伺い しても、 必ず ドア を 開けてくださった、
totsuzen oukagai shitemo, kanarazu doa o akete-kudasa-tta,
sudden visit do-POL=NOTPAST always door ACC invite in-POL-PAST

そんな 牧歌的時代 でした。
sonna bokkateki-jidai deshita.
those idyllic times POL-PAST

b. As the professors' lodgings were not equipped with a telephone, students would often call on them unexpectedly. But however sudden a student's visit might be, in those idyllic times, their professor would always invite them in.

c. 也 由于 当时 老师 宿舍 里 还没有 安装 电话，
Yě yóuyú dāngshí lǎoshī sùshè li hái-méiyǒu ānzhuāng diànhuà,
also due to that time teacher dormitory inside not yet install telephone

所以 常常 都是 无事 先告知 的 突然造访，
suǒyǐ chángcháng dōushì wúshì xiān-gàozhī de tūrán-zàofǎng,
so always all be without notice first NOM sudden-visit

但是 尽管如此， 老师 及其 家人 每次 也 都 一定 欣然 开 门
dànshì jǐnguǎn-rúcǐ, lǎoshī jíqí jiārén měicì yě dōu yídìng xīnrán kāi mén
but even so teacher together family every time also all certainly gladly open the door

迎客， 我也 从未 尝 过 闭门之羹。 那是 一个 如此 纯朴 的 时代！
yíngkè, wǒ-yě cóngwèi chángguo bìmén-zhīgēng. Nàshì yíge rúcǐ chúnpǔ de shídài !
welcome I also never taste-EXP refuse that be 1 CL such simple NOM period

８）

a. ある日、予約　なしに　胡先生　の　おうち　を　訪ねた　私　に、
Aruhi, yoyaku nashini Hu-sensei no ouchi o tazuneta watashi ni,
one day appointment without Professor Hu GEN home ACC arrive-PAST I DAT

ご一家は、　「ちょうど　八宝飯　(もち米　で　作った
goikka-wa, “choodo happou-han (mochigome de tsukutta
he and his family-TOP just babaofan (glutinous rice with make-PAST

８つ　の　ドライフルーツ　が　飾られた
8tu no doraifuruutsu ga kazarareta
eight kinds GEN dried fruit NOM decorate-PASS-PAST

デコレーションケーキ)　が　蒸しあがった　から、　食べなさい」　と、
dekoreeshonnkeeki) ga mushi-agatta kara, tabenasai” to,
decorated cake NOM steam-PAST so do have some QT

ふるまって　くださった　のです。
furumatte kudasatta nodesu.
welcomed me-POL-PST SE-POL-NONPAST

b. One day, when I arrived at Professor Hu’s home without an appointment, he and his family welcomed me with a “We’ve just steamed a *babaofan* (a cake made with glutinous rice, decorated with eight kinds of dried fruit), so do have some!”

c. 有一天，　又是　一个　突然的　造访。　胡老师　一家人　对　突然
Yǒu yì tiān, yòushì yí ge tūrán-de zàofǎng. Húlǎoshī yìjiārén duì tūrán
someday again 1 CL suddenly invite teacher-Hu family to suddenly

出现　的　我　说道：　“正好　有　蒸好的　八宝饭，　吃了
chūxiàn de wǒ shuōdào： “zhènghǎo yǒu zhēnghǎo-de bābǎofàn, chī le
appear NOM I say just have make-past babaofan, have-past

再　走　吧！”，　一边　拿出　八宝饭　招待　我　这个　不速之客。
zài zǒu ba!”, yìbian náchū bābǎofàn zhāodài wǒ zhège búsùzhīkè.
again leave SFP with take out babaofan, serve I this CL uninvited guest

９）

a. 蒸したて　の　八宝飯　の　「やさしく、　柔らかく、　幸福な　甘さ」　は、
Mushitate　no　happou-han　no　“yasashiku,　yawarakaku,　koufukuna　amasa”　-wa,
freshly steamed　GEN　babaofan　GEN　gentle　delicate　blissful　sweetness-TOP

忘れる　こと　が　できません。
wasureru　koto　ga　dekimasen.
forget　NMLZ　NOM　NEG-POL-NOTPAST

b. I will never forget the ‘gentle, delicate, blissful sweetness’ of that freshly steamed *babaofan*.

c. 刚　蒸好　的　八宝饭　所带有　的　那种　“软软、　热热、
Gāng　zhēnghǎo　de　bābǎofàn　suǒ-dàiyǒu　de　nàzhǒng　“ruǎnruǎn,　rèrè,
just now　make-past　NOM　babaofan　with　NOM　that CL　soft-soft　hot-hot

甜甜”　的　幸福　滋味，　到　现在　仍然　记忆犹新。
tiántián”　de　xìngfú　zīwèi,　dào　xiànzài　réngrán　jìyì-yóuxīn.
sweet　NOM　happiness　taste　until　now　still　in memory

１０）

a. その後、　中国料理店　で、　八宝飯　を　みつける　と、
Sonogo,　chuugoku-ryouri-ten　de,　happou-hann　o　mitukeru　to,
thereafter　Chinese restaurant　DAT　babaofan　ACC　find-NOTPAST　whenever

必ず　注文し、　胡先生　の　おもてなし　を　思い出す　のです。
kanarazu　chuumonnshi,　Hu-sennsei　no　omotenashi　o　omoidasu　nodesu.
always　order-NOTPAST Professor Hu　GEN　kind hospitality ACC　recall　SE-POL-NOTPAST

b. Thereafter, whenever I go to a Chinese restaurant and find *babaofan* on the menu, I always make a point of ordering it, and recall the kind hospitality which Professor Hu extended to me.

c. 从那以后， 每当 在 中国 餐馆里 看到 八宝饭， 我 一定 会
Cóngnàyǐhòu, měidāng zài Zhōngguó cānguǎn-li kàndào bābǎofàn, wǒ yídìng huì
after that everytime at Chinese restaurant inside see babaofan I certainly can

点来 品尝， 不为 别的， 就 只 为 想 再回味 一次 胡老师
diǎnlái pǐncháng, búwèi biéde, jiù zhǐ wèi xiǎng zài-huíwèi yícì Hú-lǎoshī
order to taste not for other just only for want recall again once teacher-Hu

和 他 家人 的 待客之道。
hé tā jiārén de dàikè-zhīdào.
and he family GEN the way of hospitality

11）
a. 論文 の 個人 指導 は、 蒲団 が
Ronnbun no kojinn-shidou-wa, futon ga
thesis GEN individual guidance session-TOP bed cover NOM

ロールケーキ のように 巻かれ、
roorukeeki noyouni makare,
Swiss roll like rolled up-PASS-NOTPAST

整えられて 長椅子 と 化した 胡先生 の ベッド に
totonoerarete nagaisu to kashita Hu-sennsei no beddo ni
arrange-PASS-NOTPAST sofa turn into-PAST Professor Hu GEN bed DAT

座って 行われ ました。
suwatte okonaware mashita.
sit conduct-PASS do-POL-PAST

b. An individual guidance session on a student's thesis would be conducted seated on Professor Hu's bed, which, with the bed cover rolled up like a Swiss roll, was turned into a sofa.

c. 每次 上课 时，老师 都 会 将 棉被 卷成 像
Měicì shàngkè shí, lǎoshī dōu huì jiāng miánbèi juǎnchéng xiàng
everytime class time teacher all can will cotton wadding batching like

西式 卷心蛋糕 似 的 长条状， 然后 将 床铺
xīshì juǎnxīn-dàngāo sì de chángtiáozhuàng, ránhòu jiāng chuángpù
western roll-cake like NOM rectangular then will bed

整理得 如同 一条 长凳子， 要 我 坐在 上面 上课。
zhěnglǐ-de rútóng yìtiáo chángdèngzi, yào wǒ zuò-zài shàngmian shàngkè.
collative like 1 CL bench let I sit on have class

12）
a. 私 が ベッド に 座る と、 胡先生は まず、 龍井茶 を
Watashi ga beddo ni suwaru to, Hu-sennsei-wa mazu, Longjing-cha o
I NOM bed DAT seat down as soon as Professor Hu-TOP first Longjing green tea ACC

蓋付き の 中国 式 マグカップ に ひと つまみ 入れて、
futa-tsuki no chuugoku-shiki-magukappu ni hito tsumami irete,
with a lid GEN Chinese-style mug DAT a few leaves place-NOTPAST

魔法瓶 から お湯 を いれ、 お茶 を 淹れて くださいました。
mahoubin kara oyu o ire, ocha o irete kudasaimashita.
thermos from hot water ACC add-NOTPAST green tea ACC serve POL-PAST

b. As soon as I had sat down on Professor Hu's bed, Professor Hu would place a few leaves of Longjing green tea in a Chinese-style mug with a lid, add some hot water from a thermos, and serve it to me.

c. 我 一坐 定 后， 老师 会 先在 一个 传统 中国式的、 带 盖子的
Wǒ yízuò dìng hòu, lǎoshī huì xiān-zài yíge chuántǒng Zhōngguóshì-de, dài gàizide
I take seat after teacher will first 1 CL traditional Chinese with cap

茶杯里 放入 一小撮的 龙井茶叶， 然后 从 热水瓶 里 倒 出
chábēi-li fàngrù yìxiǎozuǒde Lóngjǐng-cháyè, ránhòu cóng rèshuǐpíng li dào chū
tea-pot put in a little bit Longjingchaye then from thermos inside pour-out

热开水， 为 我 沏上 一杯 热茶。
rèkāishuǐ, wèi wǒ qīshàng yìbēi rèchá.
boiled-water for me make a cup of hot tea

13）

a. そして、 結婚式 の 引き出物 の ような、赤い キャンディーボックス の
Soshite, kekkonn-shiki no hikidemono noyouna, akai kyandiibokkusu no
then wedding GEN gift looked as red sweet box GEN

蓋 を とって、 「キャンディー を どうぞ」 と 優しく 微笑み ながら
futa wo totte, “kyanndii wo douzo” to yasashiku hohoemi nagara
lid ACC take off-NOTPAST sweet ACC do have QT kindly smiling while

すすめて くださった のでした。
susumete kudasatta no deshi - ta.
offer me POL-PAST SE-POL-PAST

b. He would then take the lid off a red sweet box which looked as though it might have been a gift he had received as a guest at a wedding, and, smiling kindly, and with a “Do have a sweet!”, offer me one.

c. 之后，再 拿出 一个 好像 装 喜糖 用 的 大红色 的
Zhīhòu, zài náchū yíge hǎoxiàng zhuāng xǐtáng yòng de dàhóngsè-de
then again take out 1 CL like hold wedding candies use NOM red GEN

糖果 盒， 打开 盒子，亲切地 微笑着 要 我 吃 糖。
tángguǒ-hé, dǎkāi hézi, qīnqiè-de wēixiào-zhe yào wǒ chī táng.
candy case open case kindly smile-DUR let I eat candy

14）
a. とても 質素な 時代 でしたが、 胡先生 ご一家 の おもてなしは、
Totemo shissona jidai deshitaga Hu-sennsei goikka no omotenashi-wa,
very modest times POL-PAST Professor Hu his family GEN warm hospitality-TOP

いまも 宝物の ような 思い出 として、胸に 刻まれています。
imamo takaramono noyouna omoide toshite, mune ni kizam- are te-imasu.
still treasure like memory QT heart DAT engraved in-PASS-POL-NOTPAST

b. They were very modest times, but the warm hospitality which I received from Professor Hu and his family still remains like a treasure engraved in my memory.

c. 虽然 是 一个 物资 不是 很 丰裕的 时代，但是 胡老师 以及 他 家人
Suīrán shì yíge wùzī bú shì hěn fēngyù-de shídài, dànshì Hú-lǎoshī yǐjí tā jiārén
although is 1 CL materials not very wealthy period but teacher-Hu and he family

对 我的 热情款待 的 回忆， 始终 就像 一个 宝藏 一样，
duì wǒde rèqíng-kuǎndài de huíyì, shǐzhōng jiùxiàng yíge bǎozàng yíyàng,
for I GEN warm reception NOM memories always like 1 CL treasure same

永远地 深深地 埋藏在 我的 心中。
Yǒngyuǎn-de shēnshēn-de máicángzài wǒ-de xīnzhōng.
forever DE deeply DE bury in I GEN heart

[A] Translation task for Japanese native speakers (English and Chinese majors at Tokyo University of Foreign Studies) Translation into English and Chinese

Original Japanese Text (望月圭子 Keiko MOCHIZUKI)

私は、20 代から 30 代にかけて、北京、上海、ロンドン、台湾に留学したことがあります。留学時代の思い出として、今、懐かしく思い出すのは、先生方のお宅に招かれ、おもてなしを受けた思い出です。

まず、最初に、上海留学中の思い出 1)についてお話します。

東京外国語大学で中国語学の修士号を得た私は、中国政府公費留学生として、1986 年から 1988 年にかけて、復旦大学(Fudan University)に留学しました。指導教授は、著名な中国語学者であった胡　裕樹教授(Prof. HuYushu)でした。

その頃は、復旦大学の先生方には、研究室がなく、論文指導は、大学に隣接する宿舎に住んでいらっしゃるご自宅の書斎兼寝室で行われました。先生方のご自宅には電話もなく、突然訪ねていくことが多かったのですが、突然お伺いしても、必ずドアを開けてくださった、そんな牧歌的時代 2)でした。

ある日、予約なしに胡先生のおうちを訪ねた 3)私に、ご一家は、「ちょうど八宝飯(Babaofan;もち米で作った 8 つのドライフルーツが飾られたデコレーションケーキ)が蒸しあがったから、食べなさい」と、ふるまってくださったのです。

蒸したての八宝飯の「やさしく、柔らかく、幸福な甘さ」は、忘れることができません。その後、中国料理店で、八宝飯をみつけると、必ず注文し、胡先生のおもてなしを思い出すのです。

論文の個人指導は、蒲団がロールケーキのように巻かれ、整えられて長椅子と化した胡先生のベッド 4)に座って行われました。 私がベッドに座ると、胡先生はまず、龍井茶(LongJing Green Tea)を蓋付きの中国式マグカップ 5)にひとつまみ入れて、魔法瓶からお湯をいれ、お茶 6)を淹れてくださいました。そして、結婚式の引き出物のような、赤いキャンディーボックス 7)の蓋をとって、「キャンディーをどうぞ」と優しく微笑みながらすすめてくださったのでした。

とても質素な時代 8)でしたが、胡先生ご一家のおもてなしは、いまも宝物のような思い出 9)として、胸に刻まれています。

[B] Translation task for Chinese native speakers (Shanghai International Studies University & National Taiwan Normal University) Translation into Japanese and English

Original Chinese Text (Translated by 申 亜敏 Shen YaMing)

在我二三十岁的时候也曾经到北京、上海、伦敦以及台湾留学过。如今每当我回想起当时的留学生活时，总是会想起每回到老师家里做客时的情景。首先，就让我谈谈在上海留学时的一段**回忆** 1)。

从 1986 年到 1988 年，在修完东京外国语大学的硕士课程之后，我以中国政府公费留学生的身份到上海复旦大学留学了两年，我的指导教授是著名的汉语语言学家胡裕树教授。

当时复旦大学的老师们并没有个人的研究室，每次的论文指导课都是在紧邻大学的老师宿舍里的书房兼寝室里进行的。也由于当时老师宿舍里还没有安装电话，所以常常都是无事先告知的突然造访，但是尽管如此，老师及其家人每次也都一定欣然开门迎客，我也从未尝过闭门之羹。那是一个如此纯朴的**时代** 2)！

有一天，又是**一个突然的造访** 3)。胡老师一家人对突然出现的我说道：“正好有蒸好的八宝饭，吃了再走吧！”，一边拿出八宝饭招待我这个不速之客。刚蒸好的八宝饭所带有的那种“软软、热热、甜甜”的幸福滋味，到现在仍然记忆犹新。从那以后，每当在中国餐馆里看到八宝饭，我一定会点来品尝，不为别的，就只为想再回味一次胡老师和他家人的待客之道。

每次上课时，老师都会将棉被卷成像西式卷心蛋糕似的长条状，然后将床铺整理得如同**一条长凳子** 4)，要我坐在上面上课。我一坐定后，老师会先在**一个传统中国式的、带盖子的茶杯** 5) 里放入一小撮的龙井茶叶，然后从热水瓶里倒出热开水，为我沏上**一杯热茶** 6)。之后，再拿出一个好象装喜糖用的大红色的**糖果盒** 7)，打开盒子，亲切地微笑着要我吃糖。

虽然是一个物资不是很丰裕的**时代** 8)，但是胡老师以及他家人对我的热情款待的回忆，始终就像一个**宝藏**一样 9)，永远地深深地埋藏在我的心中。

[C] English Text

(Translated by Caroline Kano)

When I was in my twenties and early thirties, I myself had the opportunity of studying in Beijing, Shanghai, London and Taiwan. Of all my memories of studying abroad, what I still now remember most fondly, are the occasions when I was invited to the homes of my professors, and the warm hospitality I received. In this connection, I would first like to talk about my memories 1) of studying in Shanghai.

After receiving my M.A. in Chinese from Tokyo University of Foreign Studies, I went as a Chinese government-sponsored exchange student to Fudan University, where I studied from 1986 to 1988.

My academic supervisor was the eminent Sinologist, Professor Hu Yushu. In those days, professors at Fudan University did not have their own room, and supervision of students' theses would be conducted in their private bedroom-cum-study in the university lodgings adjoining the university building, where they lived. As the professors' lodgings were not equipped with a telephone, students would often call on them unexpectedly. But however sudden a student's visit might be, in those idyllic times 2), their professor would always invite them in.

One day, when I arrived at Professor Hu's home without an appointment 3), he and his family welcomed me with a “We've just steamed a *babaofan* (a cake made

with glutinous rice, decorated with eight kinds of dried fruit), so do have some!" I will never forget the 'gentle, delicate, blissful sweetness' of that freshly steamed *babaofan*. Thereafter, whenever I go to a Chinese restaurant and find *babaofan* on the menu, I always make a point of ordering it, and recall the kind hospitality which Professor Hu extended to me.

An individual guidance session on a student's thesis would be conducted seated on Professor Hu's bed, which, with the bed cover rolled up like a Swiss roll, was turned into a sofa 4). As soon as I had sat down on Professor Hu's bed, Professor Hu would place a few leaves of Longjing green tea in a Chinese-style mug with a lid 5), add some hot water from a thermos, and serve it 6) to me. He would then take the lid off a red sweet box which looked as though it might have been a gift he had received as a guest at a wedding 7), and, smiling kindly, and with a "Do have a sweet!", offer me one.

They were very modest times 8), but the warm hospitality which I received from Professor Hu and his family still remains like a treasure engraved in my memory 9).

References

Fukui, N. (1995). *Theory of projection in syntax.* Stanford: CSLI Publications.

Huang, C. T. J., Li, Y. H. A., & Simpson, A. (2014). *The Handbook of Chinese Linguistics*. John Wiley & Sons.

Ikegami, Y. (1991). "Do-langugage and become-langugage: Two contrasting types of linguistic representation. In Y. Ikegami (Ed.), *The Empire of Signs: Semiotic Essays on Japanese Culture.* Amsterdam: John Benjamins, pp. 285–326.

Ikegami, Y. (2007). *Japanese and Japanese Typology*. Tokyo: Chikuma Publishers. [池上嘉彦(2007). 日本語と日本語論。東京:筑摩書房].

Kageyama, T. (1996). *Verb Semantics: The Interface between Language and Cognition.* Tokyo: Kuroshio Publishers. [影山太郎(1996). 動詞意味論—言語と認知の接点。東京:くろしお出版].

Kageyama, T. (2002). *Japanese as an unbounded language.* Tokyo: Iwanami Publishers. [影山太郎(2002). けじめのない日本語。東京:岩波書店].

Kageyama, T., & Jacobsen, W. M. (Eds.). (2016). *Transitivity and valency alternations: Studies on Japanese and beyond*. De Gruyter Mouton.

Kageyama, T. (2021a). *Between lexical verbs and auxiliaries: The architecture of Japanese verb-verb complexes*. In Kageyama, Taro, P. E. Hook, & P. Pardeshi (Eds.), 2021. *Verb-Verb complexes in Asian Languages.* Oxford University Press, pp. 15–43.

Kageyama, T. (2021b). *Grammaticalization and constructionalization in Japanese lexical compound verbs.* In Kageyama, Taro, P. E. Hook, & P. Pardeshi (Eds.), 2021. *Verb-Verb complexes in Asian Languages.* Oxford University Press, pp. 70–102.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A Functional Reference Grammar (漢語語法).* Taipei: Crane Publishing Co. Lid.

Mochizuki, K. (2004). *Causative and Inchoative Alternation: Comparative Studies on Verbs in Chinese and Japanese.* Ph.D dissertation, National Tsing Hua University, Taiwan. http://140.113.39.130/cgi-bin/gs32/hugsweb.cgi/ccd=lJoInl/record?r1=2&h1=0.

Mochizuki, K. (2007). Patient-orientedness in resultative compound verbs in Chinese. In Y. Kawaguchi, T. Takagaki, N. Tomimori, & Y. Tsuruga (Eds.), *Corpus-Based Perspectives in Linguistics,* Amsterdam: John Benjamins, pp. 287–300.

Mochizuki, K. (2009). *Error Analysis in Voice by Advanced-level Chinese Learners of Japanese: Comparative Analysis with Chinese. Tokyo University of Foreign Studies Area and Culture Studies no.78*, 85–105.

Mochizuki, K., Sano, H., Shen, Y.-M., & Wu, C.-H. (2015). Cross-Linguistic Error Types of Misused Chinese Based on Learners' Corpora. *Computational Linguistics and Chinese Language Processing, 20*(1), 97–113.

Newbery-Payton, L., & Mochizuki, K. (2020). L1 Influence on Use of Tense/Aspect by Chinese and Japanese Learners of English. *Learner Corpus Studies in Asia and the World, 4*, 67–93.

Shen, Y.-M. (2009). *Resultative Compound Verbs in Chinese- From a Viewpoint of Comparative Analyses with Resultative Compound Verbs in Japanese and English Resultative Constructions.* Ph.D dissertation, Tokyo University of Foreign Studies. http://repository.tufs.ac.jp/handle/10108/56738.

Tai, J. H.-Y. (1984). Verbs and Times in Chinese: Vendler's Four Categories. *Lexical Semantics,* 289– 296, Chicago Linguistics Society.

Tai, J. H.-Y. (2003). Cognitive relativism: Resultative construction in Chinese. In *Language and Linguistics, 4*(2), 301–316.

古川裕 (2001). 外界事物的"显著性"与句中名词的"有标性"–"出现、存在、消失"与"有界、无界",《当代语言学》第3卷2001年第4期, (264–274) , 北京. [Furukawa Yutaka (2001). Cognitive saliency and nominal marke: appearance, existence, disappearance and boundedness, unboundedness, *Contemporary Linguistics*, 3–4*,* 264–274].

沈家煊 (1995). "有界"与"无界",《中国语文》, 第 5 期, 367–380. [Shen, J.-X. (1995). Boundedness and unboundedness. *Chinese Language and Writing, 5*, 367–380].

申亞敏, 望月圭子 (1997). 華語和日語的否定辭,《第五屆世界華語文教學研討會論文集, 語言分析組》(515–525). [Shen, Y.-M., & Mochizuki, K. (1997). Negative Form in Chinese and Japanese, *Selected papers on the 5th World Conference on Chinese Language Teaching, linguistic analysis part,* 515–525].

张斌, 齐沪扬等 (2002).《新编现代汉语》复旦大学出版社. [Zhang Bin, Qi Huyang et al. (2002). *Modern Chinese Grammar.* Fudan University Press].

張莉萍 (2013). TOCFL 作文語料庫的建置與應用, 崔希亮、张宝林 (主编)《第二届汉语中介语语料库建设与应用国际学术讨论会论文选集》(141–152). 北京: 北京语言大学出版社。[Li-ping Chang. (2013). Construction and applications of the TOCFL Composition Corpus, In Cui xiliang, Zhang Baolin Eds, *Selected papers on the 2nd International Symposium on the Construction and application of Chinese interlanguage corpora,* 141–152, Beijing Language and Culture University Press].

張莉萍 (2014). 不同母語背景華語學習者的用詞特徵: 以語料庫為本的研究《中文計算語言學刊》(IJCLCLP) , 19(2),53–72. [Li-ping Chang. (2014). Salient Linguistic Features of Chinese Learners with Different L1s: A Corpus-based Study, *International Journal of Computational Linguistics and Chinese Language Processing, 19*(2), 53–72].

Chinese Verb Complement Constructions of Manner and States: A Corpus-Based Comparison Between L1 and L2 Speakers

Hong Gang Jin, Jie Zhang, and Hongyin Tao

Abstract This paper deals with the acquisition of Verb Complement Constructions of Manner and States (VCM/S, 方式/情态补语) from a comparative and corpus-based perspective. An examination of L1 and L2 Chinese VCM/S production and development yields three main findings: (a) there are marked quantitative and qualitative differences between L1 and L2 VCM/S production at both construction and component levels; (b) these persistent productive differences reflect the indispensable roles of psycholinguistic factors, such as frequency, complexity, form-meaning mapping, and co-occurrence patterns of VP and VC, especially on verb choices; and (c) L2 VCM/S construction learning is like any other construction learning that follows a U-shaped learning path that consists of unique and distinctive stages. The process also involves both implicit factors and explicit classroom input and instruction. The theoretical and pedagogical implications of these findings are discussed.

Keywords Complement of Manner and States · Construction learning · Form-meaning mapping · Usage-based models · Complexity scale · L1 and L2 comparison

H. G. Jin (✉) · H. Tao
East Asian Languages and Literatures, Hamilton College, 198 College Hill Road, Clinton, NY 13323, USA
e-mail: hjin@hamilton.edu

J. Zhang
Department of Modern Languages, Literatures, and Linguistics, University of Oklahoma, 780 Van Vleet Oval, Norman, OK 73019, USA

J. Zhang · H. Tao
Department of Asian Languages and Cultures, UCLA, Los Angeles, CA 90095-1540, 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_6

1 Introduction

Complement constructions are a major syntactic pattern in the Mandarin Chinese grammatical system, where a variety of complements can be formed to express a range of meanings, such as result, direction, degree, and so forth (Li & Thompson, 1981). As such, they constitute some of the most unique features of the Chinese syntactic system (Shen, 2003). Following our L1 corpus-based study (Tao et al., 2020) on a similar construction, this paper deals with the acquisition of one type of complement construction, which we call Verb Complement Constructions of Manner and States (VCM/S, 方式/情态补语). VCM/S constructions typically consist of three key components: the verb predicate (VP), the complementizer *de* (得), and the complement of different syntactic structures. They indicate either the manner in which the action named by the verbal predicate is executed or evaluated or a state toward which the action is carried out (ibid.). Two quick examples illustrating these patterns can be found in (1) and (2).

(1) 这个人写得不好.
Zhe ge ren xie de bu hao.
This person write DE not well.
'This person does not write well.'

(2) 他变得很精神.
Ta bian de hen jingshen.
He become DE very energetic
'He becomes very energetic.'

In (1), the complement *bu hao* 'not well' can be seen as an evaluation ('how well') of the verbal predicate *xie* 'write'. In (2), on the other hand, the complement *hen jingshen* 'very energetic' can be understood to be the state toward which the action of *bian* 'change, become' is carried out.

As a construction, VCM/S involves multiple components and has posed challenges to learners of Chinese as a second language (CSL). Previous studies have shown that learners' error rate is at 25%–50% due to the uniquely grammaticalized structure and subtle functions associated with the construction (Sun, 2002; Feng, 2013; Jiang, 2019, among others). Few CSL studies, however, have examined VCM/S from the perspective of usage-based construction learning and dealt with both implicit and explicit learning factors, nor have they investigated VCM/S as an independent construction from other complement types and compared L2 learner development in connection with L1 production data. When learner data do get analyzed, however, existing studies tend to be descriptive in nature, focusing, for example, mostly on single VCM/S component and raw frequency counts of VCM/S sentences (e.g. D. Sun, 2002; Q. Sun, 2018; Zhou & Deng, 2009, among others).

The current study intends to address those shortfalls by carrying out a comparative corpus study, and it will be informed by usage-based approaches (Bybee & Hopper, 2001; Ellis, 2006, 2008, 2012) and Construction Grammar (Goldberg, 1995; Trousdale & Hoffmann, 2013), where form and meaning pairing and co-occurrence properties are argued to play a critical role in understanding grammatical patterns in both the first language (L1) and the second language (L2). Our data comprise corpora of compositions written by CSL learners in the US college setting and by L1 Chinese speakers in preparation for national university entrance examinations. By using both L1 and L2 data, we hope to (a) determine if there are similarities or differences between L1 and L2 speakers' production of the VCM/S construction in terms of frequency, form, function, form-function mapping, as well as distributional properties in the form of co-occurrence patterns (Ellis, 2002, 2012); and (b) delineate L2 construction learning paths at different stages. Our comparative empirical study will form the basis for further exploration of L2 acquisition theory and pedagogical practice.

2 Literature Review

2.1 Usage-Based Approaches to Construction Learning

Usage-based approaches to language acquisition explore how humans learn language from experience and view language acquisition as a process of learning constructions (Bybee & Hopper, 2001; Ellis, 2006, 2008, 2012; Hoey, 2005). This acquisitional model emphasizes associative and cognitive principles of learning and focuses on investigating psycholinguistic factors of construction acquisition such as frequency, contingency, and form-meaning mapping that drive the acquisition and use of linguistic constructions (Ellis, 2012). Two concepts play the central role in this model: constructions and distributional properties of constructions.

Constructions, according to Goldberg (1995) and Trousdale and Hoffmann (2013), are defined as form-meaning mappings, conventionalized in the speech community, and entrenched as language knowledge in the learner's mind. Constructions are the fundamental units of language and language acquisition. Factors affecting construction acquisition are believed to come from several dimensions: (1) form-related factors such as frequency and salience; (2) function-related factors such as prototypicality, generality, and redundancy; (3) contingency of form and function; and (4) learner related factors such as learner attention, automaticity, and transfer, among others (Ellis, 2002, 2012).

Distributional properties are found to affect language processing and learning. Research has shown that language users are sensitive to detailed distributional information at many levels of linguistic analysis and at different grain sizes: from phonemes, morphemes, words, multi-word phrases, and syntactic constructions. Both language comprehension and production are affected by distributional factors,

such as the overall frequencies of the syntactic construction (Gahl & Garnsey, 2004; Tily et al., 2009, among others), the frequency of different components in specific syntactic constructions (Clifton et al., 1984; Garnsey, et al., 1997; Arnon & Snider, 2010, among others), and co-occurrence relations between verbs and specific arguments/complements (Trueswell and Tanenhaus 1994; Tao et al., 2020).

2.2 Chinese VCM/S Acquisition Studies

CSL research on the Chinese VCM/S acquisition began mostly in the beginning of this century. Many studies focused on the umbrella Chinese *de* complements containing 5–8 different complement constructions, of which VCM/S was one of them. Among the acquisition studies of VCM/S constructions, a majority of them were descriptive in nature, focusing on the identification, categorization, and description of the VCM/S development in terms of raw frequency counts of VCM/S sentences (D. Sun, 2002; Q. Sun, 2018; Zhou & Deng, 2009), interlanguage error patterns (D. Sun, 2002; Feng, 2013; Jiang, 2019, among others), and some general developmental patterns (Q. Sun, 2018; Zhou & Deng, 2009; Feng, 2013, among others).

While most existing CSL VCM/S studies used elicited or survey data, a few utilized corpus data. In these cases, the majority of researchers used corpus data from two sources. One is the overseas students' interlanguage composition corpus collected from a Chinese proficiency test known as *Hanyu Shuiping Kaoshi* (HSK), and the other is a self-built corpus of written samples collected from one or several institutions. For example, D. Sun (2002) analyzed 184 sample sentences containing VCM/S constructions from the Beijing Language and Culture University (BCC) corpus. Based on the error rate, the author concluded that VCM/S was the easiest for L2 learners to acquire among all verb complement constructions as fewer errors were found in their VCM/S (N = 184) collection. This claim was supported by Feng (2013) and Jiang (2019), especially in comparison with other complement types such as resultatives and potentials. Zhou and Deng (2009), by contrast, investigated two types of VCM/S constructions, where the object is in different positions (VO and OV), with both corpus data and experimental tests, and revealed that the OV structure in VCM/S constructions was actually hard to acquire and it was absent in L2 learners' production until they reached advanced proficiency levels.

Overall, we find that previous studies are descriptive in nature and that the conclusions are generally mixed. Few have examined the acquisition of the VCM/S construction learning using large-scale corpus data for both L1 and L2 with a unified L1 background, and rarely have researchers investigated both forms and functions of VCM/S constructions and the psycholinguistic factors of construction learning, such as construction frequency and complexity that may impact learner development. More importantly, previous studies have mainly looked at either the verb predicate (VP) or the verb complement (VC) before and after *de*, without examining VCM/S

as a construction with the two co-occurring open slots of VP and VC. This study will attempt to address those issues by adopting usage-based approaches to language acquisition, focusing on construction learning with corpus evidence from both L1 and L2 Chinese.

Applying usage-based approaches to construction learning, specifically that of the Chinese VCM/S construction, this study is set out to investigate three research foci: (a) VCM/S production in frequency and distribution; (b) verb choices in the predicate; and (c) VCM/S complexity scales. For each of these research foci, we seek to explore similarities and differences between L1 and L2 as well as the developmental pattern of L2 speakers' production. In the end, we will further explore the implications of the results of the comparative data for both acquisition theory and pedagogical practices in CSL.

3 Data and Methodology

3.1 The Corpora

The corpus used in this study is composed of 1,284 compositions written by CSL learners and Chinese L1 speakers with a total word count of 376,387. The learner corpus consists of 1,136 compositions written by English-speaking CSL learners at roughly four levels based on the American Council on the Teaching of Foreign Languages (ACTFL) proficiency scale[1]: (1) Novice-Mid to Intermediate-Low, (2) Intermediate-Mid to Intermediate-High, (3) Intermediate-High to Advanced-Low, and (4) Advanced-Mid and higher. For ease of discussion, we will use the short forms L2-A, L2-B, L2-C, and L2-D to represent, respectively, these four learner groups. The L2 collection came from three sources. The first two levels are a collection of student compositions at a comprehensive public university in North America. The third level comes from an intensive US study abroad program whose students represented over 20 universities and colleges in North America. Finally, the fourth level is a selection of compositions retrieved from the *Hanyu Shuiping Kaoshi (*HSK*) Dongtai Zuowen Yuliaoku* (Chinese Proficiency Test Dynamic Composition Corpus) Version 1.1.[2] The compositions culled from this collection were mostly narrative, descriptive, or argumentative by genre, and only those who registered their nationality as either the United States or Canada were included.

[1] American Council on the Teaching of Foreign Languages (ACTFL) categorizes foreign language proficiency into five major scales: Novice, Intermediate, Advanced, Superior, and Distinguished. The Novice, Intermediate, and Advanced levels each have three sublevels: Low, Mid, and High (ACTFL, 2012).

[2] The HSK Advanced was designed for CSL learners who have completed at least 4 years of Chinese instruction or who have been immersed in Chinese speaking environments for more than 3,000 h (Zhang, 2011).

Table 1 Composition of the corpus

Proficiency levels	Number of samples	Mean sample length by words	Word type	Word token	TTR
L2-A	409	110	3,175	44,994	7
L2-B	248	181	4,269	44,968	10
L2-C	206	596	7,336	122,849	6
L2-D	273	320	8,850	87,405	10
L2 Mean	**284**	**302**	**5,907**	**75,054**	**8.25**
L1	**148**	**515**	**11,203**	**76,171**	**15**
Total	**1,284**		**34,833**	**376,387**	**Ave. 9.6**

To maximize compatibility, we also include 148 compositions written by Chinese L1 speakers to serve as the L1 benchmark. This subset is a collection of compositions written by Chinese high school students taking or preparing for the National Matriculation Test (*gaokao* 高考), a high-stakes standardized test taking place annually in China. Most of the students who wrote these essays were approximately 17–20 years old in their third year of high school. The essays were downloaded from the official educational websites *Zhongguo Jiaoyu Zaixian* (China Education Online) and *Renmin Wang* (People's Daily Online). Several genres were represented, including narrative, argumentative, expository, and prose.

Table 1 presents the composition of the corpus with counts of the number of writing samples, average length of the samples (word per sample (wps)), word type, word token, and type/token ratio (TTR). In terms of the length of the writing samples, the L1 group is over 500 wps, much longer than the average L2 samples, which is at 302wps (the exception is the L2-C group, which has an average length of near 600wps). In terms of the variety of words employed, as one would expect, L1 speakers' writing averaged more word types (11,203 vs. 5,907), tokens (76,171 vs. 75,054), and a higher TTR (15 vs. 8.25). Across all proficiency levels, the L2 learner data demonstrate a growth in the number of types, tokens, and TTR as their overall language proficiency improves, with the exception of the L2-B group, which has a relatively higher TTR of 10. This may be due to the fact that this group is composed of learners enrolled in two semesters (fifth and sixth semesters of the program sequence) with the likelihood of being assigned compositions on more topics than learners over only one semester.

3.2 VCM/S Tagging and Coding Decisions

Both L1 and L2 data were first tagged for parts of speech information with the POS tool developed by the Chinese Applied Linguistics Institute (http://corpus.zhonghuayuwen.org/). Then a regular expression under AntConc (Anthony, 2019) was used to extract all the constructions fitting the VCM/S pattern, with the results being further filtered manually before coding was performed.

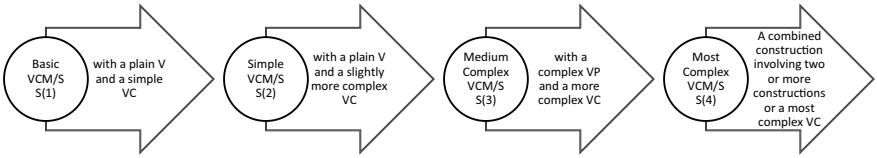


Fig. 1 Complexity scale of the Chinese VCM/S construction

Using an Excel spreadsheet, the first two authors of the paper went through all instances containing *de* and deleted cases that were deemed not the focus of the current study, such as potential verb complements, degree verb complements, and *de* used as a modal verb. Out of a total of 424 entries, there were 14 entries in discrepancy (3.3%). The two coders discussed these entries and reached an agreement for all 14 cases.

To capture how form and meaning are mapped and to explore the range of verbs and the prototypicality effect, we divided verbs in VCM/S constructions into four bands based on their raw frequencies in our corpus. A-list verbs are those with 18 or more token frequencies in the corpus; B-list includes verbs with 10–17 token frequencies; C-list verbs are common verbs with 2–9 token frequencies; in contrast, D-list includes infrequent words that have only one instance in our corpus. (For a full list of the verbs in frequency bands, please refer to Appendix.)

To capture the complexity of VCM/S constructions, we developed a coding scheme that treats each VCM/S instance as a holistic unit consisting of a verb phrase (VP) + DE + VERB COMPLEMENT (VC). Compared with previous approaches to VCM/S constructions, which focused on either the VP or the VC, this approach allows us to better capture the co-occurrence patterns of VPs and VCs, and their degrees of complexity as a construction. In Fig. 1, the complexity scales of the VCM/S construction are presented in four main categories: (1) Basic VCM/S, (2) Simple VCM/S, (3) Medium Complex VCM/S, and (4) Most Complex VCM/S. Again for ease of discussion, the short forms S(1), S(2), S(3), and S(4) will be used to refer to these complexity scales, respectively.

S(1) Basic VCM/S, represented as V + DE + (HEN) ADJ, is the most prototypical VCM/S structure. It is composed of a plain V (single syllabic or disyllabic) and an adjective as VC, denoting and evaluating the manner of the action. Because normally the unstressed adverb 很 *hen* 'very' is obligatorily required to co-occur with the adjective, we consider the use of *hen* an instance of S(1) instead of S(2). Here is an example from our corpus: 他站得很高 *Ta zhan de hen gao* 'He stands tall'.

S(2) Simple VCM/S is structurally more complex than S(1) in that different types of adverbial modifiers on the complement are used to convey different degrees. Depending on the complexity of the modifier, S(2) has two formulas: 2.1, V + DE + PREVERBAL ADV + ADJ or V + DE ADJ + POSTVERBAL ADV, such as preverbal 非常 *feichang* 'extremely', 太 *tai* 'too', and 特别 *tebie* 'especially' and postverbal 极了 *jile* 'extremely' or 一点 *yidian* 'a little'; 2.2, V DE + ADV PHRASE

+ ADJ, such as comparative 越来越 *yuelaiyue* 'more and more' /比 *bi* 'more than'/ 更 *geng* 'even more'. Examples from the corpus are 2.1 他吃得非常快 *Ta chi de feichang kuai* 'He eats extremely fast', and 2.2 他写得越来越好 *Ta xie de yuelaiyue hao* 'He writes better and better'.

S(3) Medium Complex VCM/S conveys its construction complexity through both VP and VC. The VP complexity is seen in its two forms when an object is required: VOV (taking an Object after the V and another reduplicated V) or OV[3] (preposing the Object before the V). The VC complexity is achieved through a variety of means. In addition to using an adverb or adverbial phrase as modifiers in S(2), it can also use a wider range of grammatical and lexical forms to achieve VCM/S functions, such as a complex adjective (two adjectives combined), a VP, or an idiom. Taking into consideration the complexities of both VP and VC, S(3) has three forms: 3.1, VP (VOV) + DE + (ADV/ADV PHRASE) + ADJ; 3.2, VP (OV) DE + (ADV/ADV PHRASE) + ADJ; 3.3, VP (VOV/OV) + DE + COMPLEX ADJ/VP/IDIOM. An example of 3.1 is 他写字写得非常快 *Ta xie zi xie de feichang kuai* 'He writes very quickly'; an example of 3.2 is 他字写得不太好 *Ta zi xie de butai hao* 'His handwriting is not very good'; and examples of 3.3 include 妈妈变得不通人情 *Mama bian de butong renqing* 'Mom becomes unsympathetic', 他们要吃得上点档次 *Tamen yao chi de shang dian dangci* 'They want to eat fancier food', and 他同学长得又高又胖 *Ta tongxue zhang de you gao you pang* 'His classmate is tall and chubby'.

S(4) Most Complex VCM/S achieves its highest complexity in two ways: the use of combined constructions and of a clause as a complement. Combined constructions allow two constructions to co-occur in a VCM/S. The added construction is often a disposal/passive/causative construction, introduced by 把 *ba or* 将 *jiang*[4] (the disposal markers), 被 *bei* (the passive marker), or 让 *rang,* 使 *shi or* 将 *jiang* (the causative markers). The additional information is to further specify the manner of the action via disposal, passive, and causative means. The use of a clause as a complement further enhances the speaker's evaluative stance through an expressed agent or patient of the action. The three S(4) formula are 4.1, V DE + (ADV/ADV PHRASE) CLAUSE, 4.2, CONSTRUCTION (BA/JIANG/BEI) + VCM/S OF (O)V + DE + (ADV/ADV PHRASE) + ADJ/IDIOMS/VP/CLAUSE (ADV) and 4.3, CONSTRUCTION (JIANG/RANG/SHI) + (O)V + DE + (ADV /ADV PHRASE) + ADJ/IDIOMS/VP/CLAUSE. Each combined construction can take different forms depending on the specific configuration of VPs and Cs. Here are some examples from the corpus: 4.1, 风吹得人站不住脚跟 *Feng chui de ren zhan bu zhu jiaogen* 'The wind blows so hard that people cannot stand still'; 4.2, 把巧克力的甜腻细滑刻画得淋漓尽致 *Ba qiaokeli de tiannixihua kehua de linlijinzhi* 'give the fullest and most vivid depiction of the sweetness and smoothness of chocolates'; and 4.3, 你使自己

[3] See more detailed discussion on the comparison of the two structures in Sect. 5.1 on VCM/S complexity.

[4] 将 *jiang* has two functions in Chinese: one is a disposal marker as 把 *ba* and the other is a causative marker as 让 *rang* or 使 *shi.* It thus can appear in two different constructions, denoting different functions.

Table 2 Complexity scales and formula of VCM/S constructions

Complexity	Formula
S(1) Basic VCM/S	1. V DE + (HEN) ADJ
S(2) Simple VCM/S	2.1. V DE + ADV + A (ADV) 2.2. V DE + ADV PHRASE + ADJ
S(3) Medium complex VCM/S	3.1. VP (VOV) DE + (ADV/ADV PHRASE) + ADJ 3.2. VP (OV) DE + (ADV/ADV PHRASE) + ADJ 3.3. VP (VOV/OV) DE + (ADV/ADV PHRASE) COMPLEX ADJ/VP/IDIOM
S(4) Most complex VCM/S	4.1. V DE + (ADV/ADV PHRASE) CLAUSE 4.2. C (BA/JIANG/BEI) + (O)V + DE + (ADV/ADV PHRASE) + ADJ/IDIOMS/VP/CLAUSE (ADV) 4.3. C (JIANG/RANG/SHI) + (O)V + DE + (ADV/ADV PHRASE) + ADJ/IDIOMS/VP/CLAUSE

的生活变得丰富有趣 *Ni shi ziji de shenghuo bian de fengfu youqu* 'You make your life richer and more fun'.

As summarized in Table 2, we proposed a four-scale VCM/S complexity scheme with 9 structural formulas to reflect the VCM/S complexity on form-meaning mapping and distributional properties between VP and VC as observed in the corpora. Later, we will testify to the validity of the scheme as it is taken as a measure to discriminate L1 and L2 speakers' VCM/S usage patterns both contrastively and developmentally.

4 Results

With the data and coding systems in place, this study has yielded a number of interesting results about L2 learners' VCM/S construction acquisition in areas of construction frequency distribution, verb choices in VPs, complexity scales, as well as differences between L1 and L2 production. In the following sections, we will report these findings along the lines of our proposed research foci.

4.1 VCM/S Frequency Distribution by L1 and L2 Speakers

As shown in Table 3, a total of 424 instances of VCM/S constructions were identified in the corpus out of 1,284 composition samples.[5] The L1 data had the most VCM/S instances (117). L2 learners were able to use VCM/S constructions as early as at

[5] To compensate for the varying sizes of the sub-corpus, we calculated mean frequencies of VCM/S constructions by a number of samples.

Table 3 Frequency distribution of VCM/S constructions in the corpus

Proficiency levels	Number of samples	Number of VCM/S	Mean VCM/S per sample
L2-A	409	74	0.18
L2-B	248	59	0.24
L2-C	206	101	0.49
L2-D	273	73	0.27
L2 Mean	**1136**	**307**	**0.27**
L1	**148**	**117**	**0.79**
Total	**1,284**	**424**	

the Novice-Mid to Intermediate-Low level, during which the VCM/S constructions were introduced to learners where the data were collected.

The comparative data demonstrated that L1 speakers' VCM/S production had a relatively higher frequency (at 0.79 per sample), while L2 learners in general exhibited a tendency of VCM/S underuse (at 0.27 per sample on average). A two-sample *Z*-test for the L1 and L2 data shows that the result is significant ($Z = 12.7$, $p = 0.01$).[6] The underuse pattern is conspicuous in the L2-A group (at 0.18 per sample). L2-B learners only increased their VCM/S production moderately (at 0.24 per sample). L2-C and L2-D learners had relatively higher VCM/S production per sample (at 0.49 and 0.27) among all learner groups.

The type and token frequencies of VCM/S constructions were tabulated in Table 4. Several trends emerged here. To begin with, there was a clear difference between L2 learners' usage of VCM/S constructions and that of L1 speakers. L1 speakers used 34 VCM/S construction types and 117 construction tokens, while L2-D, our highest proficiency L2 group, had 24 types and 73 tokens. Among L2 learners, the higher the proficiency level, the more construction varieties were exhibited. This is reflected by the raw frequencies of construction types (11 for L2-A, 14 for L2-B, 14 for L2-C, and 24 for L2-D) and the type-token ratio (0.149 for L2-A, 0.237 for L2-B, 0.139 for L2-C, and 0.329 for L2-D).

The data on VCM/S production frequency and its distribution indicated a significant difference between L1 and L2 speakers who exhibited underperformance in VCM/S types, tokens, and production quantity.

[6] We wish to thank Johnny Lin of UCLA Institute for Digital Research and Education for his statistical advice.

Table 4 VCM/S constructions by type and token frequency

Proficiency levels	VCM/S construction type	Number of samples	Average number of types by sample	VCM/S construction token	Average number of tokens by sample
L2-A	11	409	0.0269	74	0.1486
L2-B	14	248	0.0565	59	0.2373
L2-C	14	206	0.0680	101	0.1400
L2-D	24	273	0.0879	73	0.3288
L1	34	148	0.2297	117	0.2906

4.2 Verb Choices in the VCM/S Construction by L1 and L2 Speakers

Next, we report the results of verb choices in terms of type and token frequencies in VCM/S constructions. In our corpus, both L1 and L2 speakers' data included, there are altogether 117 verbs by type and 423 verbs by token.[7]

Based on our verb coding system, Fig. 2 shows the distribution of verb types by frequency band. Compared with the L2 data, the L1 speakers produced a markedly larger proportion of D-list verbs (63.1%), indicating a large lexical repertoire of verbs. In contrast, the proportions of A-, B-, and C-list verbs in the L1 data are quite small, about 36.9% combined. The L2 learner data, on the other hand, have an overall large proportion of A-, B-, and C-list verbs. The proportion of C-list verbs is noticeably high, indicating L2 learners' limited lexical repertoire of verbs. Compared with L1 speakers, L2 learners mainly relied on common, high-frequency, and general verbs while lacking a wide range of specific and abstract verbs. L2-D learners' distribution pattern is approaching L1 speakers but differed in quantity (44.1% vs. 24.6% for C-list and 32.4% vs. 63.1% for D-list), meaning the advanced learners like L2-D are still expanding their verbal repertoire and are beginning to accumulate more specific verbs.

The distribution of verb tokens by frequency band is presented in Fig. 3. Two opposite trends are worth noting. First, we can see more clearly that L2 learners across proficiency levels used a noticeably large number of A- and B-list verbs (64.8% for L2-A, 45.8% for L2-B, 75% for L2-C, and 57.5% for L2-D). Considering that there are only 11 verbs in the A- and B-list combined, the repeated use of these verbs (变 *bian* 'become', 考 *kao* 'test', 吃 *chi* 'eat', 说 *shuo* 'speak', 做 *zuo* 'do', 看 *kan* 'look', 过 *guo* 'lead [a life]', 长 *zhang* 'grow', 玩 *wan* 'play, have fun', 发展 *fazhan* 'develop', and 学 *xue* 'study') by L2 learners indicates learners' attention toward prototypical verb choices in VCM/S constructions. Second, on the opposite end of the scale, we notice

[7] The total number of verbs are 423 instead of 424 because in the L2-C group, there is a case of missing verb.

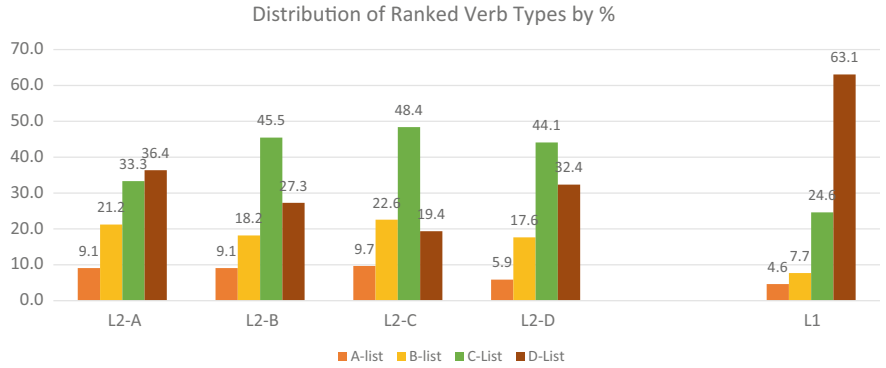


Fig. 2 Distribution of verb types by frequency band

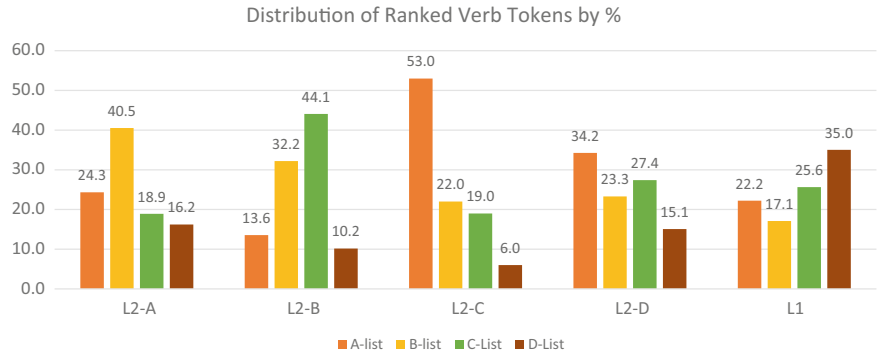


Fig. 3 Distribution of verb tokens by frequency band

that L2 learners' production of D-list verb tokens is almost negligible (16.2% for L2-A, 10.2% for L2-B, and 6% for L2-C). Even L2-D, the highest proficiency group in the corpus, only produced 15.1% worth of D-list verbs.

4.3 Complexity Scales of VCM/S Constructions by L1 and L2 Speakers

While the frequency data of VCM/S constructions and verb choices yielded interesting distributional and developmental patterns, we now turn to the issue of VCM/S complexity based on our proposed VCM/S complexity scales described in Sect. 3.2.

Here, we report three complexity-related results: (a) the VCM/S complexity distribution, (b) VP co-occurrence patterns, and (c) VC form-meaning mapping mechanisms. These three parameters are examined quantitatively and qualitatively as well as contrastively and developmentally.

In terms of the VCM/S complexity distribution, Fig. 4 reveals several useful trends about differences in production complexity. First, there is again a substantial difference between L1 and L2 production in terms of the VCM/S complexity. A chi-square test shows statistically significant between-group differences ($X^2 = 26.2$, $df = 12$, $p < 0.0001$). The L1 data exhibited an even and growing trend in the distribution of the four scales with S(4) VCM/S constructions, making the largest proportion. 59% of VCM/S constructions produced by Chinese L1 speakers belonged to S(3) and S(4), which are high-level complexity VCM/S constructions to convey more sophisticated and nuanced meanings. On the other hand, the L2 learner data showed quite different distribution patterns from L1 speakers as well as among themselves across proficiency levels. Unsurprisingly, L2 learners were generally restricted to S(1) and S(2) VCM/S constructions, especially at the first three levels, with the proportion of S(2) VCM/S constructions being particularly noticeable among intermediate and advanced learners (41.9% for L2-A, 47.5% for L2-B, and 62.4% for L2-C). S(3) VCM/S constructions presented an interesting U-shaped trajectory, which began with a higher percentage of 39.2% for L2-A, and later dipped down to 15.3%, 19.8%, and 16.4% for L2 B, C, and D, then went up to 23.9% for L1 speakers. Such a trend will be further discussed in Sect. 5.

Given the lower numbers in some of the categories (e.g. zero of the S(4) scale in L2-A), Fisher's Exact Test was conducted to identify significant intra- and inter-group differences. The results, as shown in Table 5, confirm two visual impressions obtained from inspecting the chart in Fig. 4: S(2) and S(4) VCM/S constructions seem to show important statistical information for intra- and inter-group differences.

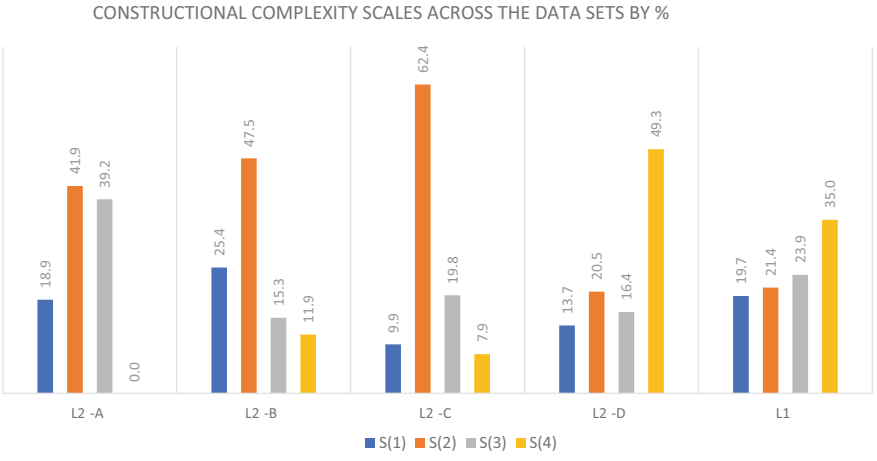


Fig. 4 Distribution of VCM/S constructional complexity scales by percentage

Table 5 Inter- and intra-group differences in VCM/S constructional complexity scales

	L2-A	L2-B	L2-C	L2-D	L1
S(1)	0.612	0.090	0.033	0.495	0.386
S(2)	0.511	0.148	**<0.0001**	**0.001**	**<0.0001**
S(3)	**0.001**	0.137	0.418	0.169	0.798
S(4)	**<0.0001**	0.060	**<0.0001**	**<0.0001**	**<0.0001**

Fisher's Exact Test, *p*-values: values displayed in bold are significant at the level alpha = 0.01.

We take the data to show that while S(2) VCM/S constructions clearly differentiate learners of L2-A, B, and C from L2-D and L1 speakers, S(4) VCM/S constructions further align L2-D learners with L1 speakers in terms of construction complexity.

In terms of VP co-occurrence patterns, our data showed an interesting difference between L1 and L2 speakers in their use of VO and OV in a VCM/S. While L1 speakers in our data used 100% OV form and 0% VO when an object is required, L2 learners' data, however, presented a mixed picture across proficiency levels. L2-A used VO and OV about 50% each with many repeats and errors, L2-B used VO 100%, L2-C VO 80% and OV 20%, and L2-D's pattern was VO 40% and OV 60%. This clearly indicates developmental changes as proficiency increases; L2 learners at higher levels seemed to have gone through a process of structure switching from VO to OV at the VP level. A detailed discussion on the peculiar L2-A data and switching process is offered in Sect. 5.

In terms of VC form-meaning mapping mechanisms, the data revealed a differential preference between L1 and L2 speakers along complexity scales. As described in Sect. 3.2, VCM/S has a variety of ways to map VC forms to their functions, such as the manner of action and the speaker's evaluative stance. Our qualitative analysis indicated that L1 and L2-D speakers preferred to use 5 high complexity types of VC pairings, some are lexical and others are structural mechanisms, to achieve nuanced VCM/S functions, especially for evaluative and affective stances. Examples of VC pairings from our data include (a) complex adjectives as 变得开朗与乐观 *bian de kailang yu leguan* 'become outgoing and optimistic', (b) verb phrases as 长得很像新疆人 *zhang de hen xiang xinjiangren* 'look very much like a Uyghur', (c) idioms as 听得耳熟能详 *ting de er-shu-neng-xiang* 'hear something frequently to the extent that it becomes very familiar', (d) clauses as 风吹得人站不住脚跟 *feng chui de ren zhanbuzhu jiaogen* 'The wind blew so hard that a person could not stand on their feet', and (e) double constructions of VCM/S co-occurring with disposal/passive/causality structures as 你使自己的生活变得丰富有趣 *ni shi zijide shenghuo biande fengfu youqu* 'you made your life both rich and interesting'. Among them, idioms were the most frequently used form by L1 speakers and L2-D learners for subtle meanings of VCM/S constructions. However, lower level L2 speakers (L2-A, -B, and -C) were found to prefer VC pairings within S(1) and S(2) and lexical modifiers for VCM/S function mapping. Examples of VC modifiers from our data include (a) adverbs as 太 *tai* 'too', 非常 *feichang* 'very', 特别 *tebie* 'especially', or 最 *zui* 'most', (b) adverbial phrases as 越来越 *yuelaiyue* 'more and more', 一天比一天 *yitian bi yitian* 'day by

day', and (c) comparative structures as 比…A *bi…A* 'more than', and 跟… (不)一样 *gen…(bu)yiyang* '[not] the same as', among others. As proficiency increased from L2-A to L2-B, and finally to L2-C, these mechanisms gradually expanded to include types with S(3) and S(4) complexity that use structural means to achieve VCM/S functions (as seen in Fig. 4).

To sum up, while L2 learners developed some fundamental abilities to use VCM/S constructions with increased complexity, it takes a long time for them to reach the L1 native level.

5 Discussion

The goal of the current study is to investigate the usage of the Chinese VCM/S construction within usage-based approaches to language acquisition. To our knowledge, this is the first comparative corpus study on construction learning that addresses psycholinguistic factors of frequency, co-occurrence properties, and form-meaning mapping between L1 and L2 speakers. Our discussion will focus on two areas based on our results and our proposed research foci: (a) similarities and differences between L1 and L2 VCM/S production in terms of the VCM/S frequency distribution, verb choices, and construction complexity; and (b) L2 VCM/S developmental patterns.

5.1 Similarities and Differences in VCM/S Production Between L1 and L2 Speakers

Regarding the VCM/S frequency distribution, our results indicate a marked difference between L1 and L2 VCM/S production. L2 learners produced substantially fewer VCM/S constructions than L1 speakers and their average token and type ratio per sample are much lower as compared with L1. This pattern persists across all proficiency levels, and the gap is not closed even when advanced proficiency is reached. Such results can be interpreted in several ways. First, our results of L2 persistent underperformance do not support D. Sun (2002), Feng (2013), and Jiang (2019)'s claim that VCM/S constructions are among the easiest verb complement constructions to learn for L2 learners but provide partial support to Zhou and Deng (2009)'s finding that VCM/S is a complex construction and certain features may not be acquired even if learners reach advanced proficiency. We postulate that different conclusions may be due to different sample sizes used, different analytical foci on VCM/S constructions, such as dynamic statistical information or static grammatical structures, and the interpretation of 'easiness' through limited measures (such as raw frequency and errors) without further examining construction complexity, verb choices, forms, and functions. Second, VCM/S learning is indeed a process of construction learning, whose usage experience and input exposure can lead to

quantitatively and qualitatively different production between L1 and L2 speakers. It is thus important to examine the VCM/S learning from usage-based approaches and to recognize the roles of both implicit (e.g. frequency, form-meaning mapping) and explicit (classroom instruction) learning factors. Additionally, the persistent L2 underperformance signals unique L2 learning needs in areas of systematic language use, expansion of lexical repertoire on construction verbs and complements, targeted attention-directing on nuanced VCM/S form-function mapping, and effective instruction at different stages.

Regarding verb choices, both similarities and differences are observed between L1 speakers and L2 learners. The differences are shown in that L2 learners tend to choose verbs which are limited in number but high in frequency (A- and B-lists), while L1 speakers have a balanced and diverse range of high- and low-frequency verbs but a clear preference for highly specific verbs (C- and D-lists) for their VCM/S constructions. The similarity is seen in L2 learners' repeated use of 变 *bian* 'become' among other 11 high-frequency verbs. First of all, such results point to L2 learners' increasing sensitivity toward and purposeful selection of prototypical verbs in VCM/S constructions as L1 speakers. The fact that prototypical verbs are used by L2-C and L2-D learners corroborates two existing findings in the field of second language acquisition: (a) highly frequent, salient, and prototypical verbs help L2 learners extract shared typical features among learned constructions that eventually help anchor the abstract construction category (Casenhiser and Goldberg, 2005; Childers & Tomasello, 2001); and (b) the prototypicality effect often does not take effect until certain input exposure or proficiency level is reached, such as Intermediate-Mid or higher in our study (Kellerman, 1979; Year & Gordon, 2009). Second, the results provide important L2 evidence to Tao et al. (2020)'s L1 study that shows the verb *bian* is ranked first in frequency and is the most prototypical verb in L1 VCM/S production, accounting for 17% of VCM/S tokens. Third, our data indicate a clear difference in vocabulary range that exists between L1 speakers and L2 learners. L2 learners' overall weakness to use specific verbs for VCM/S constructions and their limited verb choices point to the need for L2 learners to expand their lexical repertoire to convey more nuanced evaluative meanings in more complex VCM/S constructions.

Regarding construction complexity, our results again demonstrate a significant difference between L1 and L2 data measured by a 4-scale complexity scheme proposed by this study. Such a new complexity measure system allows researchers to examine the VCM/S development at both construction and component levels and also to delineate factors in VCM/S complexity distribution patterns, VP co-occurrence patterns, and VC form-function mapping in relation to complexity scales.

With VCM/S complexity distribution, L1 speakers and L2 learners are divided along the complexity scales. While lower level L2 learners tend to adhere to basic S(1) and simple S(2) complexity VCM/S constructions, L1 and L2-D speakers prefer to explore a wide variety of VCM/S constructions with low to high complexity, especially on the high end of S(3) and S(4). Such results uncovered several important findings. First, L2 VCM/S learning is closely linked to VCM/S complexity. Extensive usage and experience as reflected in proficiency levels are the major driving force

in increasing construction complexity (Ellis, 2012). Second, our L2 data analyses on complexity demonstrate that the construction complexity scheme proposed by this study can be used reliably in measuring VCM/S learning. The two high-yield scales are simple S(2) and most complex S(4). S(2) is the best indicator for VCM/S developmental changes at different stages and S(4) signals the gap and necessary alignment between L2 and L1 production. Third, our complexity analyses unveil two unusual and seemingly counter-intuitive VCM/S learning patterns. The first has to do with the U-shaped S(3) data distribution (gray bars in Fig. 4). The second is the L2-C data (orange bars in Fig. 4), which seems to deviate from the rest. We argue that these patterns reflect unique second language developmental stages commonly found in SLA studies (Gass & Selinker, 2013; N. Ellis, 2012). A detailed discussion will be provided in Sect. 5.2, which focuses on developmental patterns.

With VP co-occurrence patterns, an interesting L2 learning pattern emerges, showing differences from L1 speakers. When an object is required in a VCM/S, L2 learners, especially at lower levels, tend to use the VO form to co-occur with its complement, while L1 speakers and L2-D learners prefer the OV form. First, such results match the findings by Zhou and Deng (2009) in that L2 learners whose L1 languages are either alphabetical (Thai, Vietnamese) or non-alphabetical (Japanese and Korean) also tend to use VO in VCM/S constructions at all levels and even at the more advanced levels. Second, both our study and Zhou and Deng (2009) agree that L2 learners' VCM/S learning seems to have started in an unnatural and incorrect order, that is, from a complex VOV form before switching to a simpler and native-like OV form. Grammatically, either VO or OV can co-occur with a complement in a VCM/S, except that VO requires an extra step of verb reduplication into VOV before it can co-occur with other VCM/S components. VOV is thus argued to be more complex than OV structurally. VOV requires a reduplication transformation, whereas OV involves simple movement from postverbal to pre-verbal positions (Zhou & Deng, 2009). Based on our data and observation, we argue that the unnatural VOV usage is a possible result of formal and explicit instruction that needs to be further investigated. As is claimed by Bley-Vroman (1991) and DeKeyser (2003), adult L2 learners of formal classrooms are exposed to two types of input: (a) explicit input from text-books and classroom instruction, and (b) implicit input from language use involving frequency, co-occurrence patterns, and form-meaning mapping processes. Adult L2 learners' VOV preference is likely induced by their explicit formal instruction and textbook exposure. As L2 learners experience more Chinese, they begin to notice the gap between theirs and the native version. This awareness will trigger a restructuring process to help switch the usage to a more native-like version. This restructuring process is in line with SLA studies in Lightbown (1985) and McLaughlin and Heredia (1996).

With VC form-function mapping mechanisms, a differential preference is observed between L1 speakers and L2 learners. While lower level L2 learners like to use low complexity lexical means (adverbs and adverbial phrases) to map VC pairings to VCM/S functions, such as the manner of the action, L1 speakers and advanced L2 learners prefer to use highly complex structural means (idioms, clauses, and combined constructions) for VC pairings to map a full range of nuanced VCM/S

functions. Idioms and combined constructions (the *ba/bei/shi/rang* constructions) are also found to be abundant in L1 production. Such results directly point to several acquisitional implications. First, L1 speakers and L2 learners seem to use different mapping mechanisms (lexical vs. structural) for VCM/S functions. This result is in line with findings by Casenhiser and Goldberg (2005), who discovered that their L2 learners' construction forms first mapped to concrete functions before mapping onto more abstract and subtle functions. Second, our study finds that idiom use is a good indicator for judging whether a VCM/S is intermediate or advanced VCM/S usage. This result also corroborates the findings by Tao et al. (2020) that Chinese native speakers prefer to use complex VC pairings, such as idioms, VPs, and clauses, to convey more subtle and sophisticated evaluative and affective stances.

In summary, our study found similarities but marked differences between L1 and L2 VCM/S production in areas of the VCM/S frequency distribution, verb choices, and construction complexity. These differences point to the fact that VCM/S construction learning is experience-based and learners are sensitive to psycholinguistic factors of frequency, form-meaning mapping, and co-occurrence patterns, but it is also influenced by L2 learners' explicit formal exposure.

5.2 L2 VCM/S Developmental Patterns

The second focus of our discussion is on L2 VCM/S developmental patterns within usage-based approaches to language learning. We address two development-related issues: (a) the L2 VCM/S U-shaped learning patterns, and (b) two unique but independent stages of construction learning: the formulaic stage and the input-induced conservative stage.

As for L2 VCM/S learning patterns, we alluded in Sects. 4.3 and 5.1 that the L2 VCM/S learning process follows a typical U-shaped learning pattern known in the field of second language acquisition. This learning model was proposed and researched mainly by two groups of scholars under different frameworks and at different times.[8] We combine the two approaches to help explain our data. According to Gass and Selinker (2013), U-shaped learning refers to a L2 learning curve across three distinctive stages. The learning normally begins with a high-performance level (Stage 1) and over time it descends to a lower level (Stage 2). After another period of time, the performance once again ascends to a higher level qualitatively (Stage 3). N. Ellis (2012) further expanded the model with new elements and details. He adapted the U-shaped learning model to a three-stage process of construction development, which has distinct and unique stage characteristics. That is from formula to low-scope slot-and-frame pattern, to creative construction (see Fig. 5). Stage 1 is characterized by seemingly high performance in formulaic sequences, Stage 2 is characterized by

[8] Gass and Selinker represent the initial studies of the model within the framework of universal grammar and N. Ellis and others represent the construction learning research under usage-based approaches in recent times.

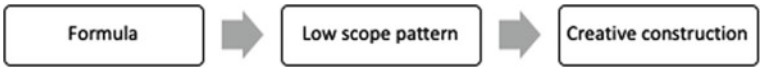


Fig. 5 The three-stage sequence model of construction learning

limited low-scope patterns and open slots to place elements with shared similarities, Stage 3 is characterized by extensive use of constructions creatively and in a wide variety. The transition from one stage to another is motivated by the need of forming and extending construction categories.

We now use the U-shaped construction learning model by Ellis (2012) to explain our L2 statistical data represented by gray bars (Scale 3 complexity) in Fig. 4 of Sect. 4.3. As is shown by the data, the learning curve starts with L2-A learners' high performance of 39.2% S(3) VCM/S constructions, then dips down to the bottom at 15.3% for L2-B, and gradually rises to 19.8% for L2-C, and finally ends at 20.5% for L2-D. We argue that our data reflect exactly a U-shaped VCM/S learning pattern. During this process, L2-A learners start with a rich repertoire of VCM/S formulaic sequences, possibly accumulated from their classroom input and textbooks (Stage 1). Because these formulaic sequences are used with rather high frequency, such as 变得很快 *bian de hen kuai* 'change quickly', and with prototypical verbs, such as *bian* 'change', these sequences quickly gain special statistical status and become concrete similarities (V *de* A) for a prototypical pattern that seeds a VCM/S category. This categorical formation then allows open slots in VP and VC to be substituted with similar elements in form and function, e.g. 变得很多 *bian de hen duo* 'change a lot' or 吃得很快 *chi de hen kuai* 'eat fast', generating different types of low-scope VCM/S patterns (Stage 2). The frequent usage of these low-scope patterns will soon increase in number and be extended to a full range of VCM/S constructions to be used productively and creatively (Stage 3). However, State 3 will last a long time before reaching the native level. We believe that this is the construction learning path that our L2 learners have gone through for VCM/S constructions and have resulted in the production data in our study.

As for the formulaic stage, we attempt to verify if L2-A learners' seemingly high performance on complex S(3) production is full-grown VCM/S constructions or memorized formulaic sequences. Given that L2-A consists of adult learners in formal classrooms and their VCM/S exposure is limited to 2–3 semesters, the complex S(3) VCM/S constructions are theoretically unlikely but in reality exist in the data. We argue that the S(3) production by L2-A is not full-fledged VCM/S constructions but are predominantly formulaic sequences taken from the input. Our evidence comes from three areas. First, the large repertoire of VCM/S constructions produced at this stage is mostly short, repetitive, and not productive. 70% of these VCM/S constructions occur only once, and many are obvious set phrases taken directly from textbooks or classroom instruction. Second, the VCM/S constructions produced by L2-A learners at this stage, though complex on the surface, are rather fixed sequences that are not breakable and cannot generate new forms. For example, 甜食吃得越来越少 *tanshi chide yuelaiyue shao* 'gradually, sweets are consumed less and less' is

a S(3) with a OV + ADV PHRASE structure that is repeated verbatim 7 times. There are no traces of open slots to substitute for new elements in these VCM/S constructions. According to Ellis (2012), the use of open slots is a crucial signal for moving beyond the formulaic stage and forming a construction category. We find S(3) VCM/S constructions are in fact memorized and unanalyzed formulaic strings taken verbatim from the textbooks after verifying from classroom teachers, e.g. the idiom use of 吃得津津有味 *chide jinjinyouwei* 'taste deliciously'. For purpose of further verifying our claim, we call for online empirical studies on VCM/S constructions.

As for the input-induced conservative stage, we attempt to use our L2-C data to explain this unique acquisition stage and its developmental trajectory. According to Goldberg and Boyd (2015), input-induced conservatism refers to a statistical preemption process where learners avoid using certain well-formed but slightly different constructions because similar constructions have been systematically witnessed. We argue that L2-C data on VCM/S frequency distribution, verb choices, and construction complexity mirror exactly this type of conservative stage. As Table 3 and Figs. 2, 3, and 4 indicate, L2-C has the highest VCM/S tokens (101), but these VCM/S constructions carry low types (14) as L2-B. Their verb choices fall on mostly high-frequency ones (75% of A-, B-, and 25% C-, D-lists). Their VCM/S complexity remained predominantly on S(1) basic to S(2) simple scales (72.3%). Clearly, L2 learners at this stage tend to clutch on familiar VCM/S uses and avoid using VCM/S constructions that are different. This seemingly low performance manifests a typical input-induced conservatism according to Goldberg and Boyd (2015) and is found in similar studies by Clark and Clark (1979) and Ellis (2012), where learners resort to limited but familiar VCM/S types and complexity. Ellis (2012) terms this conservatism as 'Teddy Bear phenomenon' and argues that the clutching to 'teddy bear constructions' indicates the learner's efforts to further form and consolidate the construction category. On the other hand, Markman and Gentner (1993) and Goldberg et al. (2007) argue that this is a typical instructed L2 developmental stage with restricted production diversity. Explicit instruction and classroom exemplars tend to lead L2 learners to narrowly focus on their familiar criteria governing the category membership. As L2 learners' proficiency level goes up, they will break this conservatism and expand the construction category widely as L2-D learners did.

To sum up, our findings above contribute to the field in three ways: (a) the U-shape learning exists in VCM/S development and the model by N. Ellis is valid in explaining the construction learning stages and their transitions from one to another; (b) adult L2 construction learning involves both implicit associative learning and explicit classroom instruction, which often dictate the course and characteristics of the formulaic stage and input-induced conservative stage, and (c) construction complexity analyses are the key to understanding construction development, and our proposed VCM/S complexity scale scheme is an important contribution to the field.

6 Conclusions

This comparative study on L1 and L2 Chinese VCM/S production and development contributes to the research on construction learning with three findings: (a) there are marked quantitative and qualitative differences between L1 and L2 VCM/S production at both construction and component levels; (b) these persistent productive differences reflect the indispensable roles of psycholinguistic factors, such as frequency, complexity, form-meaning mapping, and co-occurrence patterns of VP and VC, especially on verb choices; and (c) L2 VCM/S construction learning is like any other construction learning that followings a U-shape learning path that consists of unique and distinctive stages. The process also involves both implicit factors and explicit classroom input and instruction. Such a study has important theoretical and pedagogical implications for the field of second language learning and teaching in general and for Chinese as a second language field in particular.

Theoretically, to our knowledge, this study represents the first attempt that examines the Chinese VCM/S as a construction using usage-based approaches to language acquisition. While existing literature has mainly focused on a single component of VCM/S in isolation, in particular either the VP or VC, this study examined VCM/S acquisition as constructions both at the construction level and at the component level. At the construction level, we proposed a complexity scale of VCM/S constructions that encompasses four major categories and nine formulae. At the component level, we examined the two open slots of VP and VC jointly. This approach not only allowed us to capture the finer path of construction development, from simple to increasing complexity of VCM/S constructions structurally and semantically, but it also provided us a window to observe L2 learners' prototypical choices of VP and co-occurrence patterns on VC in the two open slots, as well as the range of form-meaning mapping mechanisms by L1 and L2 speakers. Second, this study utilized a large-size corpus with both L1 and L2 speakers' production data. The L2 learners' proficiency levels covered a wider range of the spectrum, from Novice-Mid to Advanced-Mid and higher. This allowed us to examine L2 speakers' usage patterns of VCM/S constructions both in comparison with L1 speakers and across developmental stages. The production data supported our proposed categorization and complexity scales of VCM/S as a construction. This study proves that corpus can be used in a fruitful way in examining the L2 acquisition of linguistic constructions in the usage-based paradigm.

Pedagogically, several implications can be drawn from the present study. First, CSL instructors should understand the complexity involved in the CSL learning process of VCM/S construction and should constantly monitor learners' language use experience and encourage learners to notice the possible open slots and their co-occurrence rules for form-meaning pairs with each VCM/S. Second, CSL instructors should use adequate input exemplars with strategic frequency (containing fewer number of construction types but an abundance of tokens) to help learners establish a correct mental category of basic and simple VCM/S constructions early on. Then at a later stage, use diverse input and purposeful attention-direction to expand

the VCM/S category to include many different types of nuanced VCM/S functions. Finally, explicit classroom instruction can lead learners to a possible narrow or excessive focus on a single or irrelevant dimension of the construction. CSL instructors should be more conscious of exposing learners to a wide range of native-like VCM/S construction types and at different complexity scales, especially at advanced levels.

Appendix: List of Verbs of Different Frequency Bands (Type = 117 and token = 423)

A	变	85
	考	23
	吃	22
B	说	17
	做	16
	看	16
	过	13
	长	13
	玩	12
	发展	11
	学	10
C	开	9
	走	9
	写	6
	打	6
	进行	6
	恢复	5
	站	5
	听	4
	唱	4
	表现	4
	跑	4
	弹	3
	搞	3
	活	3
	演	3
	体现	2
	去	2
	发	2

(continued)

(continued)

	安排	2
	弄	2
	惹	2
	扩大	2
	死	2
	涨	2
	熏	2
	爬	2
	睡	2
	管	2
	经过	2
	耐	2
	锻炼	2
	骂	2
D	买	1
	争	1
	住	1
	决定	1
	准备	1
	刮	1
	到	1
	刷	1
	刻画	1
	办	1
	包	1
	反应	1
	变化	1
	叫	1
	吓	1
	吹	1
	呕吐	1
	呛	1
	困	1
	失败	1
	定	1
	害	1
	差	1
	帮	1

(continued)

(continued)

	干	1
	心-信	1
	忙不迭	1
	想	1
	想象	1
	战	1
	打扫	1
	打扮	1
	扫	1
	折	1
	抛	1
	捏	1
	接近	1
	摔	1
	旅游	1
	显	1
	晒	1
	来	1
	梳	1
	模仿	1
	洗	1
	演戏	1
	烘	1
	生	1
	生活	1
	用	1
	画	1
	留	1
	皱	1
	相处	1
	穿	1
	笑	1
	管理	1
	绽放	1
	落	1
	行	1
	表演	1
	装点	1

(continued)

(continued)

	见	1
	订	1
	讲	1
	读	1
	谈谈	1
	起	1
	跳	1
	踮	1
	蹭	1
	运算	1
	进步	1
	适应	1
	逼	1
	飞	1
	骑	1

References

ACTFL. (2012). ACTFL Proficiency Guidelines 2012. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf.

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62*(1), 67–82.

Bybee, J., & Hopper, P. J. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. John Benjamins.

Bley-Vroman, R. (1991). The logical problem of foreign language learning. *Linguistic Analysis, 20*(1–2), 3–49.

Casenhiser, D., & Goldberg, A. E. (2005). Constructional fast mapping. *Berkeley Linguistic Society*.

Chao, Y. R. (1968). *A Grammar of spoken Chinese*. University of California Press.

Childers, J. B., & Tomasello, M. (2001). The Role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology, 37*(6), 739–748.

Clark E. V., & Clark, H. (1979). When nouns surface as verbs. *Language, 55*(4), 767–811.

Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 23*, 696–708.

DeKeyser, R. (2003). Implicit and explicit learning. In J. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 313–348). Oxford, MA: Blackwell.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17–44.

Ellis, N. C., & Larsen-Freeman, D. (2009). Language as a complex adaptive system (Special Issue). *Language Learning, 59*(Supplement 1).

Ellis, N. C. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned-attention, and the limited L2 end state. In P. Robinson & N. C. Ellis

(Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372–405). New York & London: Routledge.
Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*(1), 1–24.
Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188.
Feng, L. 冯丽莉. (2013). "De Zi Buyu Leixing de Xide Nandu yanjiu" 得字补语类型的习得难度研究 (Research on learning difficulties of *De* complement constructions). Shanghai Normal University MA Thesis.
Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language, 80*, 748–775.
Garnsey, S., Pearlmutter, N., Myers, E., & Lotocky, M. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.
Gass, S., & Selinker, L. (2013). *Second Language Acquisition*. New York & London: Routledge.
Goldberg, A., & Boyd, J. (2015). A-adjectives, statistical preemption, and evidence: Reply to Yang (2015). *Language, 91*(4), 184–187.
Goldberg, A., E., Casenhiser, D., & White, T. (2007). Constructions as categories of language. *New Ideas in Psychology, 25*, 70–86.
Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7*(5), 219–224.
Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago & London: University of Chicago Press.
Hoey, M. P. (2005). *Lexical priming: A new theory of words and language*. London & New York: Routledge.
Jiang, T. 姜天琦. (2019). "Qingtai Buyu de Xide Yanjiu" 情态补语的习得研究 (Research on complement constructions of manner). Yangzhou Normal University MA thesis.
Kellerman, E. (1979). Transfer and non-transfer: Where we are now. *Studies in Second Language Acquisition, 2*, 37–57.
Li, C., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley, Los Angeles, & London: University of California Press.
Lightbown, P. (1985). Great expectations: Second language acquisition research and classroom teaching. *Applied Linguistics, 6*, 173–189.
Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*(4), 431–467.
McLaughlin, B., & Heredia, R. (1996). Information-processing approaches to research on second language acquisition and use. In W. Ritchie & T. Bhatia (Eds.), *Handbook of Second language acquisition* (pp. 213–228). Academic Press.
Shen, J. (2003). Xiandai Hanyu dongbu jiegou de leixingxue kaocha [A typological investigation of verb complement constructions in modern Chinese]. *Shijie Hanyu Jiaoxue, 2003*(3), 17–23.
Sun, D. 孙德金. (2002). "Waiguo Liuxuesheng Hanyu '*De*' Zi BuyuXide Kaocha" 外国留学生汉语"得"字补语句习得情况考察 (An investigation on CFL learners' acquisition of 'DE' complements). 语言教学与研究 *Yuyan Jiaoxue yu Yanjiu, 6*, 42–50.
Sun, Q. 孙群. (2018). Oumei Xuesheng Hanyu Qingtai Buyu Ju Xide Yanjiu 欧美学生汉语情态补语句习得研究 (L2 acquisition on Chinese complement constructions of manner by CSL learners of European languages). 海外华文教育 *Haiwai Huawen Jiaoyu,* 2018.101/6. 59–67.
Tao, H., Jin, H., & Zhang, J. (2020). A corpus-based investigation of manner/state complement constructions in Mandarin Chinese. To appear in Bianca Basciano, Franco Gatti, and Anna Morbiato (Eds.), *Corpus-based research on Chinese language and linguistics, Sinica venetiana* (pp. 1–40). Edizioni Ca' Foscari Digital Publishing, Venice, Italy: Università Ca' Foscari Venezia (Ca' Foscari University) Press.

Trueswell, J., & Tanenhaus, M. (1994). Toward a lexical framework of constraint-based syntactic ambiguity resolution. In C. Clifton & L. K. R. Frazier (Eds.), *Perspectives on sentence processing* (pp. 155–179). Erlbaum.

Trousdale, G., & Hoffmann, T. (Eds.). (2013). *Oxford handbook of construction grammar*. Oxford: Oxford University Press.

Tily, H., Gahl, S., Arnon, I., Kothari, A., Snider, N., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language & Cognition, 1*, 147–165.

Year, J., & Gordon, P. (2009). Korean speaker's acquisition of the English ditransitive construction: The role of verb prototype, input distribution, and frequency. *The Modern Language Journal, 93*, 399–417.

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

Zhou, X. 周小兵, & Deng, X. 邓小宁 (2009). Liangzhong 'De' Zi Buyuju de Xide Kaocha (An investigation on two types of complement constructions of manner). 汉语学习 *Hanyu Xuexi, 2*, 65–71.

The Development of Relative Clauses in L2 Chinese: A Corpus-Based Study

Li-ping Chang

Abstract Acquisition of relative clauses (RCs) in a second language (L2) has long been a popular research focus, particularly in determining whether L2 learners' acquisition of RCs conforms to the Noun Phrase Accessibility Hierarchy (NPAH) (Keenan & Comrie, 1977), which proposes that subject-extracted RCs are the easiest to learn because they are the most commonly produced RC type with the fewest error rate. Early studies have mostly focused on Indo-European languages, especially English. In this study, we adopt a corpus-based approach to analyze the distribution of subject-extracted RCs (SRCs) and object-extracted RCs (ORCs) by Chinese learners with six different L1s and at two proficiency levels to test whether SRCs are easier than ORCs for Chinese L2 learners. The corpus we used is the Test of Chinese as a Foreign Language (TOCFL) Learner Corpus comprised 4,709 compositions written by test-takers of the writing section. A total of 2,055 RCs are analyzed, including 1,362 RCs at the CEFR-B1 (intermediate-high) level and 693 RCs at the CEFR-B2 (advanced) level by native speakers of English, Spanish, Japanese, Korean, Vietnamese, and Indonesian, representing three different language typologies. From the perspectives of RCs occurring in the grammatical position in the matrix sentence and the animacy of the head noun, the results show that ORCs for Chinese L2 learners are easier than SRCs. These results go against the NPAH hypothesis. In addition, no matter what branching types (i.e. left, right, or left-and-right) the learner's native language was, all lower-proficiency level language learners produced more ORCs than SRCs. These results coincide with the development pattern of RCs for L1 Chinese acquisition. Therefore, we propose that the dominant factor in learning Chinese RCs is word order, since ORCs have the same SVO word order as Chinese simple sentences. Regardless of learners' language background, learners can produce ORCs more naturally and with more ease. After the L2 language proficiency increases, SRCs will take over that advantage and learners' language use will become gradually closer to the target language.

L. Chang (✉)
Graduate Program of Teaching Chinese As a Second Language, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 106, Taiwan
e-mail: lchang@ntu.edu.tw

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_7

Keywords Relative clause · L2 acquisition · Interlanguage · Learner corpus · Mandarin Chinese

1 Introduction

Ever since Keenan and Comrie (1977) proposed their Noun Phrase Accessibility Hierarchy (NPAH), the relative clause (RC) has received special attention from linguists and language acquisition researchers. Using the linguistic typology approach, Keenan and Comrie conducted a thorough survey of over 50 different languages and proposed the following hierarchy of relativized noun phrases: SU>DO>IO>OBL>GEN>OCOMP. This supposes that the easiest noun phrase to relativize in a sentence is the subject, followed by the direct object, indirect object, prepositional noun, subordinate noun clause, and the object of a comparative sentence. Scholars later applied the NPAH to language acquisition research. Many previous studies indicate that, as learners acquire each type of RC, the difficulty order conforms to the NPAH. That is, the subject-extracted RC (SRC) is the easiest to be acquired (Doughty, 1991; Eckman et al., 1988; Gass, 1979; Izumi, 2003). However, the results predicted by the NPAH apply mainly to Indo-European languages, particularly English. In these languages, the RC follows the head noun that it modifies, as in example (1a) where the head noun 'person' occurs at the left of the RC. This structure is the opposite in Chinese, as in (1b) where the head noun *ren* 'person' occurs at the right.

(1) a. the person who bought the book
b. *mai shu de ren*
buy book DE person
'the person who bought the book'

The question of whether or not languages placing the head noun at the right (e.g. Japanese or Mandarin Chinese) also conforms to the NPAH prediction that has been the subject of inquiry for decades. There are still disputes over the results of that research. Tarallo and Myhill's (1983) cross-language research indicates that, for Japanese or Chinese, the object-extracted RC (ORC) is easier than the SRC. Hasegawa (2005) also supports this result for the Japanese. However, Sakamoto and Kubota (2000) studied learners whose native language was English, Chinese, or Indonesian and found that they all conformed to NPAH. Regarding L2 Chinese RC learning, there are mixed results either supporting NPAH (Li, 2015; Xu, 2014) or rejecting it (Dai, 2010; Tarallo & Myhill, 1983).

Many previous studies have been conducted with cognitive experiments, for example, by combining two sentences into a single sentence or judging the grammaticality of RCs. Those experiments were conducted within a controlled environment. The advantage of such experiments is that a feature effect can be pinpointed and focused, but only a very limited number of samples can be observed, thus perhaps causing incomplete and disputable results. In this research, we turn to naturally

produced various interlanguages instead of using a limited sample produced in a controlled way. In order to deal with such a large amount of influencing features, we adopted a corpus-based approach using the Test of Chinese as a Foreign Language (TOCFL) Learners' Corpus comprising 2,259 compositions written by learners of different proficiency levels and various L1s, including English, Japanese, Korean, Vietnamese, Indonesian, and Spanish, which provides naturally composed data for our analysis. The goal of our research is to uncover the patterns of RC acquisition of L2 Chinese through a corpus approach. These are the three research questions we ask:

1. What is the distribution between SRCs and ORCs produced by L2 learners?
2. Is there any different distribution among learners of different L1 backgrounds?
3. Is there any difference? If so, what is the difference?

In order to present the results and uncover influencing factors, we provide a comprehensive review of the related issues.

2 Related Research on Chinese RC Acquisition

2.1 Disputed Results on the Acquisition of SRCs and ORCs

As mentioned in the previous section, results that conform to the NPAH prediction apply mainly to Indo-European languages, the head nouns of which occur at the left of RCs. The question of whether left-branching languages (e.g. Japanese, Korean, and Mandarin Chinese) also conform to the NAPH prediction has led to increased research within the last decade. Sakamoto and Kubota (2000) investigated the RC acquisition of Japanese L2 learners with different L1s (i.e. English, Mandarin Chinese, and Indonesian) by using a sentence-combining task. The results conformed to the NAPH prediction. However, later studies on L2 learners of Japanese did not completely conform to the NAPH prediction (Ozeki & Shirai, 2007). Their results suggest that the NPAH does not predict the difficulty order of Japanese RCs. See Hasegawa (2005) for further reading. O'Grady et al. (2003) explored the second language acquisition of Korean. 53 native English speakers were asked to select a corresponding picture based on the type of RCs they heard. Participants showed that ORCs were more difficult to comprehend than SRCs. These results conform to the NAPH prediction. Tarallo and Myhill's (1983) cross-language research indicates that, for English native speakers learning languages where the RC occurs after a head noun (e.g. German, Portuguese, and Persian), the SRC proves easier than the ORC. This appears to conform to the NAPH prediction, but they also found that the reverse occurs when English native speakers are learning Japanese or Chinese. In such cases, the ORC is easier than the SRC.

Regarding RC acquisition for L2 Chinese learners, Packard's (2008) research utilizes a self-paced reading task to assess English speakers' processing difficulty

of L2 Chinese RCs. The results show that English speakers demonstrate slower processing times for SRCs. Packard suggests that Chinese instructors should teach ORCs before SRCs. Since this study targets English-speaking participants, this suggestion may only be applicable to English native speakers. A different approach used by Dai (2010) also supports that ORCs are easier than SRCs. He employed a sentence-combining task to investigate how the position and type of RC impact Chinese language acquisition. His 39 participants (intermediate to advanced level proficiency) came from various L1 backgrounds (i.e. English, Japanese, and Korean). That study concluded that the RC position has no significant effect on the acquisition of RCs. However, the type of RC has an obvious impact on learners' acquisition. The order of acquisition indicates that ORCs are the easiest, followed by SRCs, ID-RCs, OBL-RCs, etc. Therefore, the acquisition of L2 Chinese RCs does not support the NPAH hypothesis.

Despite those results, there are other studies which suggest that the NPAH acquisition theory is applicable to L2 Chinese. For example, Xu (2014) conducted a sentence-combining task for 45 native speakers of English in order to investigate if the order of difficulty conforms to the NPAH prediction. The results showed that the intermediate-high level learners preferred to produce SRCs than ORCs. In addition, she also claimed that SRCs were easier than ORCs through the analysis of learner's response accuracy. This shows that the NPAH is applicable to L2 Chinese. Li (2015) conducted a corpus-based study to analyze RC production by speakers of three L1s (i.e. English, Japanese, and Korean) in the HSK corpus (Zhang et al., 2004). The 201 RC sentences they observed show that all three groups of advanced level learners tended to produce more SRCs, and this therefore also supports the NPAH hypothesis.

Based on a review of the aforementioned studies, we find that, even with similar research methods, contradictory results were reported. This leads us to wonder whether the inconsistent results were caused by different L1s or different language proficiency levels. In Sect. 4, we will address this question further.

2.2 Effects of the Animacy of Head Nouns

Aside from the predictive power of the NPAH, analysis based on language processing has provided much insight into the study of RCs in recent years. For example, Traxler et al. (2002) used eye-tracking testing to conclude that an ORC following an animate head noun is more difficult to process, such as 'The mountaineer that the boulder hit', than an inanimate head noun, such as 'The rock which the boy threw.' This shows that the animacy of a head noun is connected to the difficulty of comprehension of an RC. The results of Ozeki and Shirai (2007) also support the effect of animacy. 1005 tokens of Japanese RCs by native speakers of English, Korean, and Chinese were collected from an oral interview corpus. They concluded that English-native and Chinese-native L2 Japanese learners made strong associations between Subject and animate heads and between Direct Object/Oblique and inanimate heads.

There is very limited research on the role of animacy in the L1 Chinese acquisition of RCs. The two most prominent studies on L1 Chinese are Cheng (1995) and Wu (2011). Cheng used elicitation tasks to examine Mandarin-speaking children's (across age groups of three-, four-, and five years old) production of RCs. Her research is based on the semantic hypothesis that an inanimate argument is easier to comprehend. She has shown that, if a head noun is inanimate, participants demonstrate a higher rate of accuracy and that the noun phrase proves easier to understand. And this tendency is more apparent in younger children. Wu (2011) analyzed 331 RCs in a news corpus. Her results show that SRCs contain more animate heads while ORCs contain more inanimate heads. She suggested that the effect of animacy found in the corpus may account for the inconsistent results of previous experimental studies.

Regarding L2 Chinese learning, by observing the HSK corpus, Li (2015) demonstrated that the animacy of nouns in RCs strongly affects the generation of RC types. He also declared that NPAH is secondary to animacy in affecting the production of RC types. However, his research did not take learners' proficiency into account and observed only a limited 201 samples. In our study, we also adopted a corpus approach, but we observed a total of 2,259 samples of RCs representing learners from various L1s and Chinese proficiency levels. Hopefully, this can provide a better profile to settle the dispute among the inconsistent results described above.

2.3 *Effects of Positions in a Matrix Sentence for SRCs and ORCs*

Some cognitive theories posit that center-embedded RCs may interrupt language processing; therefore, they are more difficult to comprehend than those (right- or left-embedded) which occur on the sides of the matrix sentence (Bever, 1970; Kuno, 1974). Mandarin Chinese is considered a left-branching language. An RC is also based on the left-branching structure to always occur before a head noun, which thus causes an embedded structure with an object position such as (2) and (3), but not with a subject position as shown in (4) and (5). In view of this, Chinese RC nominals in the subject position (either SS or SO) should be easier to process than object-position RC nominals (either OS or OO) as shown in the examples below. In addition, Sheldon (1974) also proposed that RCs with the same position and type are easier to comprehend than different structures; that is, SS and OO are easier to comprehend than SO and OS.

(2) *ta bu shi* [*na ge mai shu de ren*] $_{OS}$.
he not be that-Cl buy book DE person
'He is not [the one who bought that book].'

(3) *ta xihuan* [*Zhangsan mai de shu*] $_{OO}$.
he likes Zhangsan buy DE book
'He likes [the books which Zhangsan bought].'

(4) [*mai shu de nage ren*] SS *bu shi wo tongxue.*
buy book DE that-Cl person not be my classmate
'[The one who bought that book] is not my classmate.'

(5) [*Zhangsan mai de shu*] SO *bu jian le.*'
Zhangsan buy DE book not see ASP
'[The books Zhangsan bought] are lost.'

Dai (2010) aimed to understand how the position and type of RCs impact L2 Chinese acquisition. His study concluded that the position factor has no significant effect on producing SRCs or ORCs. Li's observations (2015) showed that the embedded structures of OS and OO are produced more often for English-native learners. Korean-native learners showed no preference, and Japanese-native learners showed the opposite tendency: SO and SS structures were produced more. However, there were no statistically significant differences among the three learner groups' RC production. It seems that current research shows that the position of RCs has no strong preference effect on the selection of RC types. What is curious is why learners do not avoid using more complicated embedding structures. We will clarify this with our statistical results in Sect. 4.

To summarize, most of the previous research on L2 Chinese RC acquisition has been based on experimental methods, either from the viewpoint of universal grammar or language processing. There is still controversy over the results of such research. Our motive is to discover the difficulty of SRCs and ORCs in order to apply the findings within pedagogical grammar. Therefore, we observed and analyzed written texts spontaneously produced by L2 Chinese learners to examine the distribution, position, and animacy effect of RCs.

3 Methodology

3.1 Research Scope

We would like to clarify some terminology and basic syntactic patterns of Chinese RCs before any further discussion. The basic formation of RC nominals in Mandarin Chinese is not different from common noun phrases, except that the modifier must be either a verb phrase or a clause. The modifier precedes a head noun and ends with the relative particle *de,* which connects with the head noun together to form an RC noun phrase. For instance, *na ben shu* 'that book' is the head noun of the RC nominal *wo xihuan de na ben shu* 'that book that I like' and *wo xihuan de* 'that I like' is the modifier. In this research, 'RC' may be used to refer to either the relative clause itself or sometimes the relative clause NP; a distinction between these two referred clauses will not be made if the ambiguity can be resolved by context or differentiation. A modifier of an RC must be understood as either a verb phrase or a clause, but within this intransitive stative verbs and adjectives are excluded. Therefore, NPs such as

congming de nühai 'a smart girl' or *hen hao de keben* 'a good textbook' are not in the category of RC. Furthermore, this study limits the scope of head nouns to only the subjects or objects of verbs of RCs, i.e. the top two roles within NPAH. For instance, even when modifiers are in a verb phrase, the following examples in which head nouns are not in the subject or object roles are excluded: (1) there is an appositional relationship between the modifier and the head noun, such as *women qu Ouzhou lüxing de jihua* 'our plan to travel to Europe'; this is because *jihua* 'plan' is not an argument of this clause; (2) the head noun is part of a clausal subject, as in *lihunlü hen gao de guojia* 'a country with a high divorce rate'; (3) any instance where the head noun is omitted, for example, *wo xihuan ni mai de (shu)* 'I like what you bought (the book)'. To put it simply, only the head noun of an NP is the subject or object argument of an active verb.

3.2 Research Method

Previous research on RC comprehension or generation for the most part has been based on individual experiments in cognitive psychology, such as online sentence generations, grouping linguistic elements together to form a grammatical RC (Wu & Sheng, 2014), or asking learners to combine two sentences into one with an RC construction (Xu, 2014). These methods use designed test questions to accomplish specific research objectives, and results may be used to test a research hypothesis. However, collected samples are often limited because the number of target subjects is constrained by budget and time. Other than experimental design, another solution is to analyze a much larger quantity of authentic language materials provided by a learner corpus. Learners' language use over different proficiency levels can also be regarded as longitudinal profiles. Hence, corpus-based or corpus-driven studies have provided a new avenue for research (Granger, 1998; Douglas, 2001; Ellis & Barkhuizen, 2005: 48; Myles, 2005).

In order to analyze a large quantity of authentic language used by learners, our research also adopts a corpus-based approach. The corpus we used is the TOCFL Learners Corpus (Chang, 2013).[1] This corpus consists of essays written by non-native Chinese-speaking participants who have taken the TOCFL from 2006 to 2012. It contains 1.6 million words from learners of 42 different language backgrounds, including 4,709 essays on 80 topics written by learners from different proficiency levels. The corpus differs from the HSK corpus used by Li (2015) in Mainland China. The TOCFL is an online test that allows participants to directly type their essays into a computer, and the data in the corpus comes from the beginning, intermediate, and high proficiency level learners (CEFR A2-C1). However, the HSK corpus collects the learners' hand-written compositions and only includes essays from advanced proficiency learners (CEFR B2).

[1] Please visit the website http://tocfl.itc.ntnu.edu.tw/(account: tocfl; pwd: demo123).

Therefore, in this research we have the advantage of using a much larger quantity of data from different native language backgrounds and varying proficiency levels in order to investigate if these groups demonstrate any clear differences when producing RCs in Chinese.

3.3 Corpus Data

Linguistic typologist Joseph Greenberg (1963) has noticed that Mandarin Chinese is different from the 30 other VO word order languages. In his analysis, RCs in the other VO languages are formed by placing the head noun to the left of the modifier (i.e. a right-branching structure); however, Chinese follows a VO word order where the head noun is placed on the right (i.e. a left-branching structure). This unique structure is distinct from other languages. Therefore, this study has observed learners whose native languages have different typologies as classified below (Chen, 2007:236). In order to ensure the generalizability of our analysis and to meet our research goals, this study selects two languages from each type, including Japanese and Korean (type 2), Indonesian and Vietnamese (type 3), and English and Spanish (type 4).

Type 1. Left-branching languages with VO word order: Chinese
Type 2. Left-branching languages with OV word order: Japanese and Korean
Type 3. Right-branching languages with VO word order: Thai, Vietnamese, and Indonesian
Type 4. Right-branching (head nouns + RCs) and left-branching (adjectival modifiers + head nouns) with VO word order: English, German, French, Spanish, and Italian.

The following example uses the NP *xuesheng mai de* (*na ben*) *shu* 'The book that the student bought' to exemplify the structure of RCs in each of the six languages. English, Indonesian, Vietnamese, and Spanish all place the head noun on the left, while Japanese, Korean, and Mandarin Chinese place the head noun on the right.

Chinese:	*xuesheng mai de (na ben) shu*	(the head on the right)
	student bought DE (that CL) book	
Japanese:	*gakusei-ga ___ katta hon*	(the head on the right)
	student-NOM bought book	
Korean:	학생이 산 책	(the head on the right)
	student-NOM bought book	
Indonesian:	buku yang siswa beli	(the head on the left)
	book which student buy	
Vietnamese:	cuốn sách học sinh mua	(the head on the left)
	CL book student buy	
English:	the book which the student bought	(the head on the left)

Table 1 Number and distribution of observed compositions for six L1s

	Japanese	English	Korean	Vietnamese	Indonesian	Spanish	Total
B1	530	344	245	152	163	90	1,524
B2	260	122	130	96	112	15	735
Total	790	466	375	248	275	105	2,259

Spanish: El libro que el estudiante compra (the head on the left)
the book which/that the student bought

In order to ascertain whether language proficiency affects learners' RC expressions, this study investigates two proficiency levels in the TOCFL corpus: B1 (CEFR B1 corresponds to the Intermediate-high level in the ACTFL scale) and B2 (advanced level in the ACTFL scale). The data from the B2 levels can be compared with results from previous studies using the HSK corpus of advanced learners (Li, 2015). Table 1 shows that the total number of observed compositions is 2,259 (1,524 for the B1 level and 735 for the B2 level). We can see that the corpus does not provide a balanced distribution of native speakers from each language background in Table 1; this is because there is not a balanced distribution among test participants in the first place. This is especially true of Spanish-speaking B2 learners who account for only 15 compositions. Therefore, this study provides a quantitative analysis of Spanish speakers as a reference rather than an observation of statistical significance.

3.4 RC Markup Principles for the Corpus

Once selected, corpus materials must be manually reviewed to mark the information of each RC. If an RC is applicable to this investigation, it is copied into a separate Excel spreadsheet. Each RC is then tagged with three pieces of information for analysis: (a) type of RC (ORC 'O' or SRC 'S'), (b) position of the RC nominal in the matrix sentence (subject or object position), and c) animacy of the RC head noun (animate '+' or inanimate '−'), as shown in Table 2.

In marking the RCs in the corpus, since the authentic materials are from language learners' interlanguages, more detailed criteria must be defined to judge partially incorrect samples, as shown in (6)–(12). Despite typos (*xi* is omitted in (6)) or incorrect verb usage in (7), the structure of the RC is still apparent in sentences (6) and (7). Since these errors do not jeopardize the judgment of the RCs, they are still marked as RCs in the statistical analysis. The errors for sentences (8)–(10) are respectively caused by lacking the auxiliary verb (*yao)* and the wrong word order position of the adverb (*zui*) as well as the determiner (*you xie*). Since these errors do not affect comprehension, they also count as RCs. However, sentences (11) and (12) lack the main verb of the RC. Though these sentences may still be comprehensible within

Table 2 Markups for sample RCs

Entry no.	RC nominal	Type	Position	Animacy
1	*mei tian yüdao de shiqing* every day encounter DE event 'the things encountered every day'	O	S	–
2	*ding cai de pengyou* order food DE friend 'the friend who ordered food'	S	S	+
3	*bu tai xiguan shuo Yingwen de xuesheng* not very used to speak English DE student 'the students who are not very used to speaking English'	S	O	+
4	*wo ti ni anpai de liang ge xuanze* I for you make DE two CL choice 'the two choices I made for you'	O	O	–

context, they lack a very important element—the verb. To avoid controversy, samples like number (11) and (12) have been excluded.

(6) *tamen xuyao de dong […]* (should be *tamen xuyao de* ***dongxi*** 'those things which they need')

(7) *tamen xiang yao kai de shengyi* (should be *ta men xiang yao* ***zuo*** *de shengyi* 'the business they would like to do')

(8) *wo zuotian xie de ren* (should be *wo zuotian* ***yao*** *xie de ren* 'those who I thanked yesterday')

(9) *wo bixu zui ganxie de ren* (should be *wo* ***zui bixu*** *ganxie de ren* 'those who I must thank the most.')

(10) *chi de you xie dongxi* (should be ***you xie chi de*** *dongxi* 'There are some things to eat.')

(11) *hen duo cong bu yi yang de guojia de xuesheng* (should be *hen duo cong bu yi yang de guojia* ***lai*** *de xuesheng* 'students from many different countries')

(12) *hen duo jianshu de ren* (should be *hen duo* ***hui*** *jianshu de ren* 'many people who know how to fence').

4 Statistical Results and Discussion

In this study, we observed a total of 2,055 RCs, including 1,362 RCs at the B1 level and 693 RCs at the B2 level. This can be compared with Li's (2015) corpus data, in which he investigated only 201 total RCs among English, Korean, and Japanese native speakers, a sample size significantly smaller than ours. Table 3 shows the total number of RCs produced by each of the six learner groups. Japanese learners produced the largest sample size of 563 RCs, but the largest number of occurrences does not indicate the most frequent use due to uneven distribution of compositions across different language groups. The Japanese essays account for the largest portion of the TOCFL corpus, totaling 187 thousand characters from B1 learners and 128

thousand characters from B2 learners. In fact, the highest frequency of RCs is found among native Korean B1 learners as shown in Table 3.

4.1 Difficulty of ORCs or SRCs

The majority of second language acquisition studies support the NPAH accessibility hypothesis that subject-RCs are easier to acquire than object-RCs. In this study, however, we analyzed more than 2,000 RCs and found that more ORCs were produced than SRCs with a statistical significance value of $p < 0.001$. Tables 4 and 5 provide detailed statistics where the number in parentheses indicates the number of tokens. Table 4 shows that, regardless of the mixed language background, all B1 learners consistently produced more ORCs than SRCs with a statistical significance value of $p < 0.01$. However, at the B2 level (see Table 5), there are some variations of the production advantage across different language backgrounds. English- and Korean-speaking B2 learners produced significantly more SRCs than ORCs with a statistical significance value of $p < 0.05$, averaging about 60%. Japanese-speaking B2 learners did not show a significant difference between the uses of SRCs and ORCs with $X^2 = 0.961$ and $p = 0.327$. As for B2 Indonesian-, Vietnamese-, and Spanish-speaking learners, the ORC still maintained an advantage with a statistical significance value of $p < 0.05$; however, Spanish speakers were excluded due to the small sample size. We also observe that there is an overall increase in the use of SRCs as the learners' language proficiency increases.

This is contrary to the corpus-based findings of Li (2015). While that study showed an advantage of SRCs among English, Japanese, and Korean native speakers, it showed no significant difference in the generation of the two types of RCs. While our data from the TOCFL B2 corpus is similar in quality and proficiency level to that of the HSK, our analysis shows similar results among English-speaking learners but opposite results among Japanese- and Korean-speaking learners. In addition, our B2 Vietnamese- and Indonesian-speaking learners produced an average of 60% more ORCs. The result is in contrast to native English learners' preference, despite the fact that these three languages all have head initial NP structures. Therefore, our investigation of the RCs produced by learners of different native languages and proficiency levels does not support the argument that SRCs are easier than ORCs in Mandarin Chinese. On the contrary, low proficiency level learners consistently produced more ORCs. Coincidentally, the same result is found in Chinese L1 acquisition research, which has indicated that the younger the child, the more likely they are to produce an ORC (Chen & Shirai, 2014; Cheng, 1995; Lee, 1992).

As a result, we hypothesize that ORCs are easier to learn in Chinese because more ORCs are produced for lower proficiency learners regardless of their native language types. The dominant factor for this result may be that the word order of RC nominals is the same as that of Chinese simple sentences. After reaching higher language proficiency, no matter what the language background is, learners gradually achieve more native-like expressions. Past Chinese L1 corpus-based studies all show

Table 3 Statistics on RCs

	Japanese		Korean		Vietnamese		Indonesian		English		Spanish		Total RCs
	Char	RC	Char	RC	Char	RC	Char	RC	Char	RC	Char	RC	
B1	187,650	350	92,650	299	57,879	170	60,660	236	131,443	239	33,312	68	1362
B2	128,697	213	67,795	146	54,876	97	34,093	108	66,902	107	6545	22	693

Table 4 The distribution of subject-RCs and object-RCs produced by B1 learners

B1	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Average
Subject-RC	24% (83)	21% (63)	24% (41)	22% (52)	23% (55)	31% (21)	23% (315)
Object-RC	76% (267)	79% (236)	76% (129)	78% (183)	77% (184)	69% (47)	77% (1,046)

Table 5 The distribution of subject-RCs and object-RCs produced by B2 learners

B2	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Average
Subject-RC	47% (95)	59% (83)	35% (33)	25% (25)	60% (62)	5% (1)	45% (299)
Object-RC	53% (109)	41% (58)	65% (62)	75% (77)	40% (41)	95% (21)	55% (368)

that the tokens of SRCs occur more often than ORCs (Hsian & Gibson, 2003; Pu, 2007; Tang, 2007), regardless of the genres of the corpus data. This might explain why B2 learners show an increased use of SRCs.

4.2 Effects of Positions of RCs in a Matrix Sentence

Our review of various theories in cognitive linguistics (e.g. Bever, 1970; Kuno, 1974) posits that the structure of the center-embedded RC may reduce processing speeds and make it more difficult to comprehend than an RC placed on the sides of the matrix sentence. This supposes that for languages with NP head final structure like Chinese, SS or SO structures should be easier to process than OO or OS structures, since OO and OS cause embedded structures while SS and SO do not. Does this hypothesis imply that RCs should occur more in a subject position than in an object position? Tables 6 and 7 show that the embedded structures of OO and OS are actually produced more than the non-embedded structures of SS and SO with a statistical significance value of $p < 0.001$. This goes against some theories in cognitive linguistics (e.g. Bever, 1970; Kuno, 1974; Sheldon, 1974). Our data shows that ease of comprehension seems not to equate to ease of production in language processing.

Li (2015) found that English-speaking Chinese L2 learners generated more embedded structures, Japanese-speaking L2 learners generated fewer embedded structures, and Korean-speaking L2 learners use both positions equally. Therefore, he claims that the position of RCs seems not to affect the preference for RC-type generation. However, he also stated that the result did not reach statistical significance due to the limited observation samples. He studied only 201 RC samples taken from learners who are equivalent to the B2 level of the TOCFL scale. On the contrary, we studied 440 samples from these three languages (i.e. Japanese, Korean, and English) and more from other languages (see Table 7). The results show a preference for the

Table 6 RC position and type distribution of B1 learners

B1	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Total
SS	10% (35)	7.8% (23)	4.7% (8)	7.7% (17)	4.2% (10)	16.2% (11)	44% (601)
SO	38% (133)	39.8% (117)	38.2% (65)	32.8% (77)	34.7% (83)	32.4% (22)	
OS	13.7% (48)	12.6% (37)	19.4% (33)	14.9% (35)	18.9% (45)	14.7% (10)	56% (755)
OO	38.3% (134)	39.8% (117)	37.6% (64)	45.1% (106)	42.2% (101)	36.8% (25)	
Sum	350	294	170	235	239	68	1,356

Table 7 RC position and type distribution of B2 learners

B2	Japanese	Korean	Vietnamese	Indonesian	English	Spanish	Total
SS	16.7% (34)	33.8% (45)	13.7% (13)	17% (17)	16.5% (17)	0% (0)	41% (270)
SO	17.6% (36)	23.3% (31)	23.1% (22)	36% (36)	12.6% (13)	27.3% (6)	
OS	29.9% (61)	24.1% (32)	21.1% (20)	8% (8)	43.7% (45)	4.5% (1)	59% (387)
OO	35.8% (73)	18.8% (25)	42.1% (40)	39% (39)	27.2% (28)	68.2% (15)	
Tokens	204	133	95	100	103	22	657

generation of embedded structures (OO, OS) over non-embedded structures (SS, SO).

So, what is the factor that causes more production of the more difficult embedded structures? Li (2015) provided the following explanations. Since L2 learners lack sufficient language proficiency, they are more inclined to generate simple RCs (p.37) and their processing of short RC NPs resembles that of idiom chunks, which does not cause difficulties for sentence generation. However, while such an explanation might satisfy his claim that 'the position of RCs seems not to affect the preference of RC type generation', it cannot explain why ORCs were generated more regardless of their position. Here, we reassert our previous hypothesis that word order is the dominant factor. Since the word order of ORCs is the same as the word order of basic Chinese sentences, i.e. SVO, it results in learners preferring to generate ORCs because it does not require additional processing effort. Tables 6 and 7 show that ORCs are generated significantly more than SRCs in both the subject position and the object position (i.e. SO and OO being generated more than OS and SS, respectively), no matter whether there is structure embedding or not.

Overall, the patterns of OO and SO have the advantage. The average distribution of OO is 39% and SO is 30%. That means the occurrence of the ORC type is not affected by the position in the matrix sentence. No matter where the ORCs are positioned,

their tokens occur more than SRCs. However, for the B2-level Korean- and English-speaking learners, SRC has an obvious advantage. OS is 43.7% for Korean speakers while SS is 33.8% for English speakers, as shown in Table 7. At present, we do not have a good explanation for this part of the data. The only possible assumption is that the B2 English speakers and Korean speakers produce more SRCs than ORCs as indicated in Table 5; therefore, they show this special tendency.

4.3 Animacy of Head Nouns

Previous research shows that 1) the animacy of a head noun is related to the comprehension of an RC and 2) SRCs tend to modify animate noun phrases while ORCs tend to modify inanimate noun phrases. The animacy effect may affect the distribution of the types of RCs. We can see this tendency clearly in Table 8 where the data shows that SRCs prefer animate heads (overall average is 67%), whereas the ORCs prefer inanimate heads (overall average is 85%). It is also consistent with research done by Ozeki and Shirai (2007). Our data also shows that the association between ORCs and inanimate heads is stronger than between SRCs and animate heads. Such a phenomenon may explain why the lower proficiency language learners produce more ORC types than SRC types. In addition to the factor of word order, the processing of inanimate nouns is easier than that of animate nouns (Cheng, 1995).

Furthermore, the animacy effect on SRCs becomes stronger as language proficiency moves from B1 to B2, regardless of the learners' native language. The overall average of SRCs with animate nouns for the B1 level and the B2 level is 58% and 77%, respectively, with a statistical significance value of $p < 0.005$. Though the overall average of ORCs with inanimate nouns from B1 (83%) to B2 (92%) shows the same tendency, there is no significance ($p = 0.0934$).

For another statistical perspective, Table 9 exhibits the analysis of the relationship between SRCs/ORCs and the animacy of the head noun. In the B1 level, English-native Chinese learners used a total of 85 animate head nouns, with 53 modified by ORCs and only 32 modified by SRCs. The tendency of Indonesian-native learners is the same as English learners, i.e. ORCs have the advantage. Korean-native learners used both structures (object- and subject-RCs) equally and do not exhibit a clear preference for animate head nouns. Japanese-native learners used 103 animate head nouns, with 54 occurring with SRCs. While this number is slightly higher than ORCs, there is only a difference of five RCs. The distribution of Vietnamese- and Spanish-speaking L2 learners is higher among SRCs. It is true that totally animate nouns occur more often with SRCs, but the different distribution is not statistically significant ($p = 0.8357$).

However, for the B2 level, there was an overall average of 89% of animate nouns occurring in the SRCs. Based on the discrepancy between the B1 and B2 levels, we may conclude that once high language proficiency has been reached, the effect of animacy takes over as the dominant factor in RC type generation. For lower proficiency learners, the factor of word order, instead of animacy, is dominant in

Table 8 The distribution of animacy of head nouns modified by RCs

Type	Animacy	Japanese		Korean		Vietnamese		Indonesian		English		Spanish	
SRC	Animate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		58% (32)	84% (52)	65% (54)	81% (77)	53% (33)	71% (59)	51% (31)	71% (20)	59% (24)	70% (23)	67% (14)	100% (1)
	Inanimate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		42% (23)	16% (10)	35% (29)	19% (18)	47% (29)	29% (24)	49% (30)	29% (8)	41% (17)	30% (10)	33% (7)	0% (0)
ORC	Animate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		29% (53)	2% (1)	18% (49)	9% (10)	14% (34)	7% (4)	18% (36)	9% (7)	5% (7)	6% (4)	10% (5)	14% (3)
	Inanimate	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2	B1	B2
		71% (131)	98% (40)	82% (218)	91% (99)	86% (202)	93% (54)	82% (164)	91% (71)	95% (122)	94% (58)	90% (43)	86% (18)

Table 9 The distribution of animate and inanimate head nouns of RCs with different types

	B1 animate nouns		B2 animate nouns	
	SRC	ORC	SRC	ORC
English	38% (32)	62% (53)	98% (52)	2% (1)
Japanese	52% (54)	48% (49)	88% (77)	12% (10)
Korean	50% (33)	50% (34)	94% (59)	6% (4)
Indonesian	46% (31)	54% (36)	74% (20)	26% (7)
Vietnamese	77% (24)	23% (7)	85% (23)	15% (4)
Spanish	74% (14)	26% (5)	25% (1)	75% (3)
Average	51% (188)	49% (184)	89% (232)	11% (29)

producing the type of RCs. Findings yielded in Tables 8 and 9 thus lead to the conclusions that (1) there is a strong association between ORCs and inanimate heads; and (2) the lack of significant association between SRCs and animate nouns at the B1 level indicates that ORC type is easier than SRC type for lower proficiency Chinese learners.

5 Conclusion and Limitations of This Study

This study analyzed the L2 Chinese learners' production of RCs among six different native languages classified into three language typologies and discussed the corpus results from multiple perspectives, including (1) the types of RCs, (2) the position of an RC in the matrix clause, (3) animacy of the RC modifying head noun, (4) L2 learners' native languages, and (5) different proficiency levels of learners. Specifically, we examined how those mingling features affect the production of different types of RCs.

We have found that B1 (intermediate-high level) learners produce significantly more ORCs than B2 (advanced level) learners. This trend consistently occurs among all language backgrounds, thus indicating that lower proficiency learners produce more ORCs than SRCs. This phenomenon also occurs in L1 Mandarin Chinese acquisition. Owing to this effect (i.e. more ORCs), the learners produced more OO structure RCs, which is in opposition to findings of previous research on language processing theories, which imply that object position RCs should be produced less due to their embedded structure. Furthermore, the animacy effect for SRCs at the B1 level apparently does not occur either. This leads us to conclude the following hypothesis: at the early stage of L2 Chinese learning, word order is the dominant factor for language processing no matter what the learner's native language is. Since SVO is the conventional word order in Mandarin Chinese and noun phrases modified by ORC have the same SVO structure, this results in the fact that ORCs are easier than SRCs. This also explains why there is no obvious effect on the position feature

causing embedded structure and preference for animate head nouns in SRCs among B1 learners. As learners' proficiency gradually increases, learners produce more SRCs and their interlanguage will approximate the target language.

While a corpus-based study may reflect a learner's natural language production, there are also some limits to this research. For example, the data for Spanish-native learners in this study is just for readers' reference since the corpus does not include a sufficient sample size to produce reliable statistics. Nevertheless, through the use of the spontaneously produced data of learners of different native languages and proficiency levels, the language development of L2 learners can be clearly observed and analyzed. These corpus-based results may be combined with the results from psychological or cognitive linguistic experiments to represent interlanguage development from more comprehensive perspectives.

Acknowledgements The author would like to thank the anonymous reviewers for their valuable comments and the support of the Ministry of Science and Technology, under the grant MOST 108-2410-H-002-117.

References

Bever, T. G. (1970). The cognitive basis for linguistic structures. In: Hayes, J. R. (Ed.), *Cognition and the development of language* (pp. 279–352). New York: Wiley.

Chang, L.-P. (2013). The construction and implication of the TOCFL corpus. In: B. Zhang & X. Cui (Eds.), *The proceedings of the 2nd International Conference of Interlanguage Corpora* (pp. 141–152). Beijing: Beijing Language and Culture University Press. [張莉萍, 2013, 〈TOCFL作文語料庫的建置與應用〉, 崔希亮、張寶林 主編,《第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集》141–152。北京: 北京語言大學出版社。].

Chen, J., & Shirai, Y. (2014). The acquisition of relative clauses in spontaneous child speech in Mandarin Chinese. *Journal of Child Language*. CJO 2014 https://doi.org/10.1017/S0305000914000051.

Chen, F. J. (2007). *Contrastive analysis and its applications in language pedagogy*. Taipei: Crane Publishing. [陳俊光, 2007,《對比分析與教學應用》。台北: 文鶴出版社。].

Cheng, Y.-Y. (1995). The acquisition of relative clauses in Chinese. MA thesis, National Taiwan Normal University, Taipei.

Dai, Y. (2010). An investigation of the relative clause acquisition by learners of Chinese as a second language. *Journal of Ocean University of China (Social Sciences Edition), 6*, 85–91. [戴運財, 2010, 〈漢語作為第二語言的關係從句習得難度調查〉。《中國海洋大學學報》(社會科學版) 6:85–91。].

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in Second Language Acquisition, 13*, 431–469.

Douglas, D. (2001). Performance consistency in second language acquisition and language testing research: A conceptual gap. *Second Language Research, 17*(4), 442–456.

Eckman, F. R., Bell, L. H., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics, 9*, 1–20.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Granger, S. (Ed.). (1998). *Learner English on Computer*. New York: Longman.

Gass, S. (1979). Language transfer and universal grammatical relations. *Language Learning, 29*, 327–344.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 73–113). Cambridge, Mass: MIT Press.

Hasegawa, T. (2005). Relative clause production by JSL children. In: M. Minami, H. Kobayashi, M. Nakayama, & H. Sirai (Eds.), *Studies in language sciences 4: Papers from the fourth annual conference of the Japanese society for language sciences* (pp. 189–204). Tokyo: Kurosio.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

Izumi, S. (2003). Processing difficulty in comprehension and production of relative clauses by learners of English as a second language. *Language Learning, 53*, 285–323.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*(1), 63–99.

Kuno, S. (1974). The position of relative clauses and conjunctions. *Linguistic Inquiry, 5*, 117–136.

Lee, T. H.-T. (1992). The inadequacy of processing heuristics–evidence from relative clause acquisition in Mandarin clause. In: Lee, T. (Ed.) *Research on Chinese linguistics in Hong Kong* (pp. 47–85). Hong Kong: The Linguistic Society of Hong Kong.

Li, J. (2015). Research on Chinese relative clause generation: A language type perspective. *Modern Foreign Language Research, 2*, 34–39. [李金滿, 2015, 〈漢語二語關係從句產出研究–類型學視角〉。《當代外語研究》2:34–39。].

Myles, F. (2005). International corpora and second language acquisition research. *Second Language Research, 21*(4), 373–391.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition, 25*(3), 433–448.

Ozeki, H., & Shirai, Y. (2007). Does the noun phrase accessibility hierarchy predict the difficulty order in the acquisition of Japanese relative clauses? *Studies in Second Language Acquisition, 29*(2), 169–196.

Packard, J. L. (2008). Relative clause processing in L2 speakers of Mandarin and English. *Journal of the Chinese Language Teachers Association, 43*(2), 107–146.

Pu, M.-M. (2007). The distribution of relative clauses in Chinese discourse. *Discourse Processes, 43*(1), 25–53.

Sakamoto, T., & Kubota, S. (2000). Nihongo no kankeisetu no syuutoku ni tuite [On the acquisition of Japanese relative clauses]. *Nanzan-Daigaku Kyoiku Sentaa Kiyoo [The Bulletin of the Center for International Education, Nanzan University], 1*, 114–126.

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior, 13*(3), 272–281.

Tang, Z. (2007). Position of relative clause: An analysis based on corpus and genres. *Modern Lingustics, 2*, 139–150. [唐正大, 2007, 關係化對象與關係從句的位置—基於真實語料和類型分析。《當代語言學》2:139–150。].

Tarallo, F., & Myhill, J. (1983). Interference and natural language in second language acquisition. *Language Learning, 33*, 55–76.

Traxler, M., Morris, R., & Seely, R. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*, 69–90.

Wu, F. (2011). Experience effect or memory effect?-Evidence from new corpus for Animacy effect of relative clause generation. *Language Sciences*, *53*, 396–408. [吳芙芸, 2011, 〈基於經驗還是基於工作記憶?–來自漢語新聞語料庫中關係從句生命度格局的證據〉。《語言科學》53:396–408。].

Wu, F., & Sheng, Y. (2014). Pre-RC determiner phrase bias and production preference for object relatives: Perspectives from L2-Chinese learners. *Foreign Language Teaching and Research*, *3*, 14–24. [吳芙芸、盛亞南, 2014,〈指量詞的前置優勢及賓語關係從句的產出優勢:漢語二語學習者視角〉。《外語教學與研究》(外國語文雙月刊) 3:14–24。].

Xu, Y. (2014). Evidence of the accessibility hierarchy in relative clauses in Chinese as a second language. *Language & Linguistics, 15*(3), 435–464.

Zhang, B., Cui, X., & Ren, J. (2004). The construction concept of HSK dynamic composition corpus. In: *The Proceedings of Symposium of the Third National Conference on Language Application*

(pp. 544–554). [張寶林、崔希亮、任傑, 2004, ⟨關於“HSK動態作文語料庫”的建設構想⟩,《第三屆全國語言文字應用學術研討會論文集》544–554。].

The Study of Error Types of Chinese Learners' Written Texts: A Chinese Written Corpus-Based Study

Jia-Fei Hong, Hsin-Tzu Jen, and Yao-Ting Sung

Abstract The present study aims to tackle the issue of error in written texts by Chinese learners from a macro perspective. Although previous research has demonstrated the significance of positive feedback and effective correction in the realm of Second Language Acquisition (SLA) (Fathman and Whalley, 1990; Ashwell, 2000; Ferris and Robers, 2001; Chandler, 2003), little consensus has been reached regarding its practical implementation in pedagogy. In particular, writing holds a crucial role among the four basic language skills for its complex construction and meaning in written language. However, with the rise of corpus linguistics, new approaches and perspectives have been added to the study of Chinese as a Second Language (CSL) writing (Chang et al., 2015; Hong et al., 2018). In hopes of improving the teaching of writing in an integrated way, this study adopts methodologies from SLA and corpus linguistics to broaden the scale of interdisciplinary research. Through the lens of error analysis, this study examines data from learners with diverse backgrounds in Chinese Written Corpus and analyzes learners' error types with reference to the categorization proposed by Dulay et al. (1982). The results of this analysis identify possible contributing factors of various types of errors, such as native language and level, which can then be further analyzed and may account for learners' error patterns. The present study's findings yield significant insights in outlining the distribution

This work was supported by National Taiwan Normal University's Chinese Language and Technology Center. The center is funded by Taiwan's Ministry of Education (MOE), as part of the Featured Areas Research Center Program, under the Higher Education Sprout Project.

J.-F. Hong (✉)
Department of Chinese As a Second Language, National Taiwan Normal University, 162, Section 1, Heping East Road, Taipei City 106, Taiwan
e-mail: jiafeihong@ntnu.edu.tw

H.-T. Jen
Department of East Asian Languages and Literatures, University of Hawaii at Manoa, 2500 Campus Road, Honolulu, HI 96822, USA

Y.-T. Sung
Department of Educational Psychology and Counseling, National Taiwan Normal University, 162, Section 1, Heping E. Rd., Taipei City 106, Taiwan

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_8

of errors in CSL writing and provide teachers and future researchers with practical advice on the study of teaching strategy, instructional setting, and teaching sequence.

Keywords Error analysis · Chinese Written Corpus · Corpus linguistics · CSL

1 Introduction

With the emerging number of Chinese learners worldwide, Chinese has become a dominant language in the twenty-first century and is gradually becoming one of the most popular languages besides English. According to data from the Department of Statistics at the Ministry of Education in R.O.C, the number of international students entering Taiwan to learn Chinese is growing exponentially, which rose from 8,182 to 18,645 between 2005 and 2015.[1]

In an attempt to help language learners develop well-rounded language competence, learners tend to be exposed to exercises that are focused on four fundamental language skills: listening, speaking, reading, and writing. During the process of learning a second language, learners tend to have difficulty with speaking and writing skills. Specifically, due to the nuanced meanings and the rather complex sentential structures of written language, writing is considered to be more difficult than speaking for second language learners. Students often need to put more effort into the process of writing, and teachers are also required to invest more time in providing feedback. The tendency of Chinese learners' error types described in this study, which is drawn from comprehensive and objective data, is provided to the current teachers and learners of a second language. To learners, the key to using a language fluently and communicating well is to understand grammar and develop language competence (Nassaji & Fotos, 2011). When learning a new language, obstacles in the acquisition of grammar often produce ungrammatical sentences. Theories of Second Language Acquisition (SLA) identify the benefits of positive feedback in helping learners to develop second language competence (Fathman & Whalley, 1990; Ashwell, 2000; Ferris and Robers, 2001; Chandler, 2003). Through both theoretical studies and practical settings, it has been discovered that learners tend to struggle more with speaking and writing than with listening and reading. Enlightened by further exploration, writing actually plays a more intractable role than speaking. Writing skills require learners to master sentential structures that are more complex, as well as be proficient in the nuance of meaning in the written text. Therefore, students must invest more time in learning. Furthermore, the teacher also needs to put more effort into correcting vocabulary and grammar. Due to the difficulties of learning a second

[1] According to statistic data in the report "number of university international students in degree programs and language programs". The report is excerpted from "important statistic data in education" that published on the website by the Department of Statistics in the Ministry of Education in R.O.C. http://depart.moe.edu.tw/ED4500/cp.aspx?n=002F646AFF7F5492&s=1EA96E4785E6838F#.

language, it is relatively hard for foreign learners to have a noticeable improvement in their writing performance (Buckingham & Pech, 1976).

In the field of Chinese as a Second Language (CSL), there are unsolved problems between theory and practice. Due to a lack of research that analyses the application and teaching strategies of CSL teaching, while accounting for learners' backgrounds and levels, theoretical perspectives often fail to address the actual challenges of learning a second language. Additionally, the existing research that studies errors of Chinese learners tends to solely concentrate on learners speaking a particular native language, learners at a particular level, or learners using a particular linguistic form. Although the outcome of these studies can indeed provide insight into the phenomenon of particular learners, a comprehensive view of learners' error types remains unseen. Considering the diverse backgrounds of CSL learners, distinct patterns of errors may emerge from individual native languages. Also, learners at different levels tend to have varying kinds of errors and learning difficulties. The current solution for students with different backgrounds is to assign them to different learning tracks, such as regular class, intensive class, theme-based class, and so forth, according to their native language or level. The drawback of this system is that the placement is solely based on the student's class level, and no attention is paid to the influence of the learner's native language. Even though the same course material, class arrangement, and teaching procedure can be provided, the influence of a student's native language may still influence the kinds of errors that are made and the different language levels.

In an attempt to address the aforementioned gaps in research, this study will take a top-down perspective to investigate students' learning and discuss the distribution of errors from learners of distinctive backgrounds in terms of native language and level. Furthermore, different error types will be analyzed to understand the pattern of grammatical errors in hopes of facilitating the instructional design and teaching strategies.

The Chinese writing corpus used in this study includes 43 written texts from learners of diverse backgrounds and levels and is built according to the framework of the ACTFL writing proficiency test (ACTFL, 2012). With help from the corpus, this study retrieves specific data based on the different "native languages" and "levels" of learners; it is then able to determine if the error types correlate with the grammatical attributes of a learner's native language via their authentic written text. The result of the current study suggests that understanding errors from learners of different levels not only offers implications for the instructional and material design of CSL (Hong and Sung, 2017), but could also improve a learner's overall performance and help them to express their thoughts in writing more effectively (Hong et al., 2018). Notwithstanding the achievement of the Auto-correct Chinese Written Text System, which has 65% accuracy in Auto-detecting Grammar System (Chang et al., 2015) and 88% accuracy in Auto-correcting Written Text Grading (Hong et al., 2014a, 2014b), the information that lies in the pattern of grammatical errors is a critical factor for further breakthrough accuracy.

Considering theories in SLA, corpus linguistics, the application of natural language processing (NLP), and perspectives from second language learners, this

study discusses how to incorporate the findings from common grammar mistakes and error types by Chinese learners in the field of CSL. Moreover, in light of interdisciplinary design, this study seeks to identify applications for the result of this study and further development. In order to achieve the goal of nationality-based differentiated instruction both accurately and effectively in a comprehensive, systematic, and objective manner, this study examines the error types of CSL learners in the written text through research methods in corpus linguistics using "Chinese Written Corpus." Meanwhile, this study also categorizes the error types from learners of different backgrounds and levels and constructs a framework of error patterns through cross-checking. When teaching a second language, teaching materials, methodologies, and teaching strategies should all be differentiated according to an individual student's native language and level. Hence, the corresponding differentiation is an inevitable question in this study. If the data of grammatical errors can be described and analyzed in a comprehensive and objective way based on learners' native languages, levels, and the linguistic forms they use, it would offer CSL teachers, learners, and textbook writers effective strategies for language learning and teaching. Thus, the present study aims to construct a framework of error patterns that is relevant to teaching Chinese writing and to accurately identify the mistakes in a written text by cross-checking grammatical errors in the corpus. These error patterns can thereby provide CSL teachers with advice on how to design teaching materials and give feedback to Chinese learners for self-learning, as well as provide strategies for teaching Chinese learners that speak different native languages and are at different levels. With the aid of this framework, the effectiveness and efficiency of learning and teaching Chinese writing would significantly improve.

2 Literature Review

This section will review the existing literature relating to second language acquisition, error distribution of CSL learners, CSL pedagogical grammar, and corpus-based methodologies.

2.1 Sla

Second language acquisition, psychology, cognitive psychology, and education are all closely related. Different approaches and theories have proposed different perspectives to account for the factors that influence language acquisition and the application of effective pedagogy. The following section includes discussions that are related to theories of language acquisition, types of errors, and the causes of errors.

2.1.1 Theories of Language Acquisition

Since 1990, cognitivism has gradually become the dominant theory in the field of language acquisition. In Universal Grammar (Chomsky, 1995), it is stated that the human brain is equipped with a device that enables humans to acquire grammar and language. This device adopts a universal principle that formulates certain language structures, which embodies diverse forms and causes the distinction between languages. Studies in cognitive linguistics also emphasize the psychological process of learning and processing information. The emergence of Universal Grammar and cognitive theories consequently put error analysis in a crucial position in the study of language acquisition and teaching.

In the article "The significance of learner errors" (Corder, 1967), Corder suggests that teachers should pay close attention to the errors that students are unaware of. Likewise, the concept of interlanguage, which was proposed by Selinker (1972), emphasized that the transition from a learner's native language to a target language is systematic and analyzable. The value of the study of interlanguage lies in the prediction of possible errors by students and the prevention of learners' fossilization. Thereafter, studies on linguistic errors have gradually received recognition and have led to an increase in methodologies, such as error analysis, contrastive analysis, and so forth. These methodologies are all dedicated to the investigation of systems and types of errors by students at different levels and aim to develop particular strategies to facilitate the teaching of a second language. Many recent studies have also discovered that there is considerable disparity in possible difficulties and error types between beginners, intermediate learners, and advanced learners.

In cognitive structure migration theory, Ausubel (1968) indicated that the existing learning experience contributes significantly to the ongoing process of learning. He stated that the existing learning experience and the ongoing learning process would interact with each other and ultimately form a new cognitive structure. A similar phenomenon can be seen in the acquisition of language. Several types of transfers between languages can be categorized as interlanguage transfer and intra-language transfer based on their source, and positive transfer and negative transfer based on their influence on the learning process. The errors that learners make when learning a new language may be a negative transfer derived from the grammatical rules of their native language. Thus, in the field of CSL, the study of a learner's native language and its influence on a second language holds a central place among various research topics. Many studies have collected, analyzed, and categorized the errors from learners speaking different native languages and have proposed corresponding teaching strategies.

From the studies above, it can be concluded that a learner's level and the different kinds of transfer from their native language are both crucial factors that lead to errors when learning a new language. Apart from the research of language acquisition and cognitive psychology, social and cultural factors are included in the study of language teaching and learning as well. Furthermore, with the rapid development of digital technology, the study of language teaching has not only had a substantial breakthrough in data processing and analysis, but has also been closely connected

with digital content. Since the teaching of language is inevitably oriented by these aspects, it should focus not exclusively on errors due to linguistic influence, but should also take into account the difficulties drawn from cultural factors, social factors, and teaching strategies.

2.1.2 Types of Errors

The terminology "error" in SLA refers to an unconscious mistake that correlates to a learner's native language when they are using the target language. In reference to the errors of learners at different stages when learning a target language, Corder (1976) categorized errors into three types: pre-systematic error, systematic error, and post-systematic error. He further explained that a learner's errors would decrease progressively as their grasp on the grammar system of the target language grew. Amidst the continuum, errors that are produced during the period of pre-system and post-system are the most systemic for learners who have not yet mastered the grammar system of the target language.

From a linguistic point of view, Dulay et al. (1982) discussed learners' error types and divided them into the categories of lexical error and syntactic error. After inspecting learners' output based on the disparity in sentential structures from their target language, the structural errors can be further categorized into four types: omission, addition, misformation, and misordering. Omission indicates that the learner left out a necessary part of the sentence or discourse. Addition refers to the error resulting from a redundant grammatical unit in a sentence or discourse. Misordering references a situation where a grammatical unit is misplaced in a sentence or discourse. Misformation refers to the embedding of an inappropriate grammatical unit in certain structures, namely, an error due to misuse of a grammatical unit. Many studies (James, 1998; Zhou et al., 2007) have analyzed error types through the framework of this categorization.

2.1.3 Cause of Errors

The cause of an error when using the target language demonstrates a learner's tendency to approach the new language with the grammar system of their native language, along with a gap in linguistic knowledge toward the target language. Selinker (1972) suggested that the emergence of interlanguage is drawn from five factors: linguistic transfer, overgeneralization, the impact of pedagogy, learning strategies, and communication strategies. In learning transfer, errors are likely influenced by negative transfers from the native language, a lack of knowledge of the target language, cultural factors, learning environment, teaching strategies, drilling methods, or strategies of interpersonal communication.

Limuria (2014) and Okuno (2018) examined errors in *bei* sentences by Chinese learners from Indonesia and Japan, respectively. Limuria (2014) discussed the difficulties that Indonesian learners encounter when learning *bei* sentences in Chinese

and discovered the cause of the errors through the lens of contrastive analysis and error analysis. In Limuria's research, it was found that addition caused the highest percentage of errors, followed by misordering and misformation. Omission was the least prevalent among the four types. Okuno (2018) also inspected the difference in *bei* sentences in Chinese and Japanese and the error types of learners. The results showed that the errors are mainly caused by the distinction in verb form in Chinese and Japanese. The second reason is the semantic discrepancy in the passive voice between Chinese and Japanese. The third reason is "empathy," which compels Japanese learners to focus on human subjects rather than putting a lifeless object as the subject of the sentence. Furthermore, the study also discovered some errors due to the omission of verb complements and the misuse of psychological verbs. Beyond the typical interference from a native language, some Chinese learners from Japan tend to interchange *rang* and *bei*, or omit *bei* in sentences.

From the studies above, universal errors can be found in learners speaking different native languages. Thus, through the contrast between Chinese and a learner's native language, researchers and teachers can target learners speaking a specific native language and then design specific pedagogy and learning strategies to prevent the possible occurrence of errors, and therefore, improve learning effectiveness.

2.2 Error Analysis of Chinese Learners

There is some research that is concentrated on the error analysis of Chinese learners based on their level, nationality, and knowledge of the four language-learning skills. The results of this research are used to develop corresponding teaching strategies.

2.2.1 Error Analysis of Chinese

In studies related to different levels of learners, Hung (2013) attempted to address the difficulties of potential complements for intermediate learners. The "Interlanguage Corpus of Potential Complement for Learners" used in the study is built with data collected from a self-designed questionnaire. The types and percentages of errors from learners are analyzed through the utilization of an interlanguage corpus relating to the acquisition of potential complements by Chinese learners. On par with the percentage of errors, the frequency, complexity, surface structures, and internal semantic structure of complements are jointly considered for the recommended arrangement of pedagogical grammar. Instructional design and teaching strategies are thereby developed to meet the needs of intermediate Chinese learners exclusively. Finally, the study proposes advice and gives recommended revisions pertinent to the design of and strategies for teaching potential Chinese complements through practical techniques in the classroom.

Huang (2014) spent two academic years collecting data from Chinese-language beginners from Japan. The pilot study analyzed the learners' systemic errors in

monosyllable words in the first year and continuously monitored learners' errors in both monosyllable and two-syllable words in the second year. The results of the research showed that, among monosyllable words, the third tone had the highest percentage of error, followed by the second tone, the first tone, and the fourth tone. As for errors in two-syllable words, the highest percentage is found in the tonal combination that begins with the third tone. Huang (2014) then designed a teaching plan based on the outcome of the research. Firstly, it incorporated the concept of pitch to help learners distinguish different tone values in Chinese, then it compared similar stresses and intonations in Japanese and Chinese, and finally, it included drilling exclusive to the third tone.

Huang (2018) inspected the common errors of intermediate Chinese learners from Korea and English learners from the United States in the construction of "one + classifier." The findings of this research indicate that learners from the United States have a stronger tendency toward using the structure of "one + classifier." Surprisingly, learners from Korea remained rather conservative with their use of the structure "one + classifier." This study highlights that errors are derived from a lack of teaching on how to identify the noun phrase in discourse when teaching classifiers, and the reference of a noun phrase is directly connected with the use of the structure "one + classifier."

To understand the impact of a learner's native language, Chen (2011) examined the reason for Thai-speaking Chinese learners' erroneous use of the structural particle "de" by collecting interlanguage data from questionnaires. The study classifies the Chinese structural particle "de" into "de1" and "de2," with eight subgroups based on pedagogical implications. According to the results of this study, the lack of similar structures, such as "pseudo-genitive" and "separable word," in their native language is the main cause of errors by Chinese learners from Thailand.

Similarly, Chuyen (2015) researched the difficulties that Chinese-language learners from Vietnam encounter when learning alternative question sentences from the aspect of grammatical structure. The study conducted a contrastive analysis of sentences in Chinese and Vietnamese with a postulation: sentence forms that are similar in two languages are rather easy to acquire, while sentences that differ in structure cause potential obstacles. With this postulation, Chuyen (2015) collected data from the questionnaire and discovered the distribution of errors made by Vietnamese learners of Chinese alternative question sentences: omission (65%), addition (17%), misformation (12%), and misordering (6%). The causes of these errors are due to the negative transfer from a native language, influence from teaching materials and pedagogy, intervention from the questionnaire, or a lack of linguistic knowledge of Chinese.

As for the teaching of writing, Wang (2011) studied the acquisition of directional complements of Chinese learners whose native language is German by analyzing students' written text. Questionnaires and error analysis were conducted based on the contrastive analysis of Chinese and German and the discussion of teaching materials. Except for misuse among different directional complements, the findings suggest that aspect markers in Chinese, for instance, *le* and *zhe*, jointly contribute to these interlanguage errors.

Liu (2016) conducted an error analysis on the use of sentential conjunctions in writing by Chinese learners from France. By contrasting the correct sentences and sentences with errors in the scope of a compound sentence, paragraph, and discourse, the study looked into the cause of errors in terms of the semantics, pragmatics, and function of each sentential conjunction. In addition to theoretical explanations, the study also provided an instructional model instantiating "ye" and "temporal conjunctions" on par with the textbook used in teaching "An Easy Approach to Chinese" and "Intermediate Chinese Vol. 1" for practical reference.

Tang (2018) retrieved and examined the use of punctuations in interlanguage sentences by learners speaking English and Japanese in TOCFL Learner Corpus, compiled diagnostic tests and related topics with reference to the standard punctuation systems of Chinese, English, and Japanese, and classified various types of misuse by native speakers. The study discovered that errors from native speakers tend to be from related punctuations, such as "" and "", while errors from learners tend to be unrelated punctuations, such as · and ◦. As specific usage often collocates with certain semantic attributes, both native speakers and learners could misapply punctuation due to the uniqueness in its form or meaning. Indeed, the form and meaning of punctuation from a speaker's native language tend to transfer to the target language. The study listed four situations in different punctuation systems that are particularly difficult for learners: punctuation that is similar in shape but has a restricted meaning, punctuation that exists in a particular language system, punctuation with the same meaning but a different shape, and punctuation with a similar shape but a different meaning. Thus, a language teacher should emphasize the correlation between punctuational attributes and linguistic content, as well as their collocation from an integrated perspective.

2.2.2 Teaching Strategies

Liang (2008) conducted research on the acquisition of Chinese classifiers by adult learners. A total of 68 participants (29 native speakers of Korean, 29 native speakers of English, and 10 native speakers of Taiwanese or Chinese) were asked to complete three types of tests (pairing up classifiers and nouns, pairing up classifiers and pictures, and sequencing classifiers based on concreteness). The results of this study showed that native speakers of Korean performed better than native speakers of English in the experiment. The reason for this is rooted in the similarities between Chinese and Korean. More specifically, classifiers also exist in Korean and the cognitive association with classifiers in Chinese and Korean overlaps. In the test of classifiers that are conceptually connected to shape, the most common images provided by native participants are also the most common images from participants with other native languages. In other words, with reference to the different systems of learners' native languages, different pedagogies should be incorporated when teaching Chinese classifiers to adult learners. Likewise, learners are also expected to have different responses to the pedagogies in terms of levels, learning progress, and types of classifiers.

Cai (2014) investigated the errors in character writing by Chinese learners from Japan through the contrastive analysis of characters in Chinese and Japanese. The study analyzed the errors of 10 Chinese learners from Japan in an advanced Chinese summer program at a university in Taiwan and then offered advice on the textbooks and teaching methods that target Chinese learners from Japan. The findings of this research identified six types of errors that are caused by the negative transfer from Japanese characters: (1) errors of same characters; (2) errors of different characters, but same meanings; (3) errors of same characters, but different meanings; (4) errors of non-Chinese characters; (5) errors of non-Japanese characters; and (6) errors of inverted co-morpheme phrases. As for the advice on teaching, "targetization" must be taken into account; concurrently, teachers should have a rather low tolerance level for errors, and they should remain vigilant in identifying them. Furthermore, with regard to the development of textbooks, materials for Chinese learners from Japan should be based on the contrast of characters in Chinese and Japanese, as well as the distinction between the two writing systems.

Chen (2016) discussed the discrepancy of errors between multilingual learners in international schools and ordinary Chinese learners from Thailand. By inspecting the source of errors from multilingual learners through the application of error analysis and the Principle of Temporal Sequence (PTS), Chen (2016) proposed the Lexical Chunk Approach as the solution to the errors in word order. With four months of practice, errors relating to word order decreased significantly, especially with the use of temporal and spatial adverbial modifiers.

2.3 Studies on Chinese Pedagogical Grammar

2.3.1 Pedagogical Grammar

The discussion of pedagogical grammar has long been central to the field of language teaching. Expanding on the foundation of grammar, pedagogical grammar is regarded as a prescriptive form of language for L2 learners to acquire the grammar of a target language in an integrated and logical way. Through progressive learning, learners are able to process information using the logic of the target language and, as a result, reach accuracy and proficiency. Through examining the performance of individual learners and their errors in written text, information can be provided on their ability to communicate in the prescriptive linguistic form.

While learners face many different challenges when learning a second language, writing is considered to be a relatively difficult skill to acquire. In order to produce written language, a learner must integrate grammar and vocabulary based on correct linguistic knowledge, as well as produce a coherent discourse by combining transitional clauses and sentences. Any error in the incorporation of these factors contributes to the production of ungrammatical sentences. Therefore, it is crucial to incorporate pedagogical grammar in the study of CSL. The present study has identified that pedagogical grammar sets out to address the practical needs of CSL in

order to facilitate a student's acquisition of Chinese grammar and leaves the theoretical aspect to linguistics (Zhou, 2002). As emphasized by Nassaji and Fotos (2011), grammar is rooted in every language system, and as such, language cannot function without grammar.

The theoretical value of pedagogical grammar was first recognized by Odlin (1994), who provided theoretical and systematic evidence for the significance of progressive teaching steps of grammar with reference to syntactical and grammatical theories. Pedagogical grammar is a student-centered approach that requires practicality and prescriptivity to address the factors that influence learners, such as intention, competence, and cognition. The goal of pedagogical grammar is to help learners acquire the target language systematically and efficiently so that they are able to communicate in an authentic context. Since the acquisition of linguistic knowledge and grammatical structure of the target language provides CSL learners with the ability to communicate clearly in all skills (writing, especially), the merit of pedagogical grammar in the study of CSL is great and deserves further recognition.

The theoretical systems that systematically extract collections of grammar have tremendous value to researchers and educators; as such, they ought to be viewed as a corpus that allows for the retrieval of needed information. Lv (2008) offered two suggestions for the choosing and arranging of grammar in CSL textbooks. Firstly, considering practicality and concision, a textbook should only include the basic and frequently-used constructions that are necessary for communication and should eliminate constructions that are unnecessary for the preliminary stage of learning through statistics of frequency. Secondly, regarding the shift of paradigm in pedagogy, a more detailed explanation should be attached to topics, vocabularies, and constructions that have been newly added to textbooks for advancing essential communication skills, such as non-subject sentences and single-word sentences. Furthermore, constructions that are more frequently used in written text, rather than in a colloquial context, ought to be removed from textbooks completely. Lv (2008) argues that the implementation of these suggestions would provide value and enhance the learning outcomes of CSL learners.

Pedagogical grammar is a very important element of CSL learning, and it is critical to helping learners to acquire knowledge. In Yang (2000), he indicated that CSL pedagogical grammar is programmable and that it is not arbitrary or orderless. Therefore, pedagogical grammar can be conducted in accordance with progressive steps, and it remains highly applicable for the instructional setting being sequenced from basic to advanced.

In order to progress the application of pedagogical linguistics, Lu (2000) offered three perspectives relating to the content of pedagogical grammar. The first perspective centers on the essence of Chinese linguistics. Specifically, it seeks to address the question, "What grammar is the most needed and necessary for students?" The second perspective elaborates on the difference between learners' native language and Chinese. Namely, it seeks to address the following questions, "What do the two languages have in common? And what is the difference? What kind of difference would influence the acquisition of Chinese?" The third perspective discusses the role of grammatical errors in language acquisition. It attempts to answer the question,

"What are the most common mistakes students make when learning Chinese?" Lu (2000) also insisted on the implementation of unplanned learning at the preliminary stage of grammar teaching and the necessity of summative "basic grammar consolidation" after learners have reached a high level. With respect to this teaching method, two suggestions are proposed by Lu (2000). Firstly, choosing and arranging teaching materials should not solely depend on the content. Instead, the text should incorporate characters, vocabularies, and grammar that need to be acquired by learners. Nonetheless, the arranging of grammar in a text should be highly regulated. Secondly, a summative "basic grammar consolidation" is necessary once students reach a certain level. All of these suggestions have been proposed with the goal of improving learners' acquisition of Chinese.

2.3.2 The Application of Chinese Pedagogical Grammar to Writing

Several studies have discussed the topic of pedagogical grammar in CSL. Hong et al. (2018) presented a student-centered learning sequence in the cluster of grammatical structures. Additionally, Hsieh (2009), Chen and Lin (2003), and Peng (2003) suggested that communication and writing competence can be cultivated by enhancing a learner's knowledge of grammar. Considering that the incorporation of pedagogical grammar in writing skills and written text is developed from a learner's awareness and metacognition, it is well-accepted that pedagogical grammar plays a crucial role in a learner's use of target language and holds a central place in the study of CSL writing.

The current technology of automatic grading systems of Chinese writing can detect 65% of grammatical errors (Chang et al., 2015) and reach 88% accuracy on the automatic revising system (Hong et al., 2014a, 2014b); however, the accuracy of the automatic grading system of writing remains relatively stagnant. The main reason is the detection of grammatical errors (Chang et al., 2015). Specifically, because the system lacks the grammar that CSL learners need, the precision of identifying errors is unable to make much progress. The appropriateness or difficulties of grammar is closely correlated with the learner's level. Thus, in order to contrast the common grammatical errors made by learners, the present study seeks to categorize and construct the structures of learners' grammatical errors based on different types of errors from the data and expects to further the application in the teaching of writing, as well as the evaluation of learners' writing competence.

2.4 Corpus-Based Studies

Although many language teachers tend to incorporate corpus into the study of language teaching, most of the existing research focuses on analyzing a single grammar rule; only a few among them are integrated studies. These studies can be divided into two kinds. Some studies summarize the frequency of grammar and

offer teaching advice through the utilization of corpus data collected from native speakers. The other studies categorize learners' error types and sequence the difficulty of grammar through the learner corpus, as well as provide advice on pedagogy.

2.4.1 The Application of Corpus in CSL

Chang (2005) implemented Sinica Treebank Version 1.1 (http://treebank.sinica.edu.tw/) to sort out linguistic forms that contain the function of comparison and discovered that "presentative comparison sentences" are the most common form. The study retrieved the frequency, collocation, and mutual information of "*bi*" in Sinica Corpus and lists several frequently used "*bi*" sentences, as well as provides teaching steps for "*bi*" with reference to theories of pedagogical grammar. Chang (2014) observed how learners at different levels and with different native languages (English, Japanese) acquire Chinese relative clauses through data in the learner corpus and offered advice regarding instructional design.

Lin et al. (2014) extracted data that contained the Chinese directional complement "*qilai*" from Chinese Learner Corpus by National Taiwan Normal University (NTNU) and analyzed learners' error distribution to discover possible difficulties and offer advice for the instructional setting.

In order to identify discrepancies in language use, as well as to extract usages that are either completely identical or completely different, Hong and Huang (2013) used WordNet, Chinese WordNet, and the Chinese Concept Dictionary. The study utilized Chinese Word Sketch Engine to examine data from the cross-strait area in Chinese Gigaword Corpus and analyzed the distribution in the corpus. The findings revealed an interesting phenomenon; distinction and mutual influence are restored in the usage of words in the cross-strait areas.

2.4.2 The Application of Corpus on Error Analysis

Wang et al. (2013) put forth that near-synonyms often cause difficulty in teaching, and thus, should be closely examined. Furthermore, they stated that with extensive data from learners' interlanguage, vocabulary errors would be tractable and analyzable. The study opted for the "Chinese Learner Corpus" by NTNU to differentiate the use and error distribution of two groups of near-synonyms, "*bang*," "*bangzhu*," "*bangmang*," and "*bian*," "*biande*," "*biancheng*." The study produced insights on instructional steps in the teaching of near-synonyms by examining the connection between textbooks and learners' errors.

Further research on the acquisition of transition words has been conducted by Tseng and Hsieh (2013). Specifically, they utilized Sinica Corpus and TOCFL Learner Corpus (http://tocfl.itc.ntnu.edu.tw/) to compare the acquisition of the transition word "*er*" by Chinese learners and native speakers of Chinese. The findings showed that, with higher language levels, the conjunctions that learners deploy in discourse appear to transfer from intra-sentence to inter-sentence. Additionally, the

cause of errors is derived from the learner's unawareness of grammatical and semantic restrictions governing different conjunctions. This study demonstrates the usefulness of studying specific aspects of grammar, as it provides tangible and actionable data that can impact student learning.

In light of the lack of research that concentrates on computer-based correction of Chinese word order, Cheng (2014) used "HSK Dynamic Composition Corpus" by Beijing Language and Culture University to collect sentences with errors by foreign learners. Then, a revised corpus was established based on the misordering marking in sentences; misordering was marked by two researchers who speak Chinese as their native language. The study extracted feature engineering from Google Chinese Web 5-g Corpus after retrieving the data set from HSK Dynamic Composition Corpus. The study then generated a series of available combinations that could contain correct sentences by using CRF to detect the possible sections of misordering in sentences. These combinations were then sequenced according to the possibility of correct word order. The research found 83.4% accuracy for identifying sectional misordering and 85.8% accuracy for correcting misordering. The findings of the study are applicable to future research, and the accuracy can be improved by expanding the database.

Further research utilizing the Chinese Learner Corpus was conducted by Tung et al. (2015); they analyzed data from A2 learners and B1 (referring to CEFR proficiency levels) learners, whose native language is English, and calculated the error distribution of "*le*" sentence. The findings of this research provided advice on teaching steps, as well as information that could be used for further examination.

Derived from the aforementioned studies related to corpus linguistics, the corpus provides us with valuable information on the attributes of vocabularies and grammar. The frequency of certain types of sentences and the wording difference in cross-strait areas can all be observed from the corpus data. In addition, through the data from "learners' interlanguage corpus," existing errors have become analyzable and serve as a reference to help understand the possible difficulties that CSL learners may encounter. The findings can also be utilized in future studies and offer implications for practical use.

3 Methodology

3.1 The Learner Corpus

"Chinese Written Corpus" (CWC) (http://140.122.63.128/Index.aspx) is a CSL/CFL written corpus that discovers error patterns from the same written text by learners at different levels. The collected data are then used to construct the self-evaluation system and the feedback system, as well as for the exploration of how a self-evaluation system can be applied to the study of CSL/CFL-based writing (Hong et al., 2014a, 2014b).

The corpus provides information on grade band and error marking from the post-evaluated text and also provides the error sentence and the revised sentence that are applicable in research and teaching, as shown in Figs. 1 and 2.

The grading system used in CWC is in accordance with the proficiency guidelines for writing (ACTFL, 1987, 2012) by the American Council on the Teaching of Foreign languages (ACTFL) and developed from the framework "Rating Scale of Testing Chinese Writing" by Sung et al. (2012). The assessment is composed of four elements: content, grammar, vocabulary, and punctuation. All of the texts are then classified into five levels: excellence, good, advanced, intermediate, and beginner. A total of 11 bands are employed within the level of beginner, intermediate, and advanced as low (band 1–3), medium (band 4–6), and high (band 7–11). The text is then given a score based on the performance of the four elements during the human assessment. When assessing each text, consistency and accuracy are assured by the

Fig. 1 The home page of CWC

Fig. 2 The search result of CWC

Table 1 The distribution of band score from the four texts

Text ACTFL band score	Text 1	Text 2	Text 3	Text 4	Total
Band 3	16	61	106	102	285
Band 4	195	184	267	233	879
Band 5	349	238	159	127	873
Band 6	139	140	94	77	450
Band 7	63	66	33	31	193
Band 8	9	20	6	10	45
Band 9	4	4	1	2	11
Total	775	713	666	582	2736

program monitoring grading criteria, sample texts, the trial assessment by the grader, alignment of the trial assessment, alignment of the assessment, and alignment after the assessment. The goal of this design is to produce meaningful and accurate results.

Most of the data in CWC are collected from Chinese learners of different native languages in the Mandarin Training Center (MTC) at NTNU and 11 other CSL/CFL institutes from September 2010 to December 2016. The existing data in the corpus have been documented with detailed information, such as the title of the text, the learners' name in Chinese and English, nationality, the learners' native language, institute, and so forth; the data has also been restored in the form of a text file or an image file. There are four texts that have been marked and graded and that are utilized in this analysis: "a place worth going," "the beach in summer," "a letter to my family," and "introducing my country." Samples that were completely off-topic or unanswered were deleted during the compilation of the database. The total number of texts is 2,736, the individual number for text 1, text 2, text 3, and text 4 is 775, 713, 666, and 582, respectively, as shown in Table 1.

The present study utilizes four texts, "a place worth going," "the beach in summer," "a letter to my family," and "introducing my country," in Chinese Written Corpus (CWC), with a distribution of grade from band 3 to band 9. The texts are composed of foreign language learners who speak 43 different native languages. Among the data collected from the learners, the number of text are arranged in descending order according to native language; the top five groups are listed as follows: Japanese, English, Vietnamese, Korean, and Indonesian. In light of the diverse background of learners and the disparity of data, the present study only analyzes and discusses the five groups of learners with the highest number of texts (see Table 2).

The error marking system in CWC is supported by WeCan (Chang et al., 2012a, 2012b; Chang et al., 2012a, 2012b) and is able to provide functions such as word segmentation, tagging parts of speech, error marking, and so forth. The system can then export files to be used with programs to support related studies and future development. As for the tagging of parts of speech, the study selects a total of 48 simplified markers that represent 46 simplified markers classified by the Chinese

Table 2 The number of texts from the five groups of learners classified by their native languages

Native language	Text 1	Text 2	Text 3	Text 4	total
Japanese	209	161	160	174	704
English	131	120	154	83	425
Vietnamese	117	113	81	61	378
Korean	99	112	55	60	360
Indonesian	38	54	53	60	175

Knowledge and Information Processing group (CKIP), as well as the items Nominalized Verb (Nv) and Unknown (b) that are manually added by this study. Regarding error marking, the study divides learners' errors into two parts: surface structure and linguistic form. Surface structure refers to "addition," "omission," "misformation," and "misordering," and linguistic form refers to "character," "word," and "punctuation" (see Fig. 3).

The following are the error sentences found in written texts, which are classified into four types of surface structures:

(1) Addition (a place worth going/ACTFL band 7)

*我 已經 離開 家 也 快 十年 了。

* I already left home already almost ten years AM

我 離開 家 也 快 十年 了。

I left home already almost ten years AM

(2) Omission (the beach in summer/ACTFL band 6)

*沙灘 上 有 好多 的 人 曬太陽。

*beach P have many de people bask (in) sun

沙灘 上 有 好多 的 人 在 曬太陽。

beach P have many de people AM bask (in) sun

(3) Misformation (the beach in summer/ACTFL band 5)

*而且 福隆 海邊 是 海水 跟 河水 見面 的 河口。

*And fulong beach SHI sea with river meet de estuary

而且 福隆 海邊 是 海水 跟 河水 相會 的 河口。

And fulong beach SHI sea with river join de estuary

*476 名 的 乘客 中 只 146 名 救助 了。

*476 C de passenger P only 146 C help AM

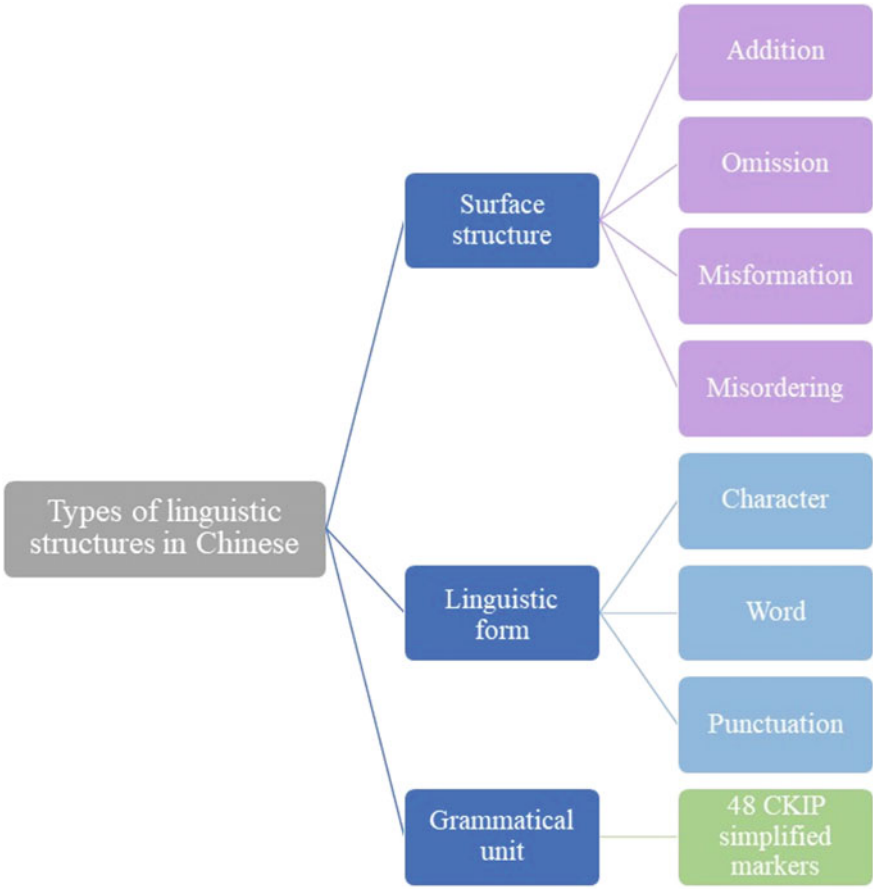


Fig. 3 The types of linguistic structures in Chinese

476 名 的 乘客 中 只 146 名 獲救 了。

476 C de passenger P only 146 C rescue AM

(4) Misordering (a place worth going/ACTFL band 7)

*讓 你 回來 以後 再 想 去 一次。

*Let you come back after again want go once

讓 你 回來 以後 想 再 去 一次。

Let you come back after want again go once

The research steps for this study are divided into two parts: fundamental studies and applied studies. These two categories are then divided into four additional subsections. Fundamental studies are divided into information on learners' errors

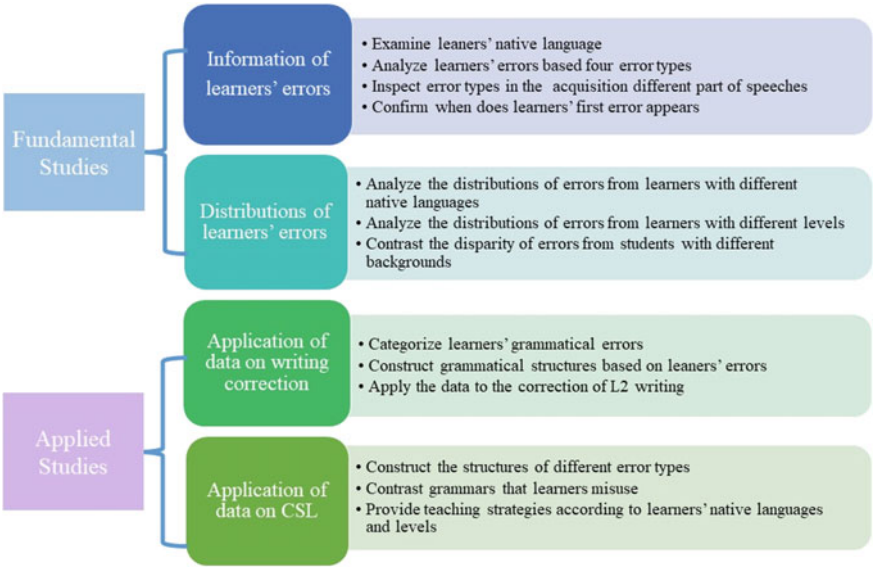


Fig. 4 Research steps of the present study

and distribution of learners' errors. Applied studies are divided into application of data in writing correction and application of data on CSL. The research framework is illustrated in Fig. 4.

3.2 The Reference Corpora

3.2.1 Sinica Corpus

"Academia Sinica Balanced Corpus of Modern Chinese version 4.0" (Chen et al., 1996, http://asbc.iis.sinica.edu.tw/), abbreviated as Sinica Corpus, contains more than ten million word tokens collected from 1981 to 2007. The database is mainly comprised of written language, and each word is segmented and tagged with part of speech. The data are retrieved from texts related to literature, social science, science, philosophy, arts, and so forth, and represent different linguistic modes (written text, manuscript), different writing styles (narrative, essay), different media (newspaper, textbook, audiovisual media), and different themes (science, literature). The corpus has collected 19,427 texts, and has 1,396,133 sentences, 11,245,330 word tokens, 239,598 word types, and 17,554,089 character tokens.

In order to examine the use of written language by native speakers with systematic tagging of parts of speech and to ensure the exclusive use of traditional Chinese in order to maintain the rigor of research, the present study retrieves data from native

speakers from Sinica Corpus. Since CWC and Sinica Corpus have the same tagging system for parts of speech, the present study can conduct a contrastive analysis through the comparison of the written text in CWC and data from native speakers in Sinica Corpus.

3.2.2 The Digital Platform of Chinese Grammar (DPCG)

"The Digital Platform of Chinese Grammar version 4.3.3." (DPCG) (http://203.64.95.103:8089/SyntaxSystem/) seeks to integrate "teaching" and "learning" in theory and practice. For teachers, it provides insight into possible obstacles that learners may encounter. For learners, the platform offers information on learning steps based on the frequency of different elements of grammar. For the development of textbooks, the platform merges teaching steps and error frequency to facilitate the compiling of teaching materials for CSL. Future research can conduct experiments pertaining to the teaching of written language and incorporate CWC as a resource and target in the study of CSL (see Fig. 5).

The DPCG brings together perspectives from native speakers, L2 learners, and textbook development by combining Chinese Gigaword Corpus (LDC, 2009) and CWC for the frequency of grammar that native speakers deploy on a daily basis and data from Chinese learners to accurately analyze the use of grammar and error frequency by learners at different levels. Through cross-checking the results and the illustration of the frequency quadrants, the platform presents a thorough analysis of the arrangement of grammar in the four textbooks that are commonly used in CSL learning: "A Course in Contemporary Chinese" (2015), "Road to Success: Threshold" (2008), "Practical Audio-Visual Chinese" (2007), and "New Practical Chinese Reader Textbook" (2002). The results that are presented in the platform offer evidence-based advice on the teaching of frequently-used grammar, as well as

Fig. 5 The home page of DPCG

Fig. 6 The frequency quadrants and sample sentences in DPCG

sentences from native speakers and error sentences from learners. Furthermore, the results are used to study the development of frequency quadrants of CSL learners (see Fig. 6).

A comparison of the data in Chinese Gigaword Corpus and CWC has led the present study to classify four quadrants that correspond to a learner's learning progress using frequency in Chinese Gigaword Corpus as the X-axis and error frequency in CWC as the Y-axis: "commonly used, high error frequency," "commonly used, low error frequency," "seldom used, high error frequency," and "seldom used, low error frequency." The four quadrants are designed to determine the appropriate steps that should be taken when teaching grammar. For example, if a grammatical construction appears in the quadrant of "commonly used, high error frequency" after comparing frequency in the two corpora, it should be taught prior to other constructions and vice versa. Likewise, teachers can understand the use of each construction by native speakers and learners and decide if certain constructions should be emphasized or underemphasized in teaching. The platform also provides error sentences by learners for instructional purposes. Overall, the four quadrants are designed to provide actionable information to teachers and learners.

4 Result and Discussion

4.1 Overall Distribution of Error Types in the Learner Corpus

The number of error sentences in the text is roughly 100,000. Among all four types of errors, misformation accounts for about 50% of the errors, which is significantly higher than other error types.

The reason for the disproportionate percentage of misformation is due to the vagueness of near-synonyms and the difficulties that arise in teaching (Hong and Sung, 2017). The semantic vagueness not only causes miscomprehension and confusion, but also leads to misuse in practice. Furthermore, misformation is prevalent among all texts by learners from different levels, which indicates that the problem of misformation is not alleviated by a learner's advancement in language competence (Cai, 2010). Hence, miscomprehension of near-synonyms ultimately gives misformation a rather salient portion of the four error types.

The possible applications of the data collected from CWC include analyzing learners' error types in written text based on the surface structure of language and examining the distribution of errors according to grammatical features, namely, parts of speech. The parts of speech of data in the present study are tagged in accordance with the 48 CKIP simplified markers in Sinica Corpus. The major categories are noun (N), verb (V), adjective (A), conjunction (C), adverb (D), interjection (I), postposition (P), particle (T), "*de*, *zhi*, *de*, *de*" (DE), "*shi*" (SHI), and foreign word (FW). Generally speaking, colloquial context and written language are primarily composed of units such as noun, verb, adjective, conjunction, adverb, and so forth. Particularly, in light of the uniqueness of its grammatical structure, *shi* not only holds a special place in the study of Chinese linguistics, but is also categorized as a transitive verb in the tagging by Sinica Corpus. Furthermore, based on observations from learners' writing proficiency, *shi* remains one of the most frequently-used linguistic errors at all levels (Hong and Sung, 2017). The words found in these six main categories tend to be the most commonly used on a daily basis. Thus, the present study aims to inspect the number of error sentences based on the parts of speech by conducting a cross-checking analysis. From the statistic results shown in Table 3, it can be seen that with addition and omission, most errors occur in the learning of adverbs, and the number of errors in the noun category is the second. As for misformation and misordering, the number of errors in the noun category dominates in both types. The second highest in terms of the number of errors in misformation and misordering are verb and adjective, respectively.

Table 3 The statistics of the error types in CWC

Types of structural errors	Number of error	Percentage of error (%)
Addition	22,496	21.61
Omission	28,874	27.74
Misformation	**48,355**	**46.46**
Misordering	4352	4.18
Total	104,077	100.00

Table 4 The statistics of error types based on parts of speech

Part of speech / Error types	Noun	Verb	Adjective	Conjunction	Adverb	*Shi*	Total
Addition	4698	2946	653	955	6547	1094	16,893
Omission	4061	2331	386	1047	7690	932	16,447
Misformation	9383	7019	2888	1791	5100	319	26,500
Misordering	563	362	443	80	159	25	1632
Total	18,705	12,658	4370	3873	19,496	2370	61,472

4.2 Distribution of Error Types Among Different Learner Variables

Many studies (Chen, 2011; Hung, 2013; Limuria, 2014; Okuno, 2018; Huang, 2018; Tang, 2018) have revealed that learners' errors tend to appear in different aspects. The present study aims to analyze the distribution of learners' errors in terms of learners' native language, level, and the use of parts of speech.

4.2.1 Native Language as the Variable

Despite classifying learners into different groups based on their native languages, according to the statistics result, the top five groups of learners (Japanese, English, Vietnamese, Korean, and Indonesian) have the same distribution and tendency for errors. As shown in Table 4, the most common type of error is misformation, followed by misordering. This suggests that, in spite of the diverse background of native languages, learners' errors in surface structure appear to be highly consistent. In addition to the impact of individual native language, the study also accounts for the reason and distribution of errors to form an integrated perspective.

4.2.2 Proficiency Level as the Variable

As with the distribution of errors by learners speaking different native languages, misformation dominates in the number of errors and remains as the main error type in all of the incorrect sentences with proficiency level as the variable. On the contrary, the number of misordering is remarkably lower than the other three error types. Addition and omission present less discrepancy in the total number of incorrect sentences. From the data in Tables 5 and 6, a universal trend can be seen in that the distribution of the four error types remains the same, regardless of a learner's native language or proficiency level.

Table 5 The statistics of errors based on learners' native languages

Native language / Error types based on surface structure	Japanese	English	Vietnamese	Korean	Indonesian
Addition	5486	2749	4044	3931	1629
Omission	6845	3438	6459	3984	1762
Misformation	11,575	6251	8165	7606	3391
Misordering	1221	494	742	686	272

Table 6 The number of sentences with different error types in different bands[2]

ACTFL band score / Error types based on surface structure	Band 3	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9	Total
Addition	2204	7055	7070	3990	1725	385	58	22,487
Omission	3546	9446	9097	4511	1831	375	61	28,867
Misformation	4576	14,960	15,575	8817	3476	768	164	48,336
Misordering	563	1375	1363	723	250	64	14	4352
Total	10,889	32,836	33,105	18,041	7282	1592	297	104,042

4.2.3 Part of Speech as the Variable

Apart from a learner's native language and proficiency level, parts of speech as the variable have the potential to provide valuable information on the overall distribution of error types to provide a holistic view of a learner's performance. Based on the data retrieved from CWC, this study will discuss how the six parts of speech, noun, verb, adjective, conjunction, adverb, and *shi*, present in the four types of errors in surface structure in the following section.

In the distribution of the first error type, addition/adverb appears to be the part of speech that is easily misused in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than in other parts of speech. Regarding other parts of speech, texts with the highest mean of sentences with the addition of noun, adjective, and conjunction are found in band 6. Also, the addition of verb and *shi* in sentences are particularly noticeable in band 7. However, the most dominant mean of sentences with the addition of adverb exists in band 8, rather than at the intermediate level. The distribution of data reveals that learners at the intermediate level tend to insert redundant units into sentences.

[2] The statistics in Table 3 are retrieved from CWC directly and constitute incorrect sentences from band 1 to band 9, and thus different from the statistics shown in Table 6, which includes data from band 3 to band 9 only. Due to the exclusion of band 1 and band 2, the number of incorrect sentences differs slightly in addition, omission, and misformation. However, the number remains identical in misordering because students in band 1 and band 2 are not exposed to long sentential structure, but instead short phrases of survival language. Hence, the error type of misordering does not exist in band 1 and band 2.

In the distribution of the second error type, omission/adverb appears to be the part of speech that learners most commonly misuse in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than in other parts of speech, which aligns with the tendency in the first error type, addition. In regards to other parts of speech, texts with the highest mean of sentences with the omission of nouns are found in band 7. The omission of verbs is particularly excessive in band 5, and the omission of adjectives is prominent in band 8. As for conjunctions, band 6 and band 7 both have the highest number of sentences with incorrect omissions. The omission of adverbs, on the other hand, is discovered to be most salient in band 6. Lastly, the omission of *shi* is particularly noticeable in band 7 and band 8. The distribution of data indicates that the error of omission is more obvious among learners at the intermediate and advanced levels.

In the distribution of the third error type, misformation/adverb appears to be the part of speech that learners most commonly misuse in texts at different levels. The number of incorrect sentences with redundant adverbs is significantly higher than other parts of speech, which aligns with the tendency of the aforementioned error types. As for other parts of speech, texts with the highest mean of sentences with the omission of nouns are also found in band 7. The misformation of verbs is detected to be excessive in band 5, and the misformation of adjectives is relatively noticeable in both band 6 and band 7. The texts with the highest mean of sentences with the misformation of adverbs are found in band 6. Finally, the misformation of *shi* is particularly dominant in band 7 and band 8. The distribution of data indicates that the error of misformation, similar to the error of omission, should receive extra attention among learners at the intermediate and advanced levels.

When examining the error of misordering, this study discovers that it appears to be the most divergent in terms of distribution among the four error types. The misordering of nouns is found to be most salient among learners from band 4 to band 6. Nevertheless, for beginner and advanced learners, the misordering of adjectives dominate in number. With respect to detailed information, the highest mean of sentences with misordering of nouns is found in the text of band 5. For the misordering of verbs and conjunctions, the highest means of sentences in the texts both appear in band 9. The misordering of adverbs, however, is relatively noticeable in band 6 and band 7. Lastly, the misordering of *shi* is especially pronounced in band 6. In conclusion, the error of misordering appears to be particularly significant among advanced learners.

The overall pattern of error distribution based on each part of speech is depicted in Table 7, which shows the mean of sentences in a text with incorrect parts of speech in different band scores and error types.

5 Conclusion

In general, sentences in written text, compared to colloquial data, appear to be more complex in terms of linguistic form and are expected to adhere to the framework

Table 7 The mean of sentences in a text with incorrect parts of speech in different band scores and error types

		Band 3	Band 4	Band 5	Band 6	Band 7	Band 8	Band 9
Addition	Noun	1.74	1.66	1.68	1.89	1.77	1.67	1.00
	Verb	1.02	1.11	1.07	1.06	1.12	0.84	0.64
	Adjective	0.25	0.24	0.23	0.26	0.24	0.07	0.09
	Conjunction	0.37	0.31	0.34	0.42	0.38	0.36	0.18
	Adverb	2.17	2.33	2.31	2.62	2.75	3.00	2.00
	Shi	0.35	0.43	0.37	0.42	0.46	0.40	0.09
Omission	Noun	1.55	1.45	1.52	1.44	1.59	1.31	0.82
	Verb	0.81	0.79	0.94	0.90	0.76	0.69	0.55
	Adjective	0.16	0.13	0.15	0.14	0.12	0.22	0.00
	Conjunction	0.34	0.34	0.41	0.44	0.44	0.24	0.18
	Adverb	2.76	2.75	2.88	2.90	2.81	2.38	1.73
	Shi	0.32	0.29	0.36	0.37	0.42	0.42	0.36
Misformation	Noun	3.47	3.14	3.42	3.82	3.82	3.80	4.82
	Verb	2.27	2.38	2.55	3.03	3.03	2.89	3.36
	Adjective	1.00	0.95	1.02	1.28	1.27	0.91	1.64
	Conjunction	0.61	0.62	0.71	0.59	0.75	0.78	0.55
	Adverb	1.76	1.83	1.84	2.04	1.94	1.73	1.00
	Shi	0.11	0.12	0.12	0.11	0.10	0.02	0.18
Misordering	Noun	0.20	0.19	0.22	0.23	0.16	0.18	0.18
	Verb	0.09	0.13	0.16	0.14	0.10	0.11	0.18
	Adjective	0.11	0.16	0.18	0.16	0.18	0.11	0.27
	Conjunction	0.02	0.03	0.04	0.03	0.01	0.02	0.09
	Adverb	0.04	0.06	0.05	0.07	0.07	0.02	0.00
	Shi	0.00	0.01	0.01	0.02	0.01	0.00	0.00

of prescriptive grammar. Ideally, in a practical context, teachers would only teach grammar that is confined to certain norms, and students would, therefore, be exclusively exposed to prescriptive usages. However, in the texts used in this study, various errors are spotted in vocabulary and grammar. Thus, the present study seeks to assist teachers in discovering students' potential grammatical errors by identifying the types and patterns of errors with the support of data from CWC. Apart from examining the existing errors, this study also attempts to improve the effectiveness of error identification. The previous research has yielded little progress in identifying errors by comparing students' written text with reference to correct grammar. Hence, this study contrasts students' written texts with the structures of grammatical errors categorized in the research and further discovers the distribution of learners' errors on parts of speech in hopes of advancing the effectiveness and efficiency of the error

identification system. The findings of the present study reveal two universal distributions in learners' error types. Firstly, among all four error types, misformation appears to be the most common error, while misordering is the rarest, regardless of a learner's background. Secondly, based on the observed association between error types and parts of speech, it appears that learners often have difficulty with adding and omitting adverbs in a sentence, and therefore, have a tendency to misform nouns and verbs.

Furthermore, since a learner's native language and level often play a crucial role in organizing teaching activities, one element of CWC is its error marking system and graded texts. Through the application of the error marking system and graded texts, future studies can conduct cross-checking based on the existing data and design teaching strategies for learners speaking different native languages or at different levels. Through the error analysis of learners' texts, as well as contrasting the distribution and frequency of various grammar errors in CWC, the present study constructs different error types and identifies shared error types among learners at different levels. The findings of the study offer insights into the implementation of teaching strategies as well as methodologies at different levels.

References

ACTFL Chinese Proficiency Guidelines. (1987). *Foreign Language Annals, 20(5),* 471–481.

ACTFL Proficiency Guidelines 2012–Writing. (2012). Retrieved from http://actflproficiencyguidelines2012.org/writing.

Ashwell, T. (2000). Patterns of teacher response to student writing in a multi-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing., 9*(3), 227–257.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart & Winston.

Buckingham, T., & Pech, W. C. (1976). An experience approach to teaching composition. *TESOL Quarterly, 10*(1), 55–65.

Cai, B. G. (2010). An investigation and analysis of errors in using synonymous action verbs in interlanguage. *Chinese Teaching in the World, 4*, 526–535.

Cai, Q. Y. (2014). The analysis of chinese character writing and words usage errors made by Chinese-Japanese learners and suggestions. *Chung Yuan Journal of Teaching Chinese as a Second Language, 17*, 53–77.

Chandler, J. (2003). The efficacy of various kinds of error feedback for improvement in the accuracy and fluency of L2 student writing. *Journal of Second Language Writing, 12*, 267–296.

Chang, L. P. (2005, July). Qiantan jiaoxueyufa–cong nijiaoju tanqi. Paper presented at the 8th Forum of Chinese Teaching. China, Beijing. [張莉萍. (2005, 7月). 浅談教學語法–從比較句談起. 第八屆漢語教學討論會. 中國北京.].

Chang, L. P. (2014). Salient linguistic features of chinese learners with different L1s: A corpus-based study. *International Journal of Computational Linguistics and Chinese Language Processing, 19*(2), 53–72.

Chang, T. H., Sung, Y. T., & Hong, J. F. (2015). Automatically detecting syntactic errors in sentences writing by learners of chinese as a foreign language. *International Journal of Computational Linguistics & Chinese Language Processing, 20*(1), 49–64.

Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012a, November). *A Chinese word segmentation and POS tagging system for readability research.* Paper presented at the 42nd Annual Meeting of the Society for Computers in Psychology (SCiP 2012), Minneapolis, MN.

Chang, T. H., Sung, Y. T., Lee, Y. T., & Hsieh, G. S. (2012b, October). Zhongwen duanci zhi keduxing yingyongyanjiu. Paper presented at The 51st Annual Conference of Taiwan Psychology Association. Taichung: Asia University. [張道行, 宋曜廷, 李垚曒, & 謝冠生 (2012, 10月). 中文斷詞之可讀性研究應用. 發表於台灣心理學會第五十一屆年會. 台中, 亞洲大學.].

Chen, C. L. (2011). An error analysis and pedagogical study of DE in Mandarin Chinese for thai learners. (Master's thesis). Department of Chinses as a Second Language at National Taiwan Normal University, Taipei.

Chen, H. H., & Lin, C. H. (2003, October). Yunyong wangluhudong jinxing huayuwen xiezuo xuexi zhi tantao. Paper presented at International Conference on Internet Chinese Education (ICICE). Taipei: Overseas Community Affairs Council. [陳懷萱, & 林金錫 (2003, 10月). 運用網路互動進行華語文寫作學習之探討. 第三屆全球華人網路教育研討會. 臺北: 僑委會.].

Chen, K. J., Huang, C. R., Chang, L. P., & Hsu, H. L. (1996, December). Sinica corpus: Design methodology for balanced corpora. In *Proceedings of the 11th Pacific Asia Conference on Language, Information and Computation* (pp. 167–176).

Chen, P. (2016). A study of chinese word order errors and its pedagogy for multilingual learners-a case study in an international school in Bangkok, Thailand. *Journal of Language and Literature Studies, 29*, 191–232.

Cheng, S. W. (2014). Chinese word ordering errors detection and correction for non-native Chinese language learners. (Master's thesis). Department of Computer Science and Information Engineering at National Taiwan University, Taipei.

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1–61.

Chuyen, V. T. (2015). The alternative question in Chinese Vietnamese language: Using Contrastive Analysis as a Reference Design in Grammar Teaching. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Corder, S. P. (1967). The significance of learner's errors. *IRAL-International Review of Applied Linguistics in Language Teaching, 5*(4), 161–170.

Dulay, H. C., Burt, M. K., & Krashen, S. D. (1982). *Language two.* New York: Oxford University Press.

Fathman, A. K., & Whalley, E. (1990). Teacher Response to Student Writing: Focus on Form versus Content. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 178–190). Cambridge: Cambridge University Press.

Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*, 161–184.

Hong, J. F., Chiu, S. W., Sung, Y. T., & Chang, T. H. (2018). Applying Chinese hierarchical grammar bank to the evaluation of chinese writing instruction. *Journal of Technology and Chinese Language Teaching, 9*(2), 40–60.

Hong, J. F., & Huang, C. R. (2013). Cross-strait lexical differences: A comparative study based on chinese gigaword corpus. *Computational Linguistics and Chinese Language Processing, 18*(2), 19–34.

Hong, J. F., & Sung, Y. T. (2017). Huayu xiezuo yuliaoku jianzhi yu fenxi. In H.-J. H. Chen (Ed.), *Corpus and teaching Chinese as a second language* (pp. 197–229). Taipei: Taiwan Higher Education Press Co. [洪嘉馣, & 宋曜廷. (2017). 華語文寫作語料庫建置與分析. In 陳浩然 (Ed.) 語料庫與華語教學 (pp. 197–229). 臺北: 高等教育出版社.].

Hong, J. F., Chang, J. Y., Chang, T. H., & Sung, Y. T. (2014a, October). Huayu wei waiyu zhi xiezuo zidongpinggu yu jiaoxue. Paper Presented at the 1st CLTA-ISCLTL. U.S., Indiana. [洪嘉馣, 張人懿, 張道行, & 宋曜廷 (2014a, 10月). 華語為外語之寫作自動評估與教學. 全美中文教師學會第一屆中文教學國際研討會 (CLTA-ISCLTL). 美國印第安納.].

Hong, J. F., Sung, Y. T., & Chang, T. H. (2014b, November). Yi fenji gainian jianzhi huayuwen xiezuo yuliaoku. Paper presented at The 10th Taiwan E-Learning Forum, 2014 (TWELF). Taipei:

National Taiwan Normal University. [洪嘉馏, 宋曜廷, & 張道行 (2014b, 11月). 以分級概念建置華語文寫作語料庫. 第十屆臺灣數位學習發展研討會. 臺北: 臺灣師範大學.].

Hsieh, C. Y. (2009). Application of controlled writing to teaching in Chinese writing: Organization of writing. *Journal of Applied Chinese, 5*, 225–262.

Huang, H. C. (2014). An analysis of the strategies of teaching mandarin tones to Japanese learners of Chinese at the basic level. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Huang, T. G. (2018). Teaching "Yi+Classifier" to native speakers of English and Korean in intermediate Chinese classes: Error analysis and the designing of a pedagogical decision tree. *Taiwan Journal of Chinese as a Second Language, 17*, 153–183.

Hung, J. W. (2013). Interlanguage analysis and a study of the teaching strategies of Chinese potential complement for the intermediate Chinese learners. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

James, C. (1998). *Errors in language learning and use: Exploring error analysis.* London, UK: Addison Wesley Longman.

Liang, N. S.-Y. (2008). The acquisition of Chinese shape classifiers by L2 adult learners. In Marjorie K.M. Chan & H. Kang (Eds.), *Proceedings of the 20th North American Conference on Chinese Linguistics (NACCL-20), 1,* (pp. 309–326). Columbus, Ohio: The Ohio State University.

Limuria, R. (2014). An error analysis on chinese passive voice produced by indonesian-speaking learners and suggested teaching notes. (Master's thesis). Department of Teaching Chinese as a Second Language at Chung Yuan Christian University, Taoyuan.

Lin, Y. T., Chen, H. J. H., & Wang, C. C. (2014). A Learner Corpus-based Study on Chinese Directional Complement "Qilai." *Journal of Chinese Language Teaching., 11*(4), 73–109.

Liu, S. C. (2016). An analysis of french learner's conjunction errors in writing with pedagogical implications—based on méthode de chinois I–II. (Master's thesis). NTU Graduate Program of Teaching Chinese as a Second Language, Taipei.

Liu, X. (Ed.). (2002). *New practical Chinese reader* (Vol. 1). Beijing: Beijing Language and Culture University Press.

Lu, J. M. (2000). Duiwai hanyu jiaozhong de yufa jiaoxue. *Language Teaching and Linguistic Studies*, *3*, 1–8. [陸儉明. (2000). 對外漢語教中的法學. 語言教學與研究. 3, 1–8.].

Lv, W. H. (2008). *Duiwai hanyu jiaoxue yufa tansuo.* Beijing: Beijing Language and Culture University Press. [呂文華. (2008). 對外漢語教學法探索. 北京: 北京 語言大學出版社.].

Nassaji, H., & Fotos, S. (2011). *Teaching grammar in second language classrooms: Integrating form-focused instruction in communicative context*. New York: Routeledge.

Odlin, T. (1994). *Perspectives on pedagogical grammar.* Cambridge University Press.

Okuno, A. (2018). Error analysis of "Bei" construction by Japanese learners. (Master's thesis). NTU Graduate Program of Teaching Chinese as a Second Language, Taipei.

Peng, N. S. (2003). A study of chinese texts, reading and writing from the systemic-functional linguistics perspective. *Journal of University of Taipei, 44*(2), 33–62.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 209–231.

Tang, N. P. (2018). An error analysis of the interlingual Chinese users from Japan, English-speaking Regions, and Chinese Speech Communities including a Contrastive Analysis of Chinese, English, and Japanese Punctuation (Master's thesis). Department of Chinese as a Second Language at National Taiwan Normal University, Taipei.

Tseng, Y. C., & Hsieh, M. L. (2013). Second language acquisition of Chinese conjunction "*er*" (and): A corpus-based study. *Taiwan Journal of Linguistics, 11*(1), 125–172.

Tung, T. Y., Chen, H. J. H., & Yang, H. M. (2015). The error analysis of "Le" based on "Chinese Learner Written Corpus." *Computational Linguistics and Chinese Language Processing, 17*, 76–95.

Wang, Y. C. (2011). Interlanguage analysis of directional complements for german learners of Chinese. (Master's thesis). Department of Chinese as a Second Language at National Taiwan Normal University, Taipei.

Wang, Y. T., Chen, H. J. H., & Pan, Y. T. (2013). Investigation and analysis of chinese synonymous verbs based on the Chinese learner corpus: Example of “bang”, “bang-zhu”, “bang-mang” and “bian”, “bian-de”, “bian-cheng.” *Journal of Chinese Language Teaching, 10*(3), 41–64.

Yang, J. Z. (2000). Duiwai hanyu jiaoxue chuji jieduan yufa xiangmu de paixu wenti. *Language Teaching and Linguistic Studies*, *3*, 9–14. [楊寄洲. (2000). 對外漢語教學初級階段語法項目的排序問題. 語言教學與研究, 3, 9–14.].

Zhang, H. (Ed.). (2008). *Road to success: Threshold*. Beijing: Beijing Language and Culture University Press.

Zhou, X. B. (2002). Chinese as a foreign language: Characteristics of instructional grammar. *Journal of Sun Yat-sen University (Social Science Edition), 42*(6), 137–142. [周小兵. (2002). 漢語第二語言教學語法的特點. 中山大學學報(社會科學版), 42(6), 137–142.].

Zhou, X. B., Zhu, Q. Z., & Zheng, X. Y. (2007). Waiguoren xue hanyu yufa piawu yanjiu. Beijing: Beijing Language and Culture University Press. [周小兵, 朱其智, & 鄭小宇. (2007). 外國人學漢語語法偏誤研究. 北京: 北京語言大學出版社.].

An Analysis on the Missing of the Adverb 都 *Dou* by CSL Learners Based on an Error-Tagged Learner Corpus

Ting-Yu Yang, Hui-Mei Yang, Wei-Jei Lee, Chen-Yu Liu, and Howard Hao-Jan Chen

Abstract Learners' difficulties in correctly using adverbs have long been reported in CSL/CFL research, and findings yielded in previous research have shed some lights on CSL/CFL learners' patterns and causes of misuse. Many of the investigated adverbs in these studies, however, were subjectively selected by the researchers and might not cover the common errors in learners' production. To more objectively identify common adverb-based errors in CSL/CFL learners' writing, this study extracted adverb-based errors from the error-tagged Chinese Learner Written Corpus of National Taiwan Normal University and discovered that missing of adverbs occurred much more frequently than other adverb-based errors. Among the 2,923 tokens of omitted adverbs, missing of the adverb 都 *dou* ranked first, which accounted for 18% of the same error type. Further analysis of the 526 tokens of omitted 都 *dou* revealed that 都 *dou* was mostly misused when serving as a scope adverb (461 tokens). In addition, the omission of 都 *dou* often occurred when the quantified NPs included 每 *mei*, 所有的 *suoyoude*, 任何 *renhe*, 隨時 *suishi*, and 到處 *daochu*, taking up 56.62% (261 tokens) of the 461 tokens. A follow-up examination of how 都 *dou* is presented in the learners' textbook indicated that the high percentage of omitting obligatory 都 *dou* in these sentences might relate to the inadequate explanation of 都 *dou*'s correct usage in the textbooks. Based on the findings, suggestions for future material writing are offered for CSL/CFL learners' better acquisition of 都 *dou.*

Keywords CSL acquisition · Adverb 都 *dou* · Error analysis

T.-Y. Yang (✉) · H.-M. Yang · W.-J. Lee · H. H.-J. Chen
Department of English, National Taiwan Normal University, 162, Section 1, Heping East Road, Taipei City 106, Taiwan
e-mail: christiney37@gmail.com

C.-Y. Liu
English Language Center, Ming Chuan University, 5, De Ming Road, Taoyuan City 333, Taiwan

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_9

1 Introduction

1.1 Adverbs in CSL/CFL Research

In Modern Chinese, adverbs are often considered complicated to use because of their abstract meanings and complex syntactical, semantic, and pragmatic functions (Duan, 2008; He, 2006; Zang, 2010), and these features have caused learners of Chinese as a second/foreign language (CSL/CFL) great difficulties in successful acquisition (Zheng, 2006). Hence, many researchers have been working on investigating CSL/CFL learners' erroneous uses of specific adverbs (e.g., Gao, 2011; Zhang, 2007), of certain types of adverbs (e.g., Tan, 2012; Zheng, 2006), and with different first languages (e.g., Jiang, 2013; Rong, 2008). Most of these studies categorized learners' errors into four major types (i.e., omission, addition, misselection, and misordering) and analyzed the causes of identified errors. Although previous research has revealed the types and causes of some adverb errors in CSL/CFL learners' writing, adverbs investigated in these studies were often subjectively selected by the researchers and might not be the most commonly misused adverbs by CSL/CFL learners. To more objectively and systematically identify common adverb errors, employment and analysis of error-tagged corpora are suggested.

In Chap. 5, we error-tagged the two-million-word Chinese Leaner Written Corpus of National Taiwan Normal University (NTNU) and identified totally 119 types (48,266 tokens) of error in the learner corpus. Among the top 10 common error types of the total errors, three out of them were adverb-based errors (i.e., missing of adverbs, incorrect selection of adverbs, and redundant adverbs), and the summation of these three error types' tokens accounted for more than 10% of the total errors, showing that adverbs are indeed difficult for CSL learners to use and worthy of further investigation. Among these adverb-based error types, missing of adverbs was more common than the other two types and ranked third among the 119 error types. There were totally 2,923 tokens of errors resulting from adverb missing, and further examination of these errors revealed that the adverb 都 *dou* was the most frequently omitted adverb, which amounted to 18.03% (527 tokens) of the 2,923 errors. Since 都 *dou* was more frequently omitted than other adverbs by CSL learners, this study hence sets out to investigate how and why CSL learners omit 都 *dou* in their writing. Furthermore, we will examine whether the way 都 *dou* presented in teaching/learning materials relates to learners' misuse of the adverb or not.

2 都 *Dou* in CSL/CFL Research

2.1 Research on 都 Dou as a Scope Adverb

In Chinese, 都 *dou* generally performs three functions, namely, scope adverb, modal particle, and time adverb (Liu, 2019; Lu, 1980; Zhang, 2003, 2005). As a scope adverb, 都 *dou* is used to quantify universally quantified noun phrases (NPs), plural NPs, bare NPs, and definite singular NPs (Lin, 1998) "to indicate that all items referred to by the subject or object noun have something in common (Teng, 2019, p. 102)", as shown in sentence (1) a–d.

(1) a. 她每件事都不喜歡。
Ta meijianshi dou bu xihuan.
'She dislikes everything.'
b. 他們都是學生。
Tamen dou shi xuesheng.
'They are students.'
c. 書都放在書架上。
Shu dou fang zai shujia shang.
'Books are on the shelf."
d. 那本練習簿我都做完了。
Naben lianxibu wo du zuo wanliao.
'I have finished that workbook."

In sentence (1)a, 都 *dou* quantifies the universally quantified NP *sheme* "everything", In sentence (1)b, 都 *dou* quantifies the plural NP *tamen* "they". In sentence (1)c, *shu* "books" is the bare NP quantified by 都 *dou*. In sentence (1)d, *naben lianxibu* "that workbook" is the definite singular NP quantified by 都 *dou*.

Liu et al. (1996) generalized about five conditions that the scope adverb 都 *dou* are often required for a well-formed sentence as follows:

1. The quantified subject includes 每 *mei*, 所有的 *suoyoude*, 一切 *yiqie*, and 任何 *renhe*, or there are 隨時 *suishi* or 到處 *daochu* in the sentence.
2. The quantified subject is a plural NP.
3. The quantified subject is formed by a *wh*-word to express universal meanings.
4. An affirmative sentence which includes 無論 *wulun*, 不論 *bulun*, or 不管 *buguan*.
5. An interrogative sentence which is formed by interrogative pronouns like 誰 *shei*, 什麼 *shenme*, 哪兒 *naer*, or 哪 *na* + quantifier.

Except for Condition 2, 都 *dou* is syntactically obligatory in the other four conditions when the quantified NPs occur in a preverbal position (cf. Cheng, 1995; Li, 2013a, 2013b; Lin, 1998; Yuan, 2009). Absence of 都 *dou* in these conditions will form ungrammatical sentences, whereas absence of 都 *dou* in Condition 2 is still syntactically grammatical (Chao, 1968; Li & Thompson, 1981; Tsai, 2014). Although

the use of 都 *dou* in Condition 2 is syntactically optional, missing of this adverb does cause a difference in meaning. That is, the appearance of 都 *dou* in Condition 2 expresses an exhaustive meaning. For example, in sentence (1)b, the presence of 都 *dou* indicates that each person in the group referred to by *tamen* "they" is a student. On the contrary, absence of 都 *dou* in (1)b does not contain the emphatic meaning on the exhaustiveness of the group. The presence of 都 *dou* in such condition is hence not semantically optional and is often suggested to be used (Liu et al., 1996; Lu, 1980).

While the above occasions require the presence of 都 *dou* to form a syntactically and/or semantically correct sentence, past research on Chinese CSL/CFL learners' use of 都 *dou* reveals learners' strong tendency of omitting 都 *dou* in their writing. Luo (2016) investigated CSL learners' use of 都 *dou* as a scope adverb based on data retrieved from corpora and data generated from a self-made questionnaire, and he identified 28 and 172 erroneous sentences of misused 都 *dou* in the corpora and the questionnaire respectively. He then categorized these misuses into four types (i.e., omission, addition, misselection, and misordering) and found missing of 都 *dou* the second common error type, which accounted for 32% and 30% out of the total errors in the corpora and the questionnaire, respectively. Other studies have even reported that omission of 都 *dou* is the most common error type than others. Li (2013a, 2013b) examined CSL learners' use of scope adverbs by analyzing a selection of 200,000 words from the HSK corpus, and her analysis of the corpus data revealed that 45.45% (i.e., 10 out of 22) of the misused 都 *dou* resulted from the missing of the adverb. Similar percentage of misused 都 *dou* resulting from omission was also reported in Liu (2014), whose investigation of CSL learners' misuse of 都 *dou* as a scope adverb via questionnaire revealed that 43.3% (i.e., 203 out of 469) of the erroneous sentences was categorized into missing of the adverb.

In a more thorough study, Yi (2016) retrieved all the misused 都 *dou* as both a scope adverb and a modal particle[1] from the error-tagged HSK corpus. Among the 362 tokens of misused 都 *dou*, 193 out of them resulted from omission, which accounted for 53.31%. In addition, 95.84% (i.e., 185 tokens) of the 193 omission errors occurred when 都 *dou* functioned as a scope adverb, suggesting that the CSL learners often omit都 *dou* as a scope adverb in their writing. To better understand the CSL learners' omission patterns, the researcher further analyzed the 185 tokens of omission into five occasions (see Table 1). Among the five occasions, the CSL learners showed a strong tendency of omission when the subject quantified by 都 *dou* included 每 *mei*, 各 *ge*, 所有 *suoyou*, 全部 *quanbu*, 任何 *renhe*, or classifier reduplication, taking up more than 60% of the 185 tokens. In addition to corpus data, the researcher also retrieved CSL learners' productive knowledge about 都 *dou* via a self-made questionnaire. Analysis of the questionnaire data revealed that the CSL

[1] In Yi (2016), the researcher employed the framework of categorizing the functions of 都 *dou* into scope adverb and modal particle, the latter of which includes functions of (1) expressing a speaker's displeasure, annoyance, or surprise toward an unexpected or unusual state of affairs and (2) expressing that an incident/situation is approaching or having already existed. The second function is categorized as a time adverb in the current study.

Table 1 Distribution of omitted 都 *dou* as a scope adverb in Yi (2016)

Occasion	Token	Percentage
The quantified subject before 都 *dou* includes 每 *mei*, 各 *ge*, 所有 *suoyou*, 全部 *quanbu*, 任何 *renhe*, or classifier reduplication	115	62.16
The quantified subject is a plural NP	24	12.97
The quantified subject is formed by a *wh*-word to express universal meanings	8	4.33
There are 無論 *wulun*, 不論 *bulun*, or 不管 *buguan* in the sentence	16	8.65
Others	22	11.89
Total	**185**	**100**

learners also tended to omit 都 *dou* when the quantified subject (e.g., 每 *mei*, 不管 *buguan*, 不論 *bulun*, etc.) required the presence of 都 *dou* to form a correct sentence.

Although previous studies have showed CSL/CFL learners' marked tendency to omit necessary 都 *dou* in their writing, researchers have different opinions on the omission rates of syntactically obligatory 都 *dou* and syntactically optional 都 *dou.* Zhou and Wang (2007) examined Chinese CSL learners' misuse of 都 *dou* and concluded that obligatory 都 *dou* is easier to acquire than optional 都 *dou.* They argued that CSL learners tend to omit 都 *dou* when it quantifies a definite singular NP or a plural NP, both of which do not require the syntactical presence of 都 *dou.* On the contrary, omission of 都 *dou* is less likely to happen when it quantifies a universal quantified NP, because learners are more aware of the syntactical necessity of 都 *dou* in such condition and will avoid making syntactically ungrammatical sentences. However, Li (2013a, b) empirically investigated English CFL learners' production of 都 *dou* via a controlled elicitation task and discovered that the learners performed better in correctly using syntactically optional 都 *dou.* She suggested that the learners' better performance in the correct use of syntactically optional 都 *dou* resulted from that they could feel the need for 都 *dou* to express the exhaustive or distributive meaning. Furthermore, her brief survey of the textbooks used by her learners disclosed a lack of introducing obligatory 都 *dou* when quantifying universal NPs in the materials. Her research thus shows that CFL learners' acquisition of optional 都 *dou* was better than that of obligatory 都 *dou*, and inadequate explanation of 都 *dou* as a scope adverb in the textbooks might be the cause of learners' misuse. Similar misuse pattern is also observed in Yi's (2016) study. As presented in Table 1, analysis based on the HSK corpus showed that the CSL learners' omission rate of syntactically obligatory 都 *dou* (i.e., 72.02%) was much higher than that of syntactically optional 都 *dou* (i.e., 12.44%), showing that the CSL learners were not aware of the necessity of employing syntactically obligatory 都 *dou* in these conditions. Findings yielded by Zhou and Wang (2007) are contradictory to those from Li (2013a, b) and Yi (2016), and more studies on CSL/CFL learners' omission rates of obligatory and optional 都 *dou* are thus suggested to better reveal learners' misuse pattern.

2.2 *Research on 都 Dou as a Modal Particle and a Time Adverb*

When 都 *dou* functions as a modal particle, it is often used to express a speaker's displeasure, annoyance, or surprise toward an unexpected or unusual state of affairs (Teng, 2019; Zhang, 2005), as illustrated in sentence 2(a) and (b).

(2) a .我都給你這麼多錢了!你還敢說我小氣?
Wo dou gei ni zheme duo qian le! Ni hai gan shuo wo xiaoqi?
'I have given you so much money! How dare you call me stingy?'
b. 我怎麼都不知道他們結婚了!
Wo zenme dou bu zhidao tamen jiehun le!
'I had no idea that they were married!'

In sentence 2(a), 都 *dou* was used to express the speaker's displeasure over the listener's criticism *xiaoqi* "stingy". In sentence 2(b), 都 *dou* was used to express the speaker's surprise of receiving a recent news *jiehun* "marriage". The two 都 *dou*s in both sentences were used as a modal particle to emphasize the speakers' emotions, and omission of the modal particle will cause the sentences to be both syntactically and semantically ungrammatical.

When 都 *dou* functions as a time adverb, its meaning is close to another Chinese adverb 已經 *yijing* "*already*", and a speaker uses 都 *dou* to express that an incident/situation is approaching or has already existed (Zhang, 2005), as illustrated in sentence 3(a) and (b).

(3) a. 都八點了!趕快起床!
Dou badian le! gankuai qichuang!
'It's almost eight o'clock. Get up now!
b. 房子都失火了!你還只顧著找錢包!
Fangzi dou shihuo le! Ni hai zhiguzhe zhao qianbao!
'The house is on fire! How come you just keep looking for your purse?

In sentence 3(a), the speaker used 都 *dou* to emphasize that it was already the time (i.e., *badian* "eight o'clock") that the listener should get up. In sentence 3(b), the speaker used 都 *dou* to express that an emergency situation (i.e., *shihuo* "on fire") is approaching and that the listener should leave the house quickly instead of looking for his/her purse. In both sentences, 都 *dou* was used as a time adverb to tell the listeners that an incident/situation is taking place.

Compared to literature on the misuse of 都 *dou* as a scope adverb, there is relatively scant research on CSL/CFL learners' use of 都 *dou* as a modal particle or a time adverb, and Yang and Yuan's (2010) study via a controlled elicitation task is one of the scant research on CSL/CFL learners' use of 都 *dou* among the three functions. In their study, the researchers designed a set of sentences to investigate 20 CSL learners' productive knowledge of 都 *dou* among the three functions, and they discovered that the learners were less likely to misuse 都 *dou* as a scope adverb. In contrast, the

learners made the most omission errors when 都 *dou* functioned as a time adverb. Their research findings suggest that CSL learners' acquisition of 都 *dou* as a scope adverb is more successful than that of other functions. Yi's (2016) corpus-based analysis on CSL learners' writing, however, revealed different results. Among the 193 omission errors, more than 95% of them resulted from the missing of 都 *dou* as a scope adverb. Omission of 都 *dou* as a model particle or a time adverb accounted for 4.16% (i.e., 8 tokens) only, and most of these errors (i.e., 6 tokens) were found in the sentence pattern 连…都… *lian…dou…* "even". Findings yielded in Yi's suggest that CSL learners have a strong tendency to omit 都 *dou* as a scope adverb as compared to the other two functions, which seems to contradict to those reported in Yang & Yuan. More studies on CSL/CFL learners use of 都 *dou* among different functions are thus required to have a more comprehensive understanding of learners' misuse of these functions.

Because of the conflicting findings regarding CSL/CFL learners' discrepancy in the omission rates of obligatory and optional scope adverb 都 *dou* as well as the limited investigation of learners' misuse of 都 *dou* as a modal particle/time adverb, this study was thus undertaken to examine CSL/CFL learners' omission of 都 *dou* among the three different functions by retrieving and analyzing errors tagged in Chinese Leaner Written Corpus of NTNU. In addition to analysis of corpus data, further examination of how 都 *dou* was introduced and explained in the learners' textbook was also conducted to see if information provided in the textbook was adequate or not.

3 Method

Data analyzed in this study was retrieved from the error-tagged Chinese Leaner Written Corpus of NTNU, a 2.14-million-character learner corpus containing 4,288 take-home essays written by CSL learners from 64 different countries and across five proficiency levels (i.e., A2, B1, B2, C1, and C2 referring to the Common European Framework of Reference for Languages). In that corpus, errors caused by the omission of adverbs are tagged with the label Madv. We hence retrieved all the errors tagged as Madv (token: 2,923) and examined these tokens one by one to identify what these omitted adverbs were. Identification of missing adverbs shows that都 *dou* was the most frequently omitted adverb in the learner corpus (token: 526), accounting for 18% of all the omitted adverbs and making itself a good subject for detailed investigation. After all the instances of omitted 都 *dou* were generated, the researchers examined what functions (i.e., scope adverb, modal particle, and time adverb) these omitted 都 *dou* served by context and counted the tokens. Distribution of omitted 都 *dou* across the three functions will be presented and discussed in the next section.

4 Results and Discussion

As shown in Table 2, there were 461 tokens of omitted 都 *dou* as a scope adverb, taking up close to 90% of the omitted 都 *dou*. By contrast, the percentage of omitted 都 *dou* as a modal particle was only 12.36%, and there was even zero occurrence of omitted 都 *dou* as a time adverb in the corpus. The distribution of omitted 都 *dou* among the three functions is pretty similar to that yielded in Yi (2016), who discovered that more than 95% of omitted 都 *dou* in the HSK corpus functioned as a scope adverb. Findings yielded in the current study and in Yi's are in accordance and suggest that CSL/CFL learners tend to omit 都 *dou* as a scope adverb much more often than other functions in their writing.

4.1 Learners' Omission of 都 Dou *as a Scope Adverb*

To further analyze how 都 *dou* as a scope adverb was omitted in the learner corpus, we employed Liu et al.'s (1996) framework to categorize the 461 tokens, and Table 3 shows the distribution of these tokens among the five conditions.

When serving as a scope adverb, the omission of 都 *dou* occurred more often in Condition 1, which accounted for more than 56% of the 461 tokens. Among the

Table 2 Distribution of omitted 都 *dou* among the three functions

Function	Token	Percentage
Scope Adverb	461	87.64
Modal Particle	65	12.36
Time Adverb	0	0
Total	**526**	**100**

Table 3 Distribution of omitted 都 *dou* as a scope adverb among the five conditions in Liu et al. (1996)

Condition	Token	Percentage
1. The quantified subject includes 每 *mei*, 所有的 *suoyoude*, 一切 *yiqie*, and 任何 *renhe*, or there are 隨時 *suishi* or 到處 *daochu* in the sentence	261	56.62
2. The quantified subject is a plural NP	164	35.57
3. The quantified subject is formed by a *wh*-word to express universal meanings	22	4.77
4. An affirmative sentence which includes 無論 *wulun*, 不論 *bulun*, or 不管 *buguan*	14	3.04
5. An interrogative sentence which is formed by interrogative pronouns like 誰 *shei*, 什麼 *shenme*, 哪兒 *naer*, or 哪 *na* + quantifier	0	0
Total	**461**	**100**

261 tokens of omitted 都 *dou* in Condition 1, 17 out of them occurred when the quantified subject included 所有的 *suoyoude*, 21 out of them occurred when the quantified subject included 隨時 *suishi* or 到處 *daochu*, three out of them occurred when the quantified subject included 任何 *renhe*, and one out of them occurred when the quantified subject included 一切 *yiqie*. When the quantified subject included 每 *mei*, the percentage of omission was the highest among all the others, taking up 84% (219 tokens) of the 261 tokens, as illustrated in concordance lines (4)–(6).

(4) *而且我也越來越喜歡學中文了, 所以每天我高興得不得了。

*Erqie wo ye yue lai yue xihuan xue zhongwenle, suoyi meitian wo *(dou) gaoxing de budele.*

'And I like learning Chinese more and more, so I am very happy every day.'

(5) *每次選擇的時候, 有好悶的感覺

*Mei ci xuanze de shihou, *(dou) you hao men de ganjue.*

'I feel so stuffy every time when I have to make choice.

(6) *每個世紀, 日本會發生幾次海嘯

*Mei ge shiji, riben *(dou) hui fasheng ji ci haixiao.*

'Every century, Japan will have several tsunamis.'

The word 每 *mei* is used as a determiner before an NP to "reinforce the sense of 'no exception'" (Teng, 2019, p. 205), and its occurrence in a sentence should always include 都 *dou* (Liu et al., 1996; Lu, 1980; Teng, 2019). The high percentage of omitting 都 *dou* in sentences with 每 *mei* in the learner corpus might indicate the learners' lack of awareness of the obligatory use of 都 *dou* in this condition.

The learners were also frequently found to omit the use of 都 *dou* when the quantified subject was a plural NP, as illustrated in concordance lines (7) and (8).

(7) *如果我遇到困難他們很願意幫助我。

*Ruguo wo yu dao kunnan tamen *(dou) hen yuanyi bangzhu wo.*

'They are willing to help me if I encounter any difficulties.'

(8) *台灣在日本和德國最重要的三個城市有代表。

*Taiwan zai riben he deguo zui zhongyao de san ge chengshi *(dou) you daibiao.*

'Taiwan has representatives in three of the most important cities in both Japan and German.'

When quantifying a plural NP, 都 *dou* is syntactically optional to express the exhaustive sense of the NP. The learners' omission of 都 *dou* might result from the fact that the use of 都 *dou* in this condition is not syntactically mandatory and they hence omitted it in these sentences.

The use of 都 *dou* in Condition 3 (i.e., a universally quantified NP formed by a *wh-* word) and Condition 4 (i.e., an affirmative sentence including 無論 *wulun*, 不論 *bulun*, or 不管 *buguan*), however, is syntactically obligatory, yet the learners still occasionally omitted 都 *dou* in these conditions. This is illustrated in concordance line (9)–(11).

(9) *我到哪裏去, 哪裏可以吃到很多好吃的東西。
*Wo dao nali qu, nali *(dou) keyi chi dao henduo hao chi de dongxi.*
'Wherever I go, I can eat much delicious food.'

(10) *現在無論別人說什麼我懂得差不多了。
*Xianzai wulun bieren shuo shenme wo *(dou) dongde chabuduole.*
'Now I understand almost everything that others tell me.'

(11) *女人不管多麼能幹不能取代男人的地位。
*Nuren buguan duome nenggan *(dou) buneng qudai nanren di diwei.*
'Women can't replace men no matter how capable they are.'

When co-occurring with 都 *dou* in the same sentence, the presence of a *wh*-word does not indicate a questions but expresses a distributive meaning. As illustrated in concordance line (9), the *wh*-word 哪裏 *nali* actually expresses the meaning of "everywhere" or "all places". Without the presence of 都 *dou* in concordance line (9), the sentence would be interpreted as a question. The use of 都 *dou* in sentences like concordance line (9) is thus mandatory. Regarding Condition 4, the presence of 都 *dou* is also obligatory when it is preceded before conjunctions like 無論 *wulun*, 不論 *bulun*, and 不管 *buguan* to express the meaning "in spite of various circumstances, the following fact [following 都 *dou*] remains unaffected" (p. 284, Teng, 2019). Omission of 都 *dou* in Condition 4 would fail to express this all-inclusive meaning and is hence syntactically incorrect.

Findings of the CSL learners' omission of 都 *dou* as a scope adverb among different conditions mostly echo with those yielded in Yi (2016). In both of the studies, learners were found to omit 都 *dou* much more often when the subject quantified by 都 *dou* included 每 *mei*, 所有(的) *suoyou(de)*, 一切 *yiqie*, 任何 *renhe*, etc., which took up 56.62% in this study and 62.16% in Yi's respectively and ranked the first in both studies. The second common condition of 都 *dou*'s omission occurs when the quantified subject was a plural noun, which accounted for 35.57% and 12.97% in our study and in Yi's respectively. Omission rates of the top two conditions constituted more than 75% of the total errors in both studies, whereas the other three conditions together amounted to no more than 25%. The similar distribution pattern of missing 都 *dou* as a scope adverb among different conditions in both studies suggests the following. First, CSL/CFL learners' omission of mandatory 都 *dou* mainly occurs when the quantified subject includes those formed by a universal quantifier (e.g., 每 *mei*, 所有 *suoyou*) modifying an NP, those expressing universal meaning (e.g., 一切 *yiqie*) and those being plural nouns, showing that learners are more likely to ignore the rule of using 都 *dou* when producing these structures. In addition, both studies found that CSL/CFL learners tend to omit 都 *dou* much more often when it quantifies a subject formed by a universal quantifier or expressing universal meaning, which contradicts to Zhou and Wang's statement (2007) that CSL learners tend to omit 都 dou when it quantifies a definite singular NP or a plural NP. Findings revealed in our study and in Yi's (2016) hence show that CSL/CFL learners' mastery of syntactically obligatory 都 *dou* might be less successful than that of syntactically optional 都 *dou*, which might result from learners' ignorance of the co-occurrence of these universal

quantifiers and the scope adverb, and more efforts should thus be made to teach learners about the necessity of using 都 *dou* in these structures (Yi, 2016).

4.2 *Learners' Omission 都* Dou *as a Modal Particle and a Time Adverb*

When used as a modal particle, there were 65 tokens of omitted 都 *dou*, and most of them occurred in the sentence pattern 連…都… *lian…dou…* "even", as illustrated in concordancing line (12)–(14).

(12) *連海洋生物家不知道遠海的深處隱藏著何種奧秘。

*Lian haiyang shengwu xue jia *(dou) bu zhidao yuanhai de shen chu yincangzhe he zhong aomi.*

'Even marine biologists don't know what mystery is hidden in the deep sea.'

(13) *這個地方很熱鬧, 連半夜找得到吃的地方。

*Zhege difang hen renao, lian banye *(dou) zhao dedao chi di difang.*

'This place is very lively, and you can find a place to eat even in the middle of the night.'

(14) *有時候可能連手機沒信號。

You shihou keneng lian shouji mei xinhao.

'My cellphone sometimes even had no reception at all.'

In this sentence pattern, "the preposition 連 *lian* introduces the focus of a sentence, highlighting a noun against all other related nouns in a given context" (Teng, 2019, p.190). The function of 都 *dou* in this sentence pattern is to emphasize the unusualness or noteworthiness of the focus after 連 *lian* (Ma, 1983), and Cao (2005) also pointed out that *dou* plays a more important role than 連 *lian* in this sentence pattern. She argued that omission of 都 *dou* in 連…都… *lian…dou…* would cause a difference in meaning, even an ungrammatical sentence, while omission of 連 *lian* might not; however, in most textbooks, focus is often placed on explaining and discussing the function of *lien* other than 都 *dou*. The high percentage of omitted 都 *dou* in the pattern 連…都… *lian…dou…* is also reported in Yi (2016), in which 75% of the omitted 都 *dou* as a modal particle occurred in this sentence pattern. The high ratios of 都 *dou*'s omission in the sentence pattern in both studies hence suggest that CSL/CFL learners' have not yet fully acquired the sentence pattern and that teachers as well as material writers should make more efforts to help learners gain a comprehensive acquisition of the structure.

When used as a time adverb to express the meaning of "already", there was zero token of omission found in the learner corpus. Although no instance of 都 *dou*'s omission as a time adverb was identified in the learner corpus, this does not mean that the CSL learners had successfully acquired its use. Instead, the zero occurrence might result from the CSL learners' avoidance of using 都 *dou* as a time adverb. In Yi's (2016) analysis of the self-made questionnaire, the researcher discovered that

the CSL learners in the study were not familiar with 都 *dou*'s function to express the meaning of "*already*". When deciding the grammaticality of sentences formed with misused 都 *dou* as a time adverb, low-level learners often misjudged incorrect sentences as correct ones, while higher-level learners would misjudge the presence of 都 *dou* as ungrammatical and cross out the adverb in the sentences. Yi's findings thus show CSL learners' inadequate knowledge about the use of 都 *dou* as a time adverb, and further suggest that the few tokens of omitted 都 *dou* as a time adverb might be caused by learners' avoidance of using it.

In general, the learners in our study omitted 都 *dou* as a scope adverb much more often than they omitted 都 *dou* as a modal particle or a time adverb, which corroborates with findings of previous studies. Our results also echo with other researchers' findings that CSL/CFL learners tend to omit syntactically obligatory 都 *dou* more often than syntactically optional 都 *dou*, and omission is highly frequent when the subject quantified by 都 *dou* includes a universal quantifier (e.g., 每 mei). This frequent omission might be due to CSL/CFL learners' unawareness of the co-occurrence of a universal quantifier and the scope adverb 都 *dou*, which is suggested to be rightly emphasized in both the teaching and learning of these universal quantifiers. Although findings in the current study and the previous ones uncover the low omission rates of 都 *dou* as a modal particle and/or a time adverb, this should not be concluded with the statement that learners' mastery of these two functions is better than that of a scope adverb. Instead, the striking difference in error percentages might arise from their underuse of these two functions, which thus lowers their chances of omission. To better understand the reasons for the learners' discrepancy in omitting 都 *dou* among the three functions, we further examined how 都 *dou* was presented in the teaching materials the learners used and whether its presentation influence the way the learners employed the adverb for different functions.

4.3 Examination on the Presentation of Dou *in Teaching Materials*

Data included in Chinese Leaner Written Corpus of NTNU contains essays written by CSL learners at the Mandarin Training Center of NTNU between 2010 and 2012, whose textbook was the second edition of *Practical Audio-Visual Chinese*. We therefore examined how 都 *dou* was presented in the five volumes of *Practical Audio-Visual Chinese* (2nd edition), and Table 4 shows the results.

Examination on the presentation of 都 *dou* in the learners' textbook revealed that 都 *dou* was mainly introduced as a scope adverb, the meaning of which was defined as *all* or *both* in English. However, the rule of using obligatory 都 *dou* when the quantified subject includes 每 *mei* was not mentioned throughout the five volumes of the textbook, neither was the occurrence of the sentence pattern 每-… 都… *mei*-…*dou*…. The meanings of 都 *dou* and 每 *mei* were defined as "all/both" (Vol.1, L3) and "every" (Vol. 1, L11) in the textbook respectively. Sun (2001) has

Table 4 Presentations of 都 *dou* in *Practical Audio-Visual Chinese*

Vol./Lesson	Page	Usage/Meaning explanation	Example sentence
V1/L3	38	都(*dou*) all; both	他們**都**很忙。 *Tamen dou hen mang* 'They are very busy.'
V2/L3	63	If one wants to express an inclusive such as "everywhere", "everyone", and "everything", or an exclusive like "nowhere", "no one" and "nothing", then he must use a question word in conjunction with the adverb 都 (*dou*). In negative expressions, the adverb 也 (*ye*) can be used in place of 都 (*dou*)	他什麼**都**知道。 *Ta shenme dou zhidao* 'He knows everything.'
V3/L2	53	連…都/也… (*lian…dou/ye…*) Even…	連半夜**都**找得到地方吃。 *Lian banye dou zhao dedao difang chi* 'I can find a place to eat even in the middle of the night.'
B3/L4	107	除了…以外/之外, 都… (*chule…yiwai/zhiwai, dou…*) Except for/other than …all the others…	除了海邊有一些平原以外, 中部**都**是山。 *Chule haibian you yixie pingyuan yiwai, zhongbu dou shi shan* 'Except for some plains along the coast, the central part is full of mountains.'
V3/L7	200	不管/不論/無論…都… (*buguan/bulun/wulun…dou…*) Regardless of whether…(all), no matter whether …(all)	以前台灣的大學, 不管公立的私立的, 學費**都**沒有你們這裡這麼高。 *Yiqian taiwan de daxue, buguan gongli de sili de, xuefei dou meiyou nimen zheli zheme gao* 'In the past, the tuition fees of universities in Taiwan, whether public or private, were not as high as yours.'
V3/L13	367	什麼都…, 就是… (*shenme dou…, jiushi…*) Everything is…, expect…	聖誕節什麼**都**好, 就是買禮物、寄聖誕卡太麻煩。 *Shengdan jie shenme dou hao, jiushi mai liwu, ji shengdanka tai mafan* 'Christmas is good, but buying gifts and sending Christmas cards is too troublesome.'
V5/L2	22	凡是…都… (*fanshi…dou…*) Every … (all)…; All … are (all)…	凡是到過歐洲旅遊的人**都**讚美歐洲的風景。 *Fanshi daoguo ouzhou luyou de ren dou zanmei ouzhou de fengjing* 'Everyone who has traveled to Europe praises the European landscape.'

pointed out that simply presenting the meanings of 都 *dou* and 每 *mei* as "all/both" and "every" in textbooks would oftentimes mislead CSL/CFL learners, especially English CSL/CFL learners, over 都 *dou*'s function of emphasizing the inclusive meaning when in conjunction with 每 *mei*. The simplified English translations of 都 *dou* and 每 *mei* and the lack of example sentences of 每-…都… *mei*-…*dou*… to illustrate the sense of "no exception" might be the reasons for the learners' frequent omission of obligatory 都 *dou* when using 每 *mei* in their writing. As for optional 都 *dou* in Condition 2, there was in fact one example sentence that formed with a plural NP (i.e., 他們 *tamen*) in volume two of the textbook; however, no further explanation of the semantical function of 都 *dou* to express an inclusive meaning was offered. This might hence cause the learners to omit 都 *dou* when the quantified subject was a plural NP in their writing so often.

都 *Dou* as a modal particle was introduced in the sentence pattern 連…都/也… *lian*…*dou/ye*… only once among the five volumes, and explanation of the sentence pattern was "the speaker thinks that the situation mentioned after 連(*lian*) is unusual or noteworthy and thus uses this pattern for emphasis. N, SV, V, VO, S-V of a simple sentence all can be place after 連(*lian*)". The explanation focused on the use of 連 *lian* only, and no effort was made to explain the function of 都 *dou*. This might result to the learners' ignorance of using 都 *dou* in this sentence pattern and thus caused the omission of this adverb.

In contrast, the function of 都 *dou* as a time adverb did not appear at all among the five volumes. The zero occurrence of 都 *dou* as a time adverb in the textbook could explain why there was also zero token of omitted 都 *dou* as a time adverb in the learner corpus. That is, since the learners merely encountered this function in their textbook, it was thus unlikely for them to use 都 *dou* to express the meaning of "already" and hence underuse the adverb in their writing.

5 Conclusion, Pedagogical Implications, and Suggestion for Future Research

The current study was conducted to investigate how and why CSL leaners omit the adverb 都 *dou* based on the error-tagged Chinese Leaner Written Corpus of NTNU. Among the 526 tokens of omitted 都 *dou*, more than 87% of the total errors occurred when 都 *dou* was used as a scope adverb, while tokens of omitted 都 *dou* as a modal adverb and a time adverb were 65 and 0 respectively. Examination of how 都 *dou* was introduced in the learners' textbook reveals the following causes of omission. Firstly, the learners were influenced by the English translation (i.e., all, both) of 都 *dou* in the textbook to misuse 都 *dou* in their writing. Secondly, the learners' low omission rates of 都 *dou* as a modal particle or a time adverb might be largely due to their underuse of the two usages, since these functions were rarely introduced in the textbook. To better improve CSL learners' productive knowledge of 都 *dou*, suggestions for material writing are offered here. First, all of the three

functions of 都 *dou* should be introduced in the textbook in the order of scope adverb, modal adverb, and time adverb, so that CSL learners could know that the adverb 都 *dou* can be used for different purposes. Second, when introducing 都 *dou* as a scope adverb, efforts should be made to specifically identify its function to express an exclusive/distributive meaning and simultaneously introduce the universal quantifiers that co-occur frequently with 都 *dou* (e.g., 每 *mei*, 所有 *suoyou*, 任何 *renhe*, etc.).

Although this study showed how CSL learners omitted the adverb 都 *dou* in their writing and identified the possible causes of learners' omission of 都 *dou*, there are still some limitations of the current study. The first limitation that should be considered is the data analyzed in the learner corpus. As described in the previous section, data in the error-tagged Chinese Leaner Written Corpus of NTNU consisted of CSL learners' writing assignments, which were produced in a context that the learners could consult different resources in their writing. Since contextual difference of datasets might influence learners' productive output, future studies are suggested to investigate CSL/CFL learners' omission of 都 *dou* in different contexts (e.g., exam scripts). In addition, the current study could only target on investigating the omission of 都 *dou* due to the length limit of the article. There are, however, other adverbs (i.e., 很 *hen* and 就 *jiu*) that were also frequently omitted in the learner corpus. To better understand the reason why CSL/CFL learners omitted these adverbs in their writing, further research on the omission of these adverbs is also suggested.

References

Cao, X. (2005). 再议"连……都/也……"句式 Zaiyi "lian…dou/ye…" ju shi. [A further discussion on the sentence pattern *lian…dou/ye*….]. *Linguistic Researches, 1*, 17–20.

Chao, Y.-R. (1968). *Grammar of spoken Chinese.* University of California Press.

Cheng, L. L. S. (1995). On *dou*-quantification. *Journal of East Asian Linguistics, 4*(3), 197–234.

Duan, S.-Y. (2008). 现代汉语副词研究综述 Xiandai hanyu fuci yanjiu zongshu. [A study of Chinese adverbs.]. *Journal of Baoshan Teachers' College*, *27*(4), 66–69.

Gao, S.-Q. (2011). 多义副词"还"的语法化顺序和习得顺序 Duo yi fuci "hai" de yufa hua shunxu he xi de shunxu. [On the grammaticalization order and acquisition order of multi-meaning adverb *hai*.]. *TCSOL Studies, 2*, 39–45.

He, S.-B. (2006). 汉语作为第二语言教学中的副词研究综述 Hanyu zuowei di er yuyan jiaoxue zhong de fuci yanjiu zongshu. [A research on adverbs in Chinese as a second language.]. *Journal of Suzhou Education Institute*, *9*(1), 71–75.

Jiang, X. (2013). 日本留学生汉语副词"还"的习得考察——基于HSK动态作文语料库的研究 Riben liuxuesheng hanyu fuci 'hai' de xide kaocha——jiyu HSK dongtai zuowen yuliaoku de yanjiu. [A study of the Japanese students' acquisition of Chinese adverb *haii*.]. *Overseas Chinese Education, 1*, 71–78.

Li, C. & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, H. (2013a). 留学生总括类范围副词习得研究及偏误分析 Liuxuesheng zonggua lei fanwei fuci xide yanjiu ji pian wu fenxi [The acquisition research and error analysis of abroad student on general scope adverb] (Unpublished master's thesis) Northwest Normal University.

Li, Y. (2013b). An empirical study on the production of *dou*: Is native-like performance attainable? *Journal of Chinese Language Teaching, 10*(3), 121–162.

Lin, J. W. (1998). Distributivity in Chinese and its implications. *Natural Language Semantics, 6*(2), 201–243.

Liu, W.-Q. (2019). “都”从总括义到惊讶义的演变 *Dou* cong zonggua yi dao jingya yi de yanbian. [Semantic analysis of *dou* from generalized meaning to surprised meaning.]. *Journal of Taiyuan Normal University*, *18*(2), 56–61.

Liu, Y., Pan, W., & Gu, W. (1996). *Modern Chinese grammar for teachers of Chinese as a second language and advanced learners of modern Chinese*. Shi Da Shu Yuan.

Liu, Y.-N. (2014). 留学生习得范围副词“都”的偏误分析 Liuxuesheng xide fanwei fuci dou de pian wu fenxi [The error analysis of international students learning scope adverb “dou”] (Unpublished master’s thesis) Lanzhou University.

Lu, S.-H. (1980). *Xiandai Hanyu Babai Ci.現代漢語八百詞* [*Eight Hundred Words of the Contemporary Chinese.*]. Shangwu Yin Shu Guan.

Luo, J.-L. (2016). 留学生习得范围副词“都”和“全”的偏误分析 Liuxuesheng xide fanwei fuci dou he quan de pian wu fenxi [An error analysis of scope adverbs dou and quan by international students]. (Unpublished master’s thesis). Heilongjiang University.

Ma, Z. (1983). 关于“都/全”所总括的对象的位置 Guanyu “dou/quan” suo zonggua de duixiang de weizhi. [On the position of the NP quantified by *dou* or *quan.*]. *Chinese Language Learning, 1*, 27–34.

Rong, H. (2008). 韩国留学生程度副词使用偏误分析 Hanguo liuxuesheng chengdu fuci shiyong pian wu fenxi. [Error analysis of degree adverbs used by South Korean students.]. *Journal of Jiangxi Institute of Education*, *29*(3), 102–105.

Sun, C.-Y. (2001). *A semantic and pedagogical study of DOU in Mandarin Chinese* [Unpublished master dissertation]. National Taiwan Normal University.

Tan, P. (2012). 汉语“已然”类时间副词的偏误分析 Hanyu yiran lei shijian fuci de pian wu fenxi. [An error analysis of time adverbs that express realis mood in Modern Chinese.]. *Survey of Education, 3*, 191–195.

Teng, S.-H. (2019). *An A to Z grammar for Chinese language learners.* Linkingbooks.

Tsai, W. T. D. (2014). *On economizing the theory of A-bar dependencies*. Routledge.

Yang, Y., & Yuan, W. (2010). 越南留学生习得“都”的偏误分析 Yuenan liuxuesheng xi de “dou” de pian wu fenxi. [An error analysis of Vietnamese learners’ use of the scope adverb *dou.*]. *Modern Chinese, 4*, 136–138.

Yi, L. (2016). 留学生副词“都”的习得调查分析及教学设计 Liuxuesheng fuci dou de xide diaocha fenxi ji jiaoxue sheji [Analysis of acquisition of adverb “dou” for foreign students and the teaching design for “dou”] (Unpublished master’s thesis) Hunan Normal University.

Yuan, B. (2009). Non-permanent representational deficit and apparent target-likeness in second language. In N. Snape, Y.-K. Leung, & M. S. Smith (Eds.), *Representational deficits in SLA: Studies in honor of Roger Hawkins* (pp. 79–103). John Benjamins Publishing Company.

Zang, C.-Y. (2010). 近十年来汉语作为第二语言教学中副词运用偏误研究综述 Jin shi nianlai hanyu zuowei di er yuyan jiaoxue zhong fuci yunyong pian wu yanjiu zongshu. [On the errors of Chinese adverb usages as the second language.]. *Journal of Liuzhou Vocational & Technical College*, *10*(2), 44–49.

Zhang, J.-B. (2007). 程度副词“很”的有关偏误分析 Chengdu fuci “hen” de youguan pian wu fenxi. [An error analysis of the adverb of degree *hen.*]. *Overseas Chinese Education, 2*, 32–39.

Zhang, Y. (2005). 副词“都”的语义及语用功能分析 Fuci “dou” de yuyi ji yu yong gongneng fenxi. [An analysis of the semantic and pragmatic functions of the adverb *dou.*]. *Theory Circle, 11*, 186–187.

Zhang, Y.-S. (2003). 论现代汉语的范围副词 Lun xiandai hanyu de fanwei fuci. [On range adverbs of Mandarin Chinese.]. *Journal of Shanghai Teachers University, 5*, 392–398.

Zheng, Y.-Q. (2006). 中介语中程度副词的使用情况分析 Zhongjie yu zhong chengdu fuci de shiyong qingkuang fenxi. [Analysis of degree adverbs usage of foreign learners.]. *Chinese Language Learning, 6*, 66–72.

Zhou, X.-B., & Wang, Y. (2007). An analysis of the grammatical errors concerning *dou* as an adverb of scope. *Chinese Language Learning, 1*, 71–76.

Acquisition of the Chinese Indefinite Determiner "One + Classifier" and English Articles in Two-Way Learner Corpora

Zhang Zheng, Laurence Newbery-Payton, and Sho Fukuda

Abstract This paper presents findings concerning use of classifiers and articles in learner corpora and the effect of learners' native languages on their acquisition of a second language. First, we use data from the Learners' Corpus of Chinese, an error-tagged two-way learner corpus of intermediate and advanced learners' written production, developed by Tokyo University of Foreign Studies (TUFS) in collaboration with National Taiwan Normal University. The corpus data reveals that 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. Third, analysis of the International Corpus of Japanese as a Second Language reveals that Chinese L1 learners of Japanese use the "one + classifier" structure more frequently than native speakers. Similarities and differences between L1 and L2 can supersede ostensible typological similarities, such as the classification of both Chinese and Japanese as classifier languages.

Z. Zheng (✉) · L. Newbery-Payton · S. Fukuda
Institute of Global Studies, Tokyo University of Foreign Studies, Evergreen 201, 3-53-16 Momijigaoka, Fuchu City, Tokyo 183-0004, Japan
e-mail: zhangzheng.apple@icloud.com

L. Newbery-Payton
Institute of Global Studies, Tokyo University of Foreign Studies, Flat 404, Nakagawa 2-9-9, Tsuzuki Ward, Yokohama, Kanagawa Prefecture 224-0001, Japan

S. Fukuda
The University of Toyama, 1535-21, Shimookubo, Toyama-shi, Toyama, Japan

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_10

1 Introduction

This study investigates the acquisition of the "one + classifier" structure by Japanese native learners of Chinese (henceforth, "JLC") and compares it with English native learners of Chinese (henceforth, "ELC"). It reveals that underuse of the "one + classifier" structure preceding nouns is more prevalent in written essays by JLC. Typical errors of omission are given below.[1]

(1) 日本 是 <φ→一个> 岛国， 四面 环海， 所以 海洋 资源 很 丰富。
Riběn shì <φ→yí ge> dǎoguó， sìmiàn huánhǎi，suǒyǐ hǎiyáng zīyuán hěn fēngfù.
(2013_146_TUFS_CH_059)
Japan is an island nation surrounded by the sea, so is rich in marine resources.

(2) <φ→一个> 团体 有 这样 的 信赖 关系 的话， 做 什么
<φ→Yí ge> tuántǐ yǒu zhèyàng de xìnlài guānxi de huà， zuò shénme
工作 都 会 成功 的。
gōngzuò dōu huì chénggōng de. (2013_233_TUFS_CH_051)
If an organization has this kind of relationship of trust, it will succeed no matter what it does.

(3) 我 妈妈 在 菲律宾 买 了 <φ→一块> 土地，我 想 帮 她 盖
Wǒ māma zài Fēilǜbīn mǎi le <φ→yíkuài> tǔdì， wǒ xiǎng bāng tā gài
<φ→一栋> 美丽 的 住房。
<φ→yí dòng> měilì de zhùfáng. (2014_057_TUFS_CH_038)
My mother bought a plot of land in the Philippines. I'd like to help her build a beautiful house there.

(4) 机场 的 旁边 有 <φ→一个> 公园。
Jīchǎng de pángbiān yǒu <φ→yí ge> gōngyuán. (2014_134_TUFS_CH_099)
There is a park by the airport.

There is a park by the airport.

English native speakers, on the other hand, are less likely to omit "one + classifier", instead displaying a slight tendency to overuse the structure, as in the following examples.

(5) 这 是 我 第一次 看到 <一座→φ> 很 大 的 雪山。 (E-B1-0140)
Zhè shì wǒ dì yī cì kàndào <yí zuò→φ> hěn dà de xuěshān.
This is the first time I have seen a tall snow-topped mountain.

(6) 恭喜！恭喜！ 我 听说 你 已经 找到 <一份→φ> 工作 了。 (E-A2-0008)
Gōngxǐ! Gōngxǐ! Wǒ tīngshuō nǐ yǐjīng zhǎodào <yí fèn→φ> gōngzuò le.
Congratulations! I heard you've already found a job.

[1] Errors are shown in the form " < error → correction > ".

Table 1 Three types of "one + Classifier"

[+referential]		[-referential]
[+specific]	[-specific]	
(7) a. 他 去年 买 了 Tā qùnián mǎi le 一 幢 房子。 yí zhuàng fángzi b. 彼は 去年 Kare-wa kyo-nen he-TOP last year 家を （1 軒） ie-o (ik-ken) house-ACC one-CL 買った。 kat-ta buy-PST c. He bought a house last year	(8) a. 他 想 买 一 幢 Tā xiǎng mǎi yí zhuàng 房子，什么 房子 fángzi，shénme fángzi 都行。 dōu xíng b. 彼は 家を Kare-wa ie-o he-TOP house-ACC （1 軒） 買いたい (ik-ken) kai-tai one-CL buy-AUX と 思っている。 to omot-teiru COMP think-DUR どのような家でも 良い。 donoyona ie-demo yoi any house-CONJ be fine c. He wants to buy a house, any house is fine	(9) a. 他 是 一 个 买卖人。 Tā shì yí ge mǎimàirén b. 彼は （??1 人の） Kare-wa (hito-ri-no) he-TOP one-CL-POSS ビジネスマン です。 bijinesu-man desu business person COP c. He is a business person

The uses of "one + classifier" can be classified into three types, based on the features of referentiality and specificity.[2] Table 1 shows examples of each alongside equivalent Japanese and English sentences. Example (7a) is a [+referential, + specific] usage. Example (8a) shows a [+referential, –specific] usage. Example (9a) is a non-referential usage and can be considered the most grammaticalized of the uses of "one + classifier".

The Japanese sentences in (7b) and (8b) may include the form "*ik-ken*", similar to the Chinese "one + classifier". However, whereas the classifier is typically required in Chinese, this is not obligatory in Japanese. The function of "one + classifier" is less grammaticalized in Japanese than it is in Chinese; the former is used mainly to express number, which restricts the scope for positive transfer in cases where "one + classifier" is not used to express purely numeral information. The non-referential use of "one + classifier", which is not expressed with a similar form in Japanese, is expected to be particularly difficult for L1 Japanese learners to acquire.

[2] All three types in Table 1 include the feature [-definite], similar to the indefinite article in English.

In contrast, the use of the "one + classifier" structure to mark an indefinite noun phrase is similar to typical uses of the English indefinite article. There is therefore the possibility for positive transfer from L1 to occur and so we predict that the "one + classifier" is relatively easy for ELC to acquire.

To summarize, we hypothesize that ELC and JLC will exhibit contrasting trends in the use of the "one + classifier" and that these differences can be plausibly explained through consideration of L1 characteristics.

In the remainder of this section, we will give an overview of functional equivalents to articles in Chinese. Chen (2004) claims that in Chinese, the demonstratives 这 zhè and 那 nà have developed definite article-like uses such as in (10a). In (10a), by using "the house"/"这栋房子", the speaker assumes that the listener knows which house George bought. The "one + classifier" structure has also undergone some degree of grammaticalization and functions in a similar way to the English indefinite article in some cases (10b, as well as the examples in Table 1). In (10b), by using "a house" / "一栋房子", the speaker assumes that the house George bought cannot be identified as a particular house and that the listener does not know which one he or she is talking about.

(10) a. George finally bought the house.
乔治 终于 买 了 这 栋 房子。
Qiáozhì zhōngyú mǎi le zhè dòng fángzi.
b. George finally bought a house.
乔 治 终于 买 了 一 栋 房子。
Qiáozhì zhōngyú mǎi le yí dòng fángzi.

(Chen 2004 p.1131)

According to Chen (2004), "one + classifier" appears in five uses equivalent to the indefinite article in English (10a)–(10e).

(11) a. numeral
这 件 事 不 难 办，我 只 要 一 个 钟头 就 够 了。
Zhè jiàn shì bù nán bàn, wǒ zhǐ yào yí ge zhōngtou jiù gòu le.
this CL thing not hard do I only need one CL hour then enough CRS
'This is not hard. I only need one/an hour for it.'

b. presentative use
一 架 飞机 从 我们 头 上 飞 了 过去。
Yí jià fēijī cóng wǒmen tóu shang fēi le guòqu.
one CL airplane from we head above fly PFV go
'An airplane flew over us.'

c. nonidentifiable specific reference
他 去年 买 了 （一） 幢 房子。
Tā qùnián mǎi le (yí) zhuàng fángzi.
he last year buy PFV one CL house
'He bought a house last year.'

d. nonidentifiable specific reference
他想 买（一） 幢 房子，什么 房子 都 行。
Tā xiǎng mǎi (yí) zhuàng fángzi，shénme fángzi dōu xíng.
he want buy one CL house any house all do
'He wants to buy a house; any house will do.'

e. nonreferential use
他 是（一）个 买卖人。
Tā shì (yí) ge mǎimài rén.
he be one CL businessman
'He is a businessman.' (Chen 2003:1171)

In the analysis below, we focus on the use of "one + classifier", its similarities with English articles, and the consequences for L2 Chinese acquisition by JLC and ELC.

2 The Present Study

2.1 Corpus Data

This paper analyzes the acquisition of the "one + classifier" structure by using the corpus of written Chinese collected by Tokyo University of Foreign Studies (TUFS corpus: https://corpus.icjs.jp). The ELC written data is provided by Taiwan Normal University and consists of essays written as part of the Test of Chinese as a Foreign Language (TOCFL). The composition of the corpus used in this paper and its size are shown in Table 2.

2.2 Methodology

Instances of the "one + classifier" structure were manually extracted from the corpus. Correct and incorrect example sentences were then distinguished and categorized based on the type of error, the linguistic context in which the error was produced, and the learner's proficiency level. The following sections will discuss possible causes of learners' under- and overuse of the "one + classifier" structure.

Table 2 Composition and size of the corpora

Subcorpus	Proficiency level	Number of essays	Number of characters	Number of words
JLC	CEFR-A2	255	110,768	66,309
	CEFR-B1	96	37,774	23,791
	CEFR-B2	56	23,225	14,938
	Total	407	171,767	105,038
ELC	CEFR-A2	225	31,216	21,985
	CEFR-B1	287	119,032	81,221
	CEFR-B2	122	61,357	36,691
	Total	634	211,605	139,897

In Sect. 3, we provide quantitative and qualitative analysis of errors in the use of classifiers by Chinese L2 learners and show how L1 appears to affect L2 acquisition. We then provide supporting evidence in the form of case studies of English L2 article use and Japanese L2 "one + classifier" use in Sects. 4 and 5 respectively.

3 Results and Discussion

This section presents results of analyzes of error trends in the use of "one + classifier" in L2 Chinese. Quantitative and qualitative findings for JLC are reported in Sects. 3.1 and 3.2 respectively. Section 3.3 briefly highlights characteristic errors by ELC. These errors additionally display parallels with the L2 English article use that we cover in Sect. 4.

3.1 Quantitative Analysis of the Use of "One + Classifier" by Japanese and English L1 Learners of Chinese

Instances of "one + classifier" produced by JLC and ELC were extracted then compared using adjusted frequencies (per 10,000 words). The results are shown in Table 3.

Table 3 Comparison of the output of "one + classifier" by Japanese/English native learners

Corpora	Frequency of occurrence	Adjusted frequency (per 10,000 words)
JLC	277	17.78
ELC	1046	74.77

($\chi 2$ test: $p < 0.01$ there is a statistically significant difference between the two groups of data)

Table 4 The correct use and misuse of the "one + classifier" by JLC

JLC				
Chinese language level	Correct use	Misuse		
		Underuse	Overuse	Replace
CEFR-A2	125 (40.06%)	184 (58.97%)	1 (0.32%)	2 (0.64%)
CEFR-B1	76 (43.93%)	93 (53.76%)	3 (1.73%)	1 (0.58%)
CEFR-B2	41 (38.68%)	62 (58.49%)	3 (2.83%)	0 (0.00%)
Total	242 (40.95%)	339 (57.36%)	7 (1.18%)	3 (0.51%)

Table 5 The correct use and misuse of the "one + classifier" by ELC

ELC				
Chinese language level	Correct use	Misuse		
		Underuse	Overuse	Replace
CEFR-A2	159 (60.16%)	6 (2.93%)	10 (4.88%)	30 (14.63%)
CEFR-B1	677 (90.63%)	12 (1.61%)	8 (1.07%)	50 (6.69%)
CEFR-B2	210 (97.22%)	2 (0.93%)	4 (1.85%)	0
Total	1,046 (89.55%)	20 (1.71%)	22 (1.88%)	80 (6.85%)

Table 3 shows that ELC produced the "one + classifier" structure 74.77 times per 10,000 words, which is significantly higher than the 17.78 times produced by JLC. This suggests that Japanese learners avoid the use of "one + classifier", and/or English learners overuse the "one + classifier". In order to confirm the above hypotheses, errors were categorized as "Underuse" (i.e., omission of "one + classifier" where it is required), "Overuse" (i.e., use of "one + classifier" where it cannot appear), or "Replace" (using the wrong classifier). The results are shown in Tables 4 and 5.

As shown in Table 4, JLC at all three proficiency levels exhibit low levels of accuracy in the use of "one + classifier". The proportion of correct use in fact decreases slightly with increasing proficiency level. Regarding error type, we observe significant underuse of "one + classifier" at all levels. Instances of underuse account for more than 50% of the errors, and this proportion does not change significantly with increased proficiency or length of language study.

The patterns of correct use and misuse of the "one + classifier" structure by ELC as shown in Table 5 are notably different. The overall frequency of misuse by English L1 speakers is low, and the proportion of errors decreases as learners' proficiency level increases. By CEFR-B2 level, ELC can be said to have acquired the "one + classifier" structure. The breakdown by error type also differs from the JLC data. The majority of errors made by ELC are of the "replace" type, but its proportion also decreases as proficiency increases. There are also slightly more instances of overuse by ELC.

In summary, JLC and ELC exhibit contrasting trends in the acquisition of "one + classifier". JLC have difficulty in acquiring "one + classifier". Errors of omission of "one + classifier" are prevalent and do not reduce significantly with increasing proficiency. In contrast, acquisition of "one + classifier" by ELC occurs more smoothly. ELC already achieve higher accuracy levels at CEFR A2 level and accuracy further improves as proficiency rises to the B1 and B2 levels. They do, however, experience some difficulty in selecting the appropriate classifier. These error trends are predictable, given that JLC lack functional equivalents to "one + classifier", whereas ELC possess functional equivalents to "one + classifier" (the indefinite article) but do not have a highly developed system of classifiers in their native language.

3.2 Qualitative Analysis of the Use of "One + Classifier" by Japanese L1 Learners of Chinese

In the previous section we showed the contrasting use of the "one + classifier" structure by JLC and ELC. Use of the "one + classifier" structure appears to be a more problematic and persistent issue for JLC than it is for ELC. In this section, we therefore focus on underuse of "one + classifier" by JLC, considering the structure's different functions.

First, we review the syntactic positions in which "one + classifier" may appear in a sentence and its function in each instance. When placed in the subject position, "one + classifier" can marks either a referential specific (12a) or non-referential (in this case, generic) noun phrase (12b).

(12) a. 一 个 警察 在 追 小偷。
Yí ge jǐngchá zài zhuī xiǎotōu.
A policeman is chasing a/the robber.
(There is a policeman chasing a/the robber.)[3]

b. 一个 警 察 应该 具有 良好 的 身体 素质。
Yí ge jǐngchá yīnggāi jùyǒu liánghǎo de shēntǐ sùzhì.
A policeman needs to be in good physical shape.

When placed in the object position, "one + classifier" marks a referential specific noun phrase, as in the following examples.

(13) a. 墙 上 有 一 张 地图。/ 墙上 挂 着 一 张 地图。/
Qiáng shang yǒu yì zhāng dìtú./ Qiángshang guà zhe yì zhāng dìtú. /
墙 上 少 了 一 张 地图。
Qiáng shang shǎo le yì zhāng dìtú.
There's a map on the wall. / There's a map hanging on the wall. / One of the maps on the wall has disappeared.

b. 他 画 好 了 一 张 地图。/ 他 拿 出来 一 张 地图。
Tā huà hǎo le yì zhāng dìtú. / Tā ná chū lai yì zhāng dìtú.
He drew a map. / He picked up a map.

c. 他 送 给 我 一 张 地图。/ 他 昨天 跟 我 要 了 一 张 地图。
Tā song gěi wǒ yì zhāng dìtú. / Tā zuótiān gēn wǒ yào le yì zhāng dìtú.
He gave me a map. / He asked me for a map yesterday.

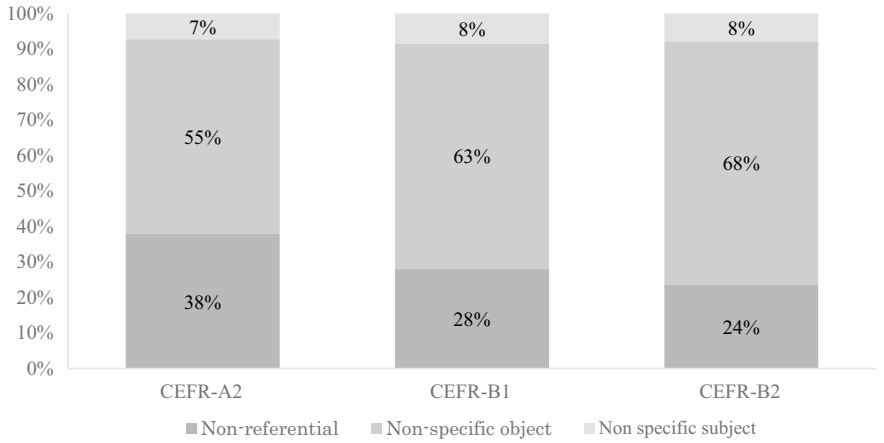


Fig. 1 Distribution of the correct use of "one + classifier" by Japanese L1 learners at each proficiency level

Finally, when used as a predictive component, "one + classifier" is used non-referentially, as in (14).

(14) 他 是 一 个 警察。
Tā shì yí ge jǐngchá.
He is a policeman.

We divided the correct uses of "one + classifier" by JLC following the above categories. The results are shown in Figs. 1 and 2. Figure 1 shows that the correct "one + classifier" structures produced by Japanese L1 learners are mainly found in cases where the indefinite noun phrase appears in object position, or in cases where it is used as a predictive element after the copula "是shi". The proportion of the former increases as proficiency increases. At all levels of proficiency, there are few instances of "one + classifier" appearing in subject position.

Figure 2 shows misuse of "one + classifier". By comparing the proportions of each use in Figs. 1 and 2, we can ascertain which uses prove to be relatively difficult at each level. First, we will consider uses of "one + classifier" in the subject position. At CEFR A2 level, these represent 7% of correct uses (Fig. 1) but 17% of omissions. At B1 and B2 levels, the proportions of correct use and omission are virtually the same. This suggests that the JLC had difficulty correctly including "one + classifier" in subject position at A2 level, but subsequently acquired this use.

Next, we will consider non-referential uses of "one + classifier". At A2 level, these uses represented a higher proportion of errors (47%, Fig. 2) than the proportion of correct use (38%, Fig. 1). As with "one + classifier" in subject position, the non-referential use, therefore, appears to be problematic at A2 level. At B1 and B2 level, the proportions are reversed, suggesting that non-referential uses of "one + classifier" become relatively unproblematic as proficiency increases.

Finally, we consider the uses of "one + classifier" in object position. In this position, "one + classifier" is used to express specific or non-specific referential

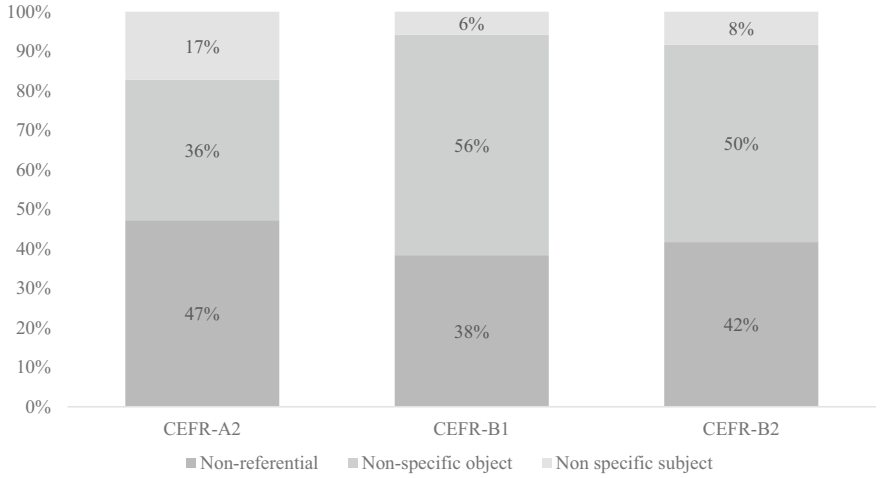


Fig. 2 Distribution of the instances of omission of "one + classifier" by Japanese L1 learners at each proficiency level

noun phrases (see Table 1 above for examples). It can perhaps be said to be the most prototypical of the uses of "one + classifier". This use represents a larger and larger proportion of total use, and misuse fluctuates around the 50% mark.

In this section we gave a breakdown of the types of sentence where "one + classifier" is used or omitted by JLC. In the following section, we turn to characteristic errors appearing in the ELC data.

3.3 Qualitative Analysis of the Use of "One + Classifier" by English L1 Learners of Chinese

In this section we briefly analyze errors made by English L1 learners of Chinese (ELC). As demonstrated in Sect. 3.1, ELC seem able to use "one + classifier" with greater ease than JLC. At intermediate (i.e., CEFR B1 and B2) levels in particular, ELC use "one + classifier" correctly in over 90% of instances. This contrasts with JLC, where the proportion of correct uses is low in elementary (A2) level learners and remains largely unchanged regardless of increases in proficiency.

Nonetheless, a small number of errors do occur in ELC production. Strikingly, errors of overuse are more prevalent than errors of omission. Examples are provided below. Superfluous uses of "one + article" are shown in brackets.

(15) 现在 我 的 家 没有 [一个] 电视， 所以 我 要 买 一个 新 电视。
Xiànzài wǒ de jiā méiyǒu [yí ge] diànshì, suǒyǐ wǒ yào mǎi yí ge xīn diànshì
I don't have a TV in my room, so I want to buy a new one.

(16) 明天 是 我 朋友 的 生日， 可是 我 还 没 给 他 买 [一个] 礼物
Míngtiān shì wǒ péngyou de shēngrì, kěshì wǒ hái méi gěi tā mǎi [yí ge] lǐwù
It is my friend's birthday tomorrow, but I haven't bought him a gift.

(17) 你也 可以 请 他 喝 [一杯] 茶、一起 打 球。
Nǐ yě kěyǐ qǐng tā hē [yìbēi] chá、yìqǐ dǎ qiú.
You can also invite him to have a cup of tea, and to play basketball.

(18) 然后 我们 可以 在 这个 电视 上 看 [一个] 节目。
Ránhòu women kěyǐ zài zhège diànshì shang kàn [yí ge] jiémù.
Then we can watch a program on this TV.

As is evident from the English translations of each sentence, ELC use "one + article" where an indefinite article would be required in English. Overall, this is an effective strategy, as "one + article" and the indefinite article are functionally equivalent in many cases, but it is inappropriate in (15–18). Note that none of these examples express realis events. (15–16) are negative sentences and (17–18) express future possibilities that may or may not occur. All four examples are incompatible with the individualizing function of "one + classifier". English articles are not affected by similar semantic considerations, so it is perhaps unsurprising that a one-to-one mapping of "one + classifier" and the indefinite article in the interlanguage of ELC would lead to overuse of the kind shown in (15–18).[3]

The results of this study are complementary to those of Crosthwaite et al. (2017), who analyzed the expression of definite discourse-new (so-called "bridging") reference by English, Korean and Japanese learners of Chinese. "Bridging" refers to

[3] Languages like Spanish require the definite article before generic noun phrases, as in (i). This contrasts with English, where a definite article, an indefinite article, or a bare plural noun phrase are all possible.

(i) *(Los) tigres comen carne.
The tigers eat meat
'Tigers eat meat.' Based on Chang (2016: 800).

On the other hand, languages like Dutch disallow articles in sentences like (14), i.e., in at least some nonreferential contexts, as in (ii).

(ii) Marie is leerkracht
Marie is teacher
'Marie is a teacher.' Based on Aguilar-Guevara, Le Bruyn & Zwarts (2014: 8)

Just as differences among "article-less languages" (i.e., differing degrees of grammaticalization of "one + classifier" in Chinese and Japanese) can affect L2 English article acquisition, differences among languages possessing articles may affect the ease of acquisition of particular uses of "one + article" in L2 Chinese. This is an empirical question that awaits further research.

situations where a new referent can be linked to a previous referent in the discourse. For example, in (19a), "waiter" can be marked with the definite article because its existence is implied by "restaurant" in the preceding sentence. Chinese does not mark definite discourse-new reference morphologically (19b), in contrast to standard uses of "one + classifier" to mark discourse new, noninferable reference, e.g., in (7–9) and (11).

(19) a. A man walks into a restaurant. The waiter gave him a menu.
b. 有 一 个 男孩 走 进 餐厅， 服务员 给 他 一 份 菜单。
Yǒu yí ge nánhái zǒu jìn cāntīng, fúwùyuán gěi tā yí fèn càidān.
'A boy entered a restaurant. The waiter gave him a menu.'
(Crosthwaite et al, 2017: 628)

ELC showed a tendency to use Chinese demonstratives and classifiers analogously to English articles. In other words, they used "demonstrative + classifier + noun" for bridging reference, and "one + classifier" for nonbridging reference (Crosthwaite et al., 2017: 644).[4] While article-like use of "one + classifier" may result in native-like use as attested in Sect. 3, in irrealis situations (the current study) or bridging situations (Crosthwaite et al.'s study), over-generalization of the functional equivalence between articles and "one + classifier" results in infelicitous use of the latter.

This section has shown how both the high level of overall acquisition of "one + article" and its overuse in specific circumstances can be explained if we assume that ELC associate "one + article" with the indefinite article in their L1. Whether this phenomenon is due to conscious strategies by individual learners, the result of L2 pedagogy or an unconscious association is a task for further research.

3.4 Analysis of the Use of "One + Classifier" by Korean L1 Learners of Chinese

In this section, we discuss the data relating to the use of "one + classifier" in essays written by Korean native learners of Chinese (henceforth, "KLC"). Table 6 shows that among instances of misuse of "one + classifier" (underuse: 78.40%, overuse: 16.00%, replace: 5.60%), the percentage of underuse (196: 78.40%) is remarkably high in the essays by KLC. This high proportion of underuse errors is similar to the results for the JLC data shown in Sect. 3.1. This may be due to the fact that Korean is typologically similar to Japanese.

[4] Native speakers of Korean in Crosthwaite et al's (2017) study did not make similar errors in their L2 Chinese.

Table 6 The correct use and misuse of the "one + classifier" by KLC

KLC (294 essays)	Correct use	Misuse			Total
		Underuse	Overuse	Replace	
	1125 (81.82%)	196 (14.25%)	40 (2.91%)	14 (1.02%)	1056 (100%)
		250 (18.18%)			
		196 (78.40%)	40 (16.00%)	14 (5.60%)	250 (100%)

That is, Korean, as well as Japanese, does not have a determiner position (DP) in syntax, which may have affected the underuse of "one + classifier". The following are some specific examples for misuse of underuse in KLC data.

(20) 几年前， 朝鲜 有 <φ→一个> 很 大 的 机会，就是 六个 国家 的 会谈，但是…。
Jǐniánqián, Cháoxiǎn yǒu <φ→yí ge> hěn dà de jīhuì, jiùshì liù ge guójiā de huìtán, dànshì….
(0051_20130311_PKU_IL_CH_008)
"A few years ago, there was a great opportunity in Korea. It was the Six-Nation Talk, but…"

(21) 所以 我 开始 了学习 汉语。 那 时候 隔壁 <φ→一个> 姐姐 的 专业 是 中文系，
Suǒyǐ wǒ kāishǐ le xuéxí Hànyǔ. Nà shíhou gébì <φ→ yí ge> jiějie de zhuānyè shì Zhōngwénxì,
除了 那 个 姐姐 教 我 汉语 以外，我 自己 学 了 汉语。
chúle nà ge jiějie jiāo wǒ Hànyǔ yǐwài, wǒ zìjǐ xué le Hànyǔ.
(0541_20130617_PKU_IL_CH_207)
"So I started to study Chinese. At that time, a girl who lived next door majored in Chinese, and except for her teaching me, I studied Chinese by myself."

The percentage of correct use is considerably higher in the KLC data than in the JLC data (KLC: 81.82%, JLC: 40.95%). This difference is presumably due to differing proficiency levels of the KLC and JLC learners. The Korean learners belonged to the Chinese language department of a university in China. In other words, they are learning Chinese not only in the classroom but also in their living environment, which means their Chinese level is likely to be higher. In contrast, Japanese learners were studying Chinese as a foreign language in Japan.

However, in spite of the higher Chinese level of KLC, the proportion of overuse is higher in the data of KLC (2.91%) than in the data of JLC (1.18%), which may be due to individual differences among learners. These errors are different from those of JLC, such as use of "one + classifier" in negative sentences, and are a potential area of future research.

(22) 我们 在 人生 的 路途 中， 总会 面对 很 大 的 危机。 但是 我 相信，
Wǒmen zài rénshēng de lùtú zhōng，zǒnghuì miànduì hěn dà de wēijī. Dànshì wǒ xiāngxìn,
拥有 <一个→φ> 梦 想 的人， 能够 克服 它。
yōngyǒu < yí ge →φ> mèngxiǎng de rén, nénggòu kèfú tā.
(0053_20130311_PKU_IL_CH_021)
"We will always face great crisis in our life. But I believe that people who have a dream can overcome it."

(23) 如果 怕 自己的 孩子 在 社会 上 淘汰，就要 让 孩子 一 上 学 就 开始 学
Rúguǒ pà zìjǐ de háizi zài shèhuì shang táotài, jiùyao ràng háizi yí shàng xué jiù kāishǐ xué
<一门→φ> 外语。
< yì mén →φ> wàiyǔ. (0062_20130311_PKU_IL_CH_045)
"If you are afraid that your child will be weeded out in society, you should have your child start learning a foreign language as soon as they enter school."

4 L2 English Article Use by Chinese L1 and Japanese L1 Learners

In this section we focus on the acquisition of articles in L2 English. In Sect. 4.4 we will refer back to error examples in Sect. 3.3, demonstrating how the presence or absence of a realis/irrealis distinction in the use of determiners affects L2 English as it does L2 Chinese. First, we introduce our data set and more general findings.

4.1 Background

In Sect. 3.1 we demonstrated how similarities and differences between L1 and L2 appear to contribute to contrasting trends in the use of the "one + classifier" in L2 Chinese. Specifically, the functional similarity between the English indefinite article and the Chinese "one + classifier" structure appears to be more conducive to the acquisition of the L2 form than the morphological similarity between "one + classifier" in Chinese and Japanese. In the following sections, we offer a preliminary investigation of similar processes in L2 article use by Chinese learners of English (CLE) and Japanese learners of English (JLE).

Research on the acquisition of English L2 articles is voluminous, with the majority of studies being based on forced elicitation tasks, self-paced reading tasks, or other experimental designs. Research has particularly focused on native speakers of "article-less" languages, including Chinese (Díez-Bedmar & Papp, 2008; Robertson, 2000; Snape, Leung, & Ting, 2006; Xu et al., 2016; Yang & Ionin, 2009) and Japanese (Butler, 2002; Hawkins et al., 2006; Snape, Leung, & Ting, 2006; Ogawa, 2007; Kume, 2016; Yamada, 2019).[5] Learners of these languages are believed to

[5] This is by no means intended to be a comprehensive list. Other "article-less" languages frequently analyzed include Korean and Russian.

find the acquisition of articles problematic due to an absence of equivalent features in L1. Nonetheless, there is little consensus about which particular uses of articles learners struggle with the most, and the underlying causes. It has been argued that learners' choices of article can be affected by factors including definiteness, specificity, countability, and reference salience. Some research has argued that individual learners whose native languages lack articles fluctuate in their L2 article use, although results are not uniform between or within languages. The differing results of previous studies have numerous potential causes, including mode- or task-related effects, learner proficiency, and even different analytical frameworks.[6]

One further factor, which will be the focus of the present study, is the influence of learners' native language. Specifically, we investigate whether L1 Chinese and L1 Japanese learners of English exhibit different patterns in their use and misuse of English articles. We introduce one study that is particularly pertinent to this research question below.

Crosthwaite (2016a) is a corpus study comparing article use by L1 Mandarin, L1 Korean and L1 Thai learners of English. While all three of these languages are regarded as "article-less", Crosthwaite (2016a: 78) asserts that in L1 language use Chinese speakers "appear to use overt syntactic means to signal (in)definiteness (e.g., overt or deliberately omitted numeral + classifiers, demonstratives) more often and in more clearly differentiated article contexts than Korean and Thai speakers". As a result, "the potential for positive L1 transfer of certain form/function relationships associated with the English article system appears to be greater" for Chinese L1 learners.

In Crosthwaite's (2016a) study, Chinese L1 learners indeed exhibited more target-like use of articles than Korean L1 and Thai L1 learners. Furthermore, Chinese L1 learners exhibited similar levels of accuracy for zero, definite and indefinite articles, in contrast to Korean and Thai L1 learners, for whom definite and indefinite articles proved to be more challenging than zero articles. Crosthwaite (2016a: 33) concludes that this phenomenon can be regarded as the effect of positive transfer, and as evidence that Chinese "does, in fact, have an article-like system". In summary, Crosthwaite's study demonstrates the practical effects on second language acquisition of the article-like uses of "one + classifier" and demonstratives noted by earlier studies (Chen, 2004; Gundel et al., 1993; Snape, Leung & Ting, 2006). It also demonstrates that differences among learners whose L1 lack articles can be observed not only in experimental contexts but also in natural language use as captured in learner corpora.

In the remainder of Sect. 4, we examine our own data for patterns similar to those observed by Crosthwaite (2016a) and parallel to the L2 Chinese data detailed in previous sections.

[6] For instance, there are at least two competing definitions of "specificity" in the context of Second Language Acquisition (SLA) research on article use. See, e.g., Ionin & Díez-Bedmar (2021) for discussion.

Table 7 Article use by Chinese L1 learners

		Correct form				
		zero	a	the	Other	**Total errors**
Original form	zero	*205*	68	60	8	**136**
	a	3	*480*	5	0	**8**
	the	24	35	*524*	30	**89**
	Total errors	**27**	**103**	**65**	**38**	**233**

4.2 Data Set

The data set referred to in the rest of Sect. 4 comprises another subsection of the three-way learner corpus developed at Tokyo University of Foreign Studies. There are two important caveats regarding this particular data set. First, we lack proficiency data of the type referred to above for the Chinese L2 data. The Japanese L1 learners were first-year English majors at the time of data collection, whereas the Chinese L1 learners were fourth-year English majors. Second, the data is taken from a translation task, in which learners translated equivalent texts from their L1 into English. The task was thus controlled for content but not for proficiency, so the results presented below cannot be compared directly to the analysis in Sect. 3.1. They can, however, be considered as another potential instantiation of L1 effects on the acquisition of L2 forms, and represent a task type that has not, to our knowledge, been employed in L2 acquisition studies concerning English articles.

4.3 Quantitative Analysis

Tables 7 and 8 show article use by CLE and JLE respectively. Raw figures have been used because both data sets consist of the same number of translations of the same source text (n = 40). Shaded cells represent correct use of either the zero, indefinite or definite article.[7]

Error patterns appear to be largely similar in the two groups of learners, with omissions of the indefinite article being most prominent, followed by omissions of the definite article. Examples of each error type are shown in (24) and (25) below.[8]

[7] "other" in Tables 6 and 7 refers to forms other than articles, such as possessive pronouns. We will not consider these forms in detail in the present paper.

[8] Abbreviations: JP = Japanese native speaker; CH: Chinese native speaker.

Table 8 Article use by Japanese L1 learners

		Correct form				
		zero	a	the	Other	**Total errors**
Original form	zero	*298*	97	71	30	**198**
	a	5	*574*	16	5	**26**
	the	13	32	*354*	18	**63**
	Total errors	**18**	**129**	**87**	**53**	**287**

(24) Omission of the indefinite article
a. My teacher would roll his quilt into (zero → a) strip like a rolled cake . (CH_38)
b. Hu put Longjing Green tea in the Chinese mug with (zero → a) lid and poured hot water from a vacant bottle. (JP_10)

(25) Omission of the definite article
a. I, at public expense by (zero → the) Chinese Government, furthered my study in Fudan University (CH_12)
b. It was at simple age but I remembered (zero → the) hospitality of the Prof. Hu vividly. (JP_39)

Furthermore, overuse of articles (i.e., where the correct form is zero) represent a very low proportion of total errors (approximately 12% for CLE and 6% for JLE). Examples for the definite article are given in (26) for reference.

(26) Article overuse:
a. the redness of the candy box was so bright that I nearly mistook it was for (the → zero) wedding candies. (CH_46)
b. his family treated me to Babaofan- (the → zero) decorated cakes made of glutinous rice with eight dried fruits. (JP_07)

The error trends illustrated above are expected, given that both Chinese and Japanese are regarded as article-less languages. Note, however, that JLE exhibit a higher total frequency of errors than CLE, and that the proportion of errors of omission is greater for JLE (69%) than it is for CLE (58%). This difference appears to mirror Japanese native speakers' omission of "one + classifier" in L2 Chinese, and also suggests that Japanese learners of English may be closer to Korean and Thai learners of English than they are to Chinese learners of English, in terms of their frequent omission of articles.

The data in Tables 7 and 8 were then used to calculate the Target Language Use (TLU) as proposed by Pica (1983) for the definite, indefinite, and zero articles. TLU takes into account both overuse and underuse of a target form and has been repeatedly adopted in previous studies on article use (Crosthwaite, 2016a, 2016b; Díez-Bedmar & Papp, 2008). TLU is calculated using the formula shown below. A TLU score of 1 represents entirely "target-like" use. The lower the TLU, the more problematic the form for learners.

(27) $$\text{TLU} = \frac{\text{no. of suppliances in obligatory contexts}}{(\text{no. of obligatory contexts}) + (\text{no. of suppliances in non-obligatory contexts})}.$$

Table 9 TLU of articles by Chinese L1 and Japanese L1 learners

	Chinese L1			Japanese L1		
	zero	a	the	zero	a	the
obligatory contexts	232	583	589	316	703	441
correct suppliances in obligatory contexts	205	480	524	298	574	354
suppliances in non-obligatory contexts	136	8	89	198	26	63
TLU	**0.56**	**0.81**	**0.77**	**0.58**	**0.79**	**0.70**

The TLU for each article is shown in Table 9. The patterns can be summarized as follows. First, for both groups of learners, TLU was lowest for the zero article and highest for the indefinite article. Comparison with Tables 7 and 8 shows that the low TLU for the zero article was due to "overuse" of the zero article, in other words, omission of the indefinite article and, to a lesser extent, omission of the definite article. Again, such errors of omission are expected given that both Chinese and Japanese lack articles. Despite this trend to omit the indefinite article, in the majority of cases learners' selection of the indefinite article was in fact appropriate. As a result, the indefinite article showed the highest TLU, for both CLE (TLU = 0.81) and JLE (TLU = 0.79).

Second, the TLU for the zero article and indefinite article was almost identical for CLE and JLE. In other words, the functional similarity between "one + classifier" and the indefinite article did not have an obvious positive effect on L2 article acquisition by CLE in the current data set.

In contrast, the TLU for the definite article was notably higher for CLE (TLU = 0.77) than for JLE (TLU = 0.70). The precise reason for this difference is not entirely clear, but it reflects a greater overall use of the definite article by CLE. Table 10, calculated from the figures in Tables 7 and 8, shows the proportion of articles used by each group of learners. In the Chinese L1 data, the definite article accounts for over 40% of overall article use, the highest of the three possible forms, whereas in the Japanese L1 data, the percentage is less than 30%, the lowest of the three possible forms in article position. Table 10 also shows that the L1 Japanese data exhibits a greater proportion of zero article use (32.8%) than that seen in the L1 Chinese data (23.6%). These figures refer to overall use irrespective of whether the usage is correct or incorrect, but they are suggestive of a tendency among JLE to avoid articles more generally.

Taken together, the data in tables from Tables 7, 8 and 9 and 10 suggest that CLE show greater mastery of articles in general and the definite article in particular compared to JLE. This may ultimately be due in part to the functional equivalence between the definite article and determiner-like uses of demonstratives in Chinese, though it is not clear why a similar phenomenon does not occur with the indefinite article in the current data set.[9] There may be task-related issues, so further studies should be conducted using different data sets to enable data triangulation.

[9] This is not to imply that CLE translated demonstratives in the source text as definite articles, and that JLE did not. We merely suggest that the propensity of use of demonstratives in L1 Chinese

Table 10 Proportion of use of each article

Article	L1 Chinese		L1 Japanese	
	Raw frequency	Percentage	Raw frequency	Percentage
zero	341	23.6	496	32.8
a	488	33.8	600	39.7
the	613	42.5	417	27.6
Total	**1442**		**1513**	

4.4 Qualitative Analysis

This section briefly discusses some concrete examples of article errors and suggests how these may have been influenced by learners' L1. As mentioned above, the current data set is a translation task and so L1 influence can potentially occur not only through learners' interlanguage, but explicitly through the source text. However, the text in question features few uses of either "one + classifier" or determiners, so L1 influence is likely to occur more generally. Below, we consider how the absence of "one + classifier" in L1 Chinese may lead to omission of articles in L2 English.

The examples in (28) show omission of articles by CLE, with the appropriate article added in brackets. These examples feature negation and are found almost exclusively in the CLE data. This is a reflection of the fact that the "one + classifier" structure is not required or even allowed in irrealis sentences such as negatives and conditionals. Crucially, this mirrors the erroneous use of "one + classifier" structure by ELC in irrealis sentences illustrated in Sect. 3.3. In other words, although "one + classifier" and articles have functional similarities and appear to be sources of positive transfer overall, restrictions on the use of the former to realize contexts appear to contribute to overuse of "one + classifier" by ELC and underuse of the indefinite article in particular contexts by CLE.

(28) a. At that time, teachers in Fudan University didn't have (∅ → an) individual laboratory (CH_57)

b. Since teachers dorm did not have (∅ → a) telephone, I mostly called without invitation. (CH_24)

Finally, we mention some other areas of difficulty for learners. Article errors are concentrated in three main areas: a. use of the definite article for bridging reference, b. non-referential use of the indefinite article, and c. inappropriate use of the definite article in bridging relations.

First, learners appear to have difficulty with the definite article used for bridging reference, i.e., where a new referent can be inferred from another referent. The translation task includes such pairs as "bed" and "futon", "box" and "lid", and "restaurant" and "menu". Learners frequently failed to mark "futon", "lid" or "menu" with

functionally similar to uses of the English definite article (Crosthwaite 2016a) primes CLE to mark definiteness grammatically with more regularity than JLE.

the definite article. In fact, these errors occur almost exclusively among Japanese learners. This is unsurprising, as Japanese neither marks bridging relations morphologically, nor distinguishes these from new referents (which are also unmarked morphologically). Therefore, even JLE who successfully use the definite article for previously mentioned referents may struggle with bridging reference. Instead, they use the indefinite article or omit the article altogether (29).

(29) Errors involving use of the definite article for bridging reference:
   a. The Professor used to give me individual tutorials, sitting on the bed with (∅ → the) "futon" rolling up like rolled cakes and turned into couch. (JP_31)
   b. Then, he took (a → the) lid off from a red candy box like a present given in a wedding ceremony (JP_43)
   c. After that, I always order Babaofan whenever I see it on (∅ → the) menu of Chinese restaurants and remember Omotenashi by Prof. Hu. (JP_42)

Second, both Chinese and Japanese learners tend to omit the indefinite article in non-referential situations. The irrealis sentences in (28) can be regarded as non-referential. In addition, there are frequent omissions of the indefinite article in contexts like those shown in (30). The example in (30a) refers to Chinese restaurants in general and the examples in (30b-d) do not to refer to actual concrete objects.[10]

(30) Omission of the indefinite article in non-referential contexts:
   a. After that Eight Treasures Rice was on my must-order list everytime I went to (∅ → a) Chinese restaurant (CH_35)
   b. When I attended the class, Mr. Hu always rolled the quilt with cotton wadding into a long strip like (∅ → a) Western cake (CH_17)
   c. Then, he would also uncap a red box, which looked like a gift for guests at (∅ → a) wedding reception, and offered me candies in it, smiling kindly. (JP_03)
   d. Private guidance of my paper carried out on his bed, which likes (∅ → a) jelly roll and imitate like (∅ → a) bench. (JP_08)

Finally, we comment briefly on errors where learners selected the definite article instead of the indefinite article. This represents the third most common error type among both groups of learners, following omission of the indefinite article and omission of the definite article (see Tables 7 and 8). This error type largely represents inappropriate use of the definite article in bridging relations. In other words, learners use the definite article for new referents despite their being no clear implication of the referents existence. For example, in (31a) there is no reason to assume the existence

[10] Note that (30b-d) all include *like* or similar expressions.

of a teacup, let alone one known to the reader, based on the text up to that point. Learners are therefore marking an indefinite, specific referent as if it were a definite, specific referent.

(31) Erroneous use of the definite article:
- a. After I sat down, my teacher would pinch some Longjing Tea to (the → a) traditional lidded Chinese teacup (CH_56)
- b. From then, I would order a Babao rice every time I came to (the → a) Chinese restaurant (CH_41)
- c. After I sat down, the teacher would first put some Longjing tea in a traditional Chinese style cup which with a lid and then poured hot boiling water from (the → a) thermos to make a cup of hot tea for me. (CH_32)
- d. Then, he took away the lid of (the → a) red candy box just like a gifts for guests at wedding ceremony (JP_32)
- e. When I sit down on the futon, Prof. Hu, firstly, put one pick of Longjing Green Tea into (the → a) Chinese mug cup with cover (JP_14)

Why, then, does this error trend emerge? We suggest that the presence of modifying elements (underlined in the examples above) gives learners the impression that there is enough information for the reader to identify the referent.[11] While this is an incorrect application of the English definite article, such behavior has been observed in previous studies. The current study, therefore, supports the idea that learners confuse the features of definiteness and specificity when choosing the appropriate English article.

Section 4 has provided some partial supporting evidence for the assertion that learners' native language affects acquisition of L2 forms. L1 influence is not uniformly positive or negative but can lead to both native-like and erroneous use of L2 forms, depending on a range of other factors. In the next section, we briefly turn to the acquisition of L2 Japanese.

5 Use of Japanese "One + classifier" in Compositions by L1 Chinese and English Learners

This section examines the use of "one + classifier" in Japanese compositions by Chinese L1 learners (CLJ) and English L1 learners (ELJ). The data used in this

[11] Nevertheless, examples like (31c) are not accompanied by modifying elements. Assuming that article choice is not random, there may be a cultural element at play. Errors with the referent "thermos" only appear among Chinese native speakers. Perhaps there is an assumption among Chinese native speakers that "tea" earlier in the narrative is sufficient for a bridging reference with "thermos". Native speakers of English may be more likely to accept "kettle" marked by the definite article in a similar context.

Fig. 3 Five-frame cartoon by I-JAS

Table 11 Modifier elements and frequency of occurrence with "inu" (dog)

Modifier for "inu" (dog)	Chinese L1	English L1	Japanese NS
zero form	421 (1539.08)	263 (2610.16)	111 (1901.66)
"one + classifier"	13 (47.53)	**0 (0.00)**	**0 (0.00)**
"aru" (a certain)	12 (43.87)	**0 (0.00)**	**0 (0.00)**
determiner	19 (69.46)	6 (59.55)	1 (17.13)
other	95 (347.30)	28 (277.89)	33 (565.36)

(Numbers in brackets indicate adjusted frequency per 100,000 words)

section is taken from the "Story writing 1" section (SW1)[12] of I-JAS.[13] We examined the use of modifying elements preceding the noun "inu" (dog) in the SW1 data for CLJ and ELJ and compared each to the use of modifying elements by native speakers of Japanese on the same task. The types of modifying elements used by L1 speakers of each language are summarized in Table 11. The table reveals the following two points.

First, Chinese L1 learners show a significantly higher frequency in the use of "one + classifier" (Chinese L1: 47.53, Japanese native: 0.00, $G^2 = 5.03$, $p = 0.025$) and of the indefinite element "aru" (Chinese L1: 43.87, Japanese native: 0.00, $G^2 = 4.64$, $p = 0.031$) as compared to Japanese native speakers. Second, in contrast to Chinese L1 learners, English L1 learners, like Japanese native speakers, do not exhibit overuse of

[12] Learners produce sentences to describe the story depicted in the five-frame cartoon shown in Fig. 3.

[13] I-JAS: "International Corpus of Japanese as a Second Language" (http://lsaj.ninjal.ac.jp/) (National Institute for Japanese Language and Linguistics),refer to Chap. 14 in this book.

"one + classifier" or "aru". Indeed, neither ELJ nor Japanese native speakers show any use of "one + classifier" or "aru".

This phenomenon suggests that although both Chinese and Japanese possess the "one + classifier" structure, its function is different in both languages, otherwise we would expect to see similar patterns of use between Japanese native speakers and CLJ. While Japanese possesses the "one + classifier" structure, in practice it is not selected by native speakers in the current context, where conveying numeral information is not necessary.

Next, we will discuss "one + classifier" from the perspective of "boundedness"[14] in cognitive structure and "the function of introducing new information" in informational structure.

Shen (1995) suggests that the function of classifiers in Chinese is to "embody the opposition between bounded and unbounded" in the human cognitive structure. We analyze the use of "one + classifier" by CLJ in the I-JAS data making reference to this concept of "boundedness". Errors related to classifiers have been corrected, with corrections shown in brackets. Unrelated errors have been left uncorrected.

"Locative Structure".

(32) a. サンドイッチを 入った バスケットの 近くに
Sandoitchi-o hait-ta basuketto-no chikaku-ni
Sandwich-ACC enter-PST basket-POSS near-DAT
犬が 一只（一匹）[15] あります。 (CCH29-SW1)
inu-ga ip-piki ari-masu
dog-NOM one-CL be-COP

b. 装 着 三明治 的 篮子 的 附近 有 一只 狗。
Zhuāng zhe sānmíngzhì de lánzi de fùjìn yǒu yì zhī gǒu.
(Chinese translation by the author)
There is a dog near the basket with the sandwiches inside.

15

(33) a. マリ と ケンの うちには 犬が 一匹 飼います。
Mari-to Ken-no uchi-ni-ha inu-ga ip-piki kai-masu
Mari-COM Ken-POSS house-DAT-TOP dog-NOM one-CL feed-COP
(CCS03-SW1)

b. 玛丽 和 小健 的家里养 了 一 条 狗。
Mǎlì hé Xiǎojiàn de jiāli yǎng le yì tiáo gǒu.
(Chinese translation by the author)
Mari and Ken have a dog. (There is a dog in Ken and Mari's house.)

[14] The term "boundedness" in linguistics was first pointed out by Langacker (1987, 1991a, b, 2001) from the perspective of cognitive linguistics. Boundedness is, in essence, generally considered to be the concept of whether or not a boundary exists within something. Among nouns, "countable nouns" with clearly defined boundaries express "boundedness", while those with no clear boundaries, such as "collective nouns", express "unboundedness". In Chinese, the restrictions on the syntactic structure of quantifiers effectively embody the basic opposition of "bounded" and "unbounded" in human cognition (Shen 1995).

[15] "只zhī" is a classifier in Chineese, whereas "匹hiki" is the correct classifier in Japanese.

"Verb-Complement Structure".

(34) a. 彼らは…、 バスケットを 開けて、 一つの 犬を
Karera-wa basuketto-o ake-te hitotsu-no inu-o
they-TOP basket-ACC open-PTC one-POSS dog-ACC
見つけて… (CCH38-SW1)
mitsuke-te
find-PTC

b. 他们…， 打开 了 篮子，看到 一 只 狗…
Tāmen …，dǎkāi le lánzi， kàndào yì zhī gǒu…
(Chinese translation by the author)
They open the basket and find a dog…

Examples of Japanese sentences written by Chinese L1 learners are shown in (35–37). The equivalent sentences in Chinese are provided for comparison. As the Chinese sentences require "one + classifier" to realize boundedness in each case, the overuse of "one + classifier" in L2 Japanese by Chinese L1 learners in I-JAS may be due to L1 transfer. On the other hand, the overuse of classifiers is also related to factors concerning information structure. In Japanese, information structure is typically expressed by marking sentence elements with either the case marker "ga" or topic marker "wa".

(35) a. ケンは 蓋を 開けって（開けて）、 食べ物を 出そう と
Ken-wa futa-o aket-te tabemono-o da-so to
Ken-TOP lid-ACC open-PTC food-ACC take out-will COMP
思って、 パット 一匹の 犬が 走ってしまった。
omot-te patto ip-piki-no inu-ga hashit-teshimat-ta
think-PTC suddenly one-CL-POSS dog-NOM run-regret modal-PAST
(CCM11-SW1)

b. 小健 打开 盖子， 想 要 取出 食物，突然 一 条 狗 跑 了 出来。
Xiǎojiàn dǎkāi gàizi， xiǎngyào qǔchū shíwù，tūrán yì tiáo gǒu pǎo le chūlái.
Ken opened the lid, wanting to take out the food, and a dog suddenly ran out.
(Chinese translation by the author)

In (35b), which is the Chinese translation of (35a), the subject "gou" (dog) is an indefinite noun expressing new information and so must be marked with "one + classifier". Because of this stipulation in L1, it is assumed that Chinese L1 learners will overuse "one + classifier" in Japanese sentences like (35a). In I-JAS, the sentences including "one + classifier" produced by Chinese L1 learners basically co-occur with "-ga" ("-ga": 10 cases, "-wo": 2 cases, "no particle": 1 case) and there are no cases of co-occurrence with the particle "-wa" in particular.

Table 12 Use of "wa" and "ga" by Japanese native speakers and Chinese L1 learners

	Japanese NS (n = 50)		Chinese (n = 100)	
	Dog (scene1)	Dog (scene2)	Dog1 (scene1)	Dog2 (scene2)
-ga	48 (96%)	50 (100%)	50 (50%)	54 (54%)
-wa	**1 (2%)**	**0 (0%)**	**35 (35%)**	**31 (31%)**
others	1 (2%)	0 (0%)	15 (15%)	15 (15%)
total	50 (100%)	50 (100%)	100 (100%)	100 (100%)

(scene1: panels① and ② in Figure#1, scene2: panel ④ in Figure#1)

(36) そして、 二人が 地図について 相談していた 所に、
Soshite futari-ga chizu-nitsuite soodanshi-tei-ta tokoroni
then two of them-NOM map-about discuss-DUR-PST at that time
一匹の 犬が こっそりと バスケットに 入って、
ip-piki-no inu-ga kossorito basuketto-ni hait-te
one-CL-POSS dog-NOM secretly basket-DAT enter-PTC
その中の 物を 食べてしまいました。 (CCM15-SW1)
sononaka-no mono-o tabe-teshimai-mashi-ta
inside-POSS food-ACC eat-regret modal-PLT-PST

Then, as the two of them were discussing the map, a dog sneaked into the basket and ate the food inside.

(37) 一匹の 犬ちゃんが 用意してあった バスケットに 入ったことは…
Ip-piki-no inu-chan-ga yoishi-teatta basuketto-ni haitta-koto-wa
one-CL-POSS doggie-NOM be prepared basket-DAT enter-thing-TOP
(CCT15-SW1)

…that a doggie had gone into the basket (they had) prepared.

Table 12 shows a comparison of the use of "-wa" and "-ga" added to "inu" (dog) in the two scenes of the story writing task. While Japanese native speakers almost exclusively use "-ga" in both scenes, CLJ use "-wa" more frequently (35% in scene 1, 31% in scene 2). This may be caused by the fact that learners have not fully acquired the distinction between "-wa" to mark old information and "-ga" to mark new information.

6 Implications for Chinese Teaching

Based on the three case studies presented above, in this section we outline the implications of our findings for teaching Chinese as a foreign or second language.

Elementary and intermediate Chinese language teaching materials currently in use in Japanese universities typically provide little or no explanation of the "indefinite" use of the "one + classifier" structure. Classifiers are treated as "units to count objects" and are usually only brought up in relation to the range of classifiers used to mark objects with different properties and shapes. Japanese does not have a grammatical form expressing indefiniteness and lacks obligatory marking of the

(in)definiteness of noun phrases. As such, Japanese learners are predicted to struggle to acquire the "one + classifier" structure unless they are taught it explicitly.

We propose that the striking tendency for JLC to omit "one + classifier" is caused by such differences. Likewise, this may explain why ELC, whose native language shows a grammatical distinction between definite and indefinite, did not tend to omit "one + classifier", instead overusing the structure on occasion.

The implications for Chinese language pedagogy aimed at L1 Japanese learners can be summarized as follows. First, the "indefinite" use of the "one + classifier" structure should be introduced from the elementary or intermediate level. Second, it may be effective to introduce the "one + classifier" structure through reference to the English indefinite article, to which all university level learners will have been exposed to. In this way, L2 (English) knowledge could potentially aid L3 (Chinese) acquisition.

7 Conclusion

This paper has introduced three case studies examining the possible influence of learners' native language on the acquisition of L2 forms. The findings can be summarized as follows. In L2 Chinese, functional similarities between the indefinite article and "one + classifier" had a beneficial effect on L2 acquisition for English native speakers, whereas morphological similarities between "one + classifier" in Japanese and Chinese led to the omission of the target form by Japanese native speakers. The same morphological similarities also contributed to the overuse of "one + classifier" in Chinese native speakers' L2 Japanese, although other factors including information structure also appear to be at play. Finally, functional similarity may also contribute to greater and more accurate use of English articles by Chinese native speakers in some contexts. Overall, this paper's findings suggest that when teaching grammatical forms there is a need for nuanced consideration of characteristics of learners' native languages, especially given that superficial morphological similarities may be just as likely to lead to errors than to native-like use.

References

Aguilar-Guevara, A., Le Bruyn, B., & Zwarts, J. (2014). Advances in weak referentiality. In A. Aguilar-Guevara, B. Le Bruyn, & J. Zwarts (Eds.), *Weak referentiality* (pp. 1–16). John Benjamins.

Butler, Y. (2002). Second language learners' theories on the use of English articles an analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition, 24*, 451–480.

Chang, H. (2016). Interpretation of bare and demonstrative noun phrases in the acquisition of Mandarin. *Language & Linguistics, 17*(6), 797–825.

Chen, P. (2004). Identifiability and definiteness. *Linguistics, 42*(6), 1129–1184.

Crosthwaite, P. (2016a). L2 English article use by speakers of article-less languages: A learner corpus study. *International Journal of Learner Corpus Research, 2*(1), 68–100.
Crosthwaite, P. (2016b). Definite article bridging relations in L2: A learner corpus study. *Corpus Linguistics and Linguistic Theory, 15*(2), 297–319.
Crosthwaite, P., Yeung, Y., Bai, X., Lu, L., & Bae, Y. (2017). Definite discourse-new reference in L1 and L2 the case of L2 Mandarin. *Studies in Second Language Acquisition, 40*(3), 625–649.
Díez-Bedmar, M., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, S. Papp, & B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 147–175). Rodopi.
Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*(2), 274–307.
Hawkins, R., et al. (2006). Accounting for English article interpretation by L2 speakers. *EUROSLA Yearbook, 6*, 7–25.
Ionin, T., & Díez-Bedmar, M. (2021). Article use in Russian and Spanish learner writing at CEFR B1 and B2 levels: Effects of proficiency, native language, and specificity. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 10–38). Cambridge University Press.
Kume, K. (2016). The role of universal semantic features in L2 English article choice by L1 Japanese speakers. *Second Language, 15*, 31–51.
Langacker, R.W. (1987). *Foundations of cognitive grammar. Vol. I: Theoretical Prerequisites.* Stanford University Press.
Langacker, R.W. (1991a). *Concept, image and symbol: The cognitive basis of grammar*. Mouton de Gruyter.
Langacker, R. W. (1991b). *Foundations of cognitive grammar. Vol II: Descriptive application.* Stanford University Press.
Langacker, R. W. (2001). The English present tense. *English Language and Linguistics, 5*(2), 251–272.
Ogawa, M. (2007). The acquisition of English articles by advanced EFL Japanese learners: Analysis based on noun types. *Osaka Prefecture University Journal of Language and Culture Language and Information, 3*, 133–151.
Pica, T. (1983). Methods of morpheme quantifications: Their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6*, 69–78.
Robertson, D. (2000). Variability in the use of the English article system by Chinese learners of English. *Second Language Research, 16*(2), 135–172.
Shen(沈家煊), J. (1995). "Youjie" yu "wujie" ("有界"与"无界"). *Zhongguo Yuwen (中国语文), 5*, 367–380.
Snape, N., Leung, Y., & Ting, H. (2006). Comparing Chinese, Japanese and Spanish speakers in L2 English article acquisition: Evidence against the fluctuation hypothesis? In *Proceedings of the 8th Generative Approaches to Language Acquisition Conference* (pp. 132–139).
Yang, M. & Ionin, T. (2009). L2 English articles and the computation of uniqueness. In *Proceedings of the 3rd Conference on Generative Approaches to Language Acquisition North America* (pp. 325–335).
Xu, Q., Shi, Y., & Snape, N. (2016). A study on Chinese students' acquisition of English articles and interlanguage syntactic impairement. *Chinese Journal of Applied Linguistics (quarterly), 39*(4), 459–480.
Yamada, K. (2019). Accounting for article interpretation in L2 English by L1 Japanese child learners. *Second Language, 18*, 71–88.

The Acquisition of Aspect in Chinese Based on Learners' L1 Typology: An Analysis Based on the TUFS Co-referential Learner Corpora of Chinese and Japanese

Keiko Mochizuki and Yasuhiro Shirai

Abstract This chapter reports on a bi-directional study that investigated the second language acquisition of telic forms in Chinese and Japanese grammar based on TUFS (Tokyo University of Foreign Studies) co- referential learner corpora of Chinese and Japanese. First, the TUFS Japanese Learner Corpus of Chinese shows that Japanese learners learning Chinese have difficulties acquiring resultative compound verbs expressing telicity and the atelic auxiliary verb "huì". Second, the TUFS Chinese Learner Corpus of Japanese shows that Chinese learners learning Japanese have difficulties acquiring aspectual compound verbs (e.g., inchoative "V-*kakeru*/*kakaru*, - *dasu*"(start to ~)and completive "V- *ageru*/*-agaru*"(complete to~), and overuse of resultative intransitive verbs in transitive/intransitive pairs in Japanese. We claim 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. We also explore effective pedagogy based on learner's native languages.

Keywords Second language acquisition of telicity · Learner error corpora of Japanese and Chinese · Aspectual compound verbs · Resultative compound verbs · Cognition of telicity · Typology of syntactic/lexical structure

List of Abbreviations

ACC	Accusative
ASP	Aspect

K. Mochizuki (✉)
Tokyo University of Foreign Studies, 3-11-1 Asahicho, Fuchu City, Tokyo, Japan
e-mail: mkeiko@tufs.ac.jp

Y. Shirai
Department of Cognitive Science, Case Western Reserve University, 10900 Euclid Ave, Cleveland, OH 44106, USA
e-mail: yasshiraijp@gmail.com

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_11

CAU	Causative suffix
DAT	Dative
DES	Desiderative form
GEN	Genitive
INS	Instrumental
ITS	Intransitive suffix
NEG	Negative
NML	Nominalizer
NOM	Nominative
NONPAST	Nonpast
PAST	Past
PP	Pragmatic particle
PSS	Passive suffix
POL	Polite suffix
QT	Quotative particle
SE	Sentence extender
SFX	Suffix
TOP	Topic marking particle
TRS	Transitive suffix

1 Introduction

The goal of this paper is to analyze how typologically different native languages affect second language acquisition by comparing the output contained in learner corpora of Japanese and Chinese. Expressions of telicity are one of the most difficult grammatical items in second language acquisition. This paper focuses on the second language acquisition of aspectual expressions in Japanese by Chinese speakers and Chinese by Japanese speakers. Next, we will discuss the connection between typological differences in the cognition of telicity and the second language acquisition of Japanese and Chinese.

In the present study, the error-tagged learner corpora of Chinese and Japanese that are publicly available were used, as shown below:

(1) Tokyo University of Foreign Studies International Center of Japan Studies "Learners' Error Corpora of Japanese/English/Chinese Searching Platform" https://corpus.icjs.jp/
(2) Tokyo University of Foreign Studies International Center of Japan Studies "The Learners Language Corpus of Japanese and Online Error Dictionary" http://cblle.tufs.ac.jp/llc/ja_wrong/index.php?m=default
(3) Tokyo University of Foreign Studies International Center of Japan Studies "Online Dictionary of Misused Chinese" https://corpus.icjs.jp/corpus_ch/index.php

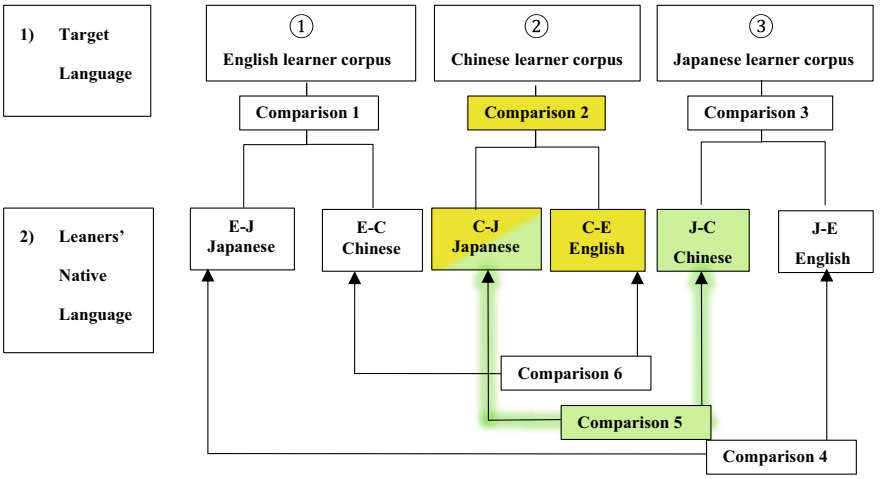


Fig. 1 Patterns of comparison between combinations of target language and Learners' native language

2 Cross-Referentiality of Multi-lingual Learner Corpora

The corpora listed above in (1)–(3) allow comparison of pairs of "target languages" and "learners' native languages" in the six patterns shown in Fig. 1. This allows second language acquisition researchers to investigate how different native languages affect the acquisition of grammatical items.

This paper first considers comparison 5 in Fig. 1. Patterns of comparison between combinations of target language and Learners' native language.

In other words it compares Japanese learners of Chinese in the Chinese learner corpus ② (C-J) with Chinese learners of Japanese in the Japanese learner corpus ③ (J-C), thus constituting a bi-directional study (e.g., Rocca, 2002). The aims of the comparison are to analyze what characteristics of Japanese affect Japanese learners' acquisition of Chinese, and to analyze what characteristics of Chinese affect Chinese learners' acquisition of Japanese. In addition, it also considers comparison 2 in Fig. 1, i.e., a comparison of the production of Japanese and English native speakers in their L2 Chinese corpus.

3 Acquisition of Chinese Lexical Aspect by Japanese Native Speakers

It has long been pointed out in the field of Chinese language education that complements are difficult for Japanese learners of Chinese to acquire.

Mochizuki (2018) presented learner corpus data showing that acquisition of resultative complements is more difficult for Japanese native speakers than for English native speakers. This chapter refines the data and provides further analysis.

Resultative complements used frequently by native speakers of Chinese include <-到 dào>, <-成 chéng> and <-完 wán>, which appear after activity verbs. We compare the production of these three complements by Japanese, English, and Chinese native speakers. Data is drawn from the three corpora below.

The first is the "Learners' Error Corpora of Chinese Searching Platform" containing data from Japanese native speakers learning Chinese at Tokyo University of Foreign Studies.[1] The second is "Corpus of English Native Learners of Chinese"[2] created in collaboration with Taiwan Normal University, and the third is the "Chinese Native Speakers Corpus".[3]

Zhang (2019: 54) uses the three corpora above to report the production of resultative complements, as shown in Table 1.

Table 1 Comparison of production of resultative complements expressing telicity in Chinese (Zhang, 2019:54)

	Japanese native speakers	Chinese native speakers	Statistical test	English native speakers	Chinese native speakers	Statistical Test
Resultative complement	Adjusted frequency (per 100,000 words)		*p* value	Adjusted frequency (per 100,000 words)		*p* value
～到 dào	393.98	3338.78	<0.0001	523.93	3338.78	<0.0001
～成 chéng	58.64	260.35	<0.0001	53.26	260.35	<0.0001
～完 wán	18.32	15.60	0.4736	82.99	15.60	<0.0001
～好 hǎo	15.58	24.84	0.0531	63.17	24.84	<0.0001
～掉 diào	5.50	14.74	0.0121	9.91	14.74	0.2597
～错 cuò	10.99	15.04	0.2779	7.43	15.04	0.0786
～开 kāi	1.83	13.85	0.0007	1.24	13.85	0.0023
～住 zhù	13.74	17.59	0.3408	8.67	17.59	0.0566
～坏 huài	3.66	27.64	<0.0001	2.48	27.64	<0.0001
～满 mǎn	4.58	25.12	<0.0001	7.43	25.12	<0.0001

The three sets of data (L2 Chinese data by L1 Japanese learners, L2 Chinese data by L1 English learners, and comparison data from native speakers) in Table 1 have not been strictly controlled for content, style, or proficiency level, but focusing on <-到 dào >, the complement with the highest frequency in three corpora, Japanese native speakers display an adjusted frequency of 393.98 per 10,000 words, significantly less than the adjusted frequency of 3338.78 by native speakers of Chinese. Table 2. shows X^2 significance testing. Results reveal that English native speakers use "V+到 dào", "V+完 wán" and "V+好 hǎo" significantly more frequently than Japanese native speakers.

We suggest one of the possible factors for this phenomenon in second language acquisition is the word order typology in a word structure. English has SVO word order and "Verb + Resultatives" structure both in syntax and lexical structure as phrasal verb, therefore both English and Chinese have the same word order in lexicon, it would be easier for English L1 learners of Chinese to acquire resultative complements in Chinese. On the other hand, Japanese has SOV word order, there is no SVOC construction, therefore this word order typology might affect the acquisition of "Verb + Resultatives" structure.

Table 2 X2 significance testing

	X^2	P	df	Significance	Corpus with higher frequency
V+到dào	17.23	0.0000	1	**Significant at 0.1% ($\chi 2 = 17.24$, $p = 0.000$)**	**English**
V+成chéng	0.15	0.6962	1	No significant difference ($\chi 2 = 0.15$, $p = 0.696$)	n/a
V+完wán	40.97	0.0000	1	**Significant at 0.1% ($\chi 2 = 40.97$, $p = 0.000$)**	**English**
V+好hǎo	28.05	0.0000	1	**Significant at 0.1% ($\chi 2 = 28.05$, $p = 0.000$)**	**English**
V+掉diào	0.70	0.4029	1	No significant difference ($\chi 2 = 0.70$, $p = 0.403$)	n/a
V+错cuò	0.30	0.5823	1	No significant difference ($\chi 2 = 0.30$, $p = 0.582$)	n/a
V+开kāi	0.00	1.0000	1	No significant difference ($\chi 2 = 0.00$, $p = 1.000$)	n/a
V+住zhù	0.64	0.4239	1	No significant difference ($\chi 2 = 0.64$, $p = 0.424$)	n/a
V+坏huài	0.00	0.9663	1	No significant difference ($\chi 2 = 0.00$, $p = 0.966$)	n/a
V+满mǎn	0.25	0.6158	1	No significant difference ($\chi 2 = 0.25$, $p = 0.616$)	n/a

What are the causes of this phenomenon, whereby Japanese native speakers produce fewer complements than native speakers of Chinese and English? First, we consider some examples from the "Learners' Error Corpora of Chinese Searching Platform". (4), (5), (6) and (7) below are examples of omission of resultative complements by Japanese native speakers. Underlined sections are instances of omission and corrections appear to the right of the arrows.

(4) a. 如果住在车站附近的话，人们→就(add, 关联副词)很容易买→买到(add, 结果补语)→他们(add, 主语)需要的东西。
Rúguǒ zhù zài chēzhàn fùjìn de huà, rénmen → jiù (add, Guānlián fùcí) hěn róngyì mǎi → mǎidào(add, Jiéguǒ bǔyǔ) → tāmen (add, Zhǔyǔ) xūyào de dōngxi.
b. If you live near the station, you can buy things you need easily.
(TUFS_CH_043: Tokyo University of Foreign Studies 2nd year. Length of study: 13 months)

(5) a. 有 (add, 动词 存现动词)学者说，因为城市生活→里(replace, 名词 短语 方位词 方位短语)有很多人，有→(delete, 动词 存现动词)很多噪音，还有有很多压力，所以脑子→ 精神(replace, 名词)→会(add, 动词 能愿動詞)受→受到(add, 结果补语)更大的影响。
Yǒu (add, Dòngcí, Cúnxiàn dòngcí) xuézhě shuō, yīnwèi chéngshì shēnghuó → lǐ (replace, Míngcí duǎnyǔ, Fāngwèicí, Fāngwèi duǎnyǔ) yǒu hěn duō rén, yǒu → (delete, Dòngcí, Cúnxiàn dòngcí) hěn duō zàoyīn, háiyǒu yǒu hěn duō yālì, suǒyǐ nǎozi → jīngshén (replace, Míngcí) → huì (add, Dòngcí, Néngyuàn dòngcí) shòu → shòudào (add, Jiéguǒ bǔyǔ) gèng dà de yǐngxiǎng.
b. One scholar says that living in the city there are lots of people, it's noisy and stressful, and it greatly affects them mentally.
(TUFS_CH_051: Tokyo University of Foreign Studies 2nd year. Length of study: 13 months)

(6) a. 所以(add, 连词)自然的→地(replace, 结构助词)人→人们(replace, 名词)→都(add, 范围副词)汇集→于此(add, 介词短语补语)，就业机会也很多，并且好大学→好的大学(replace,“的”字短语)一般→都(add, 范围副词)在城市，所以农村的青年人→年轻人(replace, 名词)希望去城市受→接受(replace, 动词)良好的教育，找 → **找 到(add, 结果补语)** 好(add, 形容词)工作。

Suǒyǐ (add, Liáncí) zìrán de → de (replace, Jiégòu zhùcí) rén → rénmen (replace, Míngcí) → dōu (add, Fànwéi fùcí) huìjí → yúcǐ (add, Jiècí duǎnyǔ bǔyǔ), jiùyè jīhuì yě hěn duō, bìngqiě hǎo dàxué → hǎo de dàxué (replace, “de” zì duǎnyǔ) yìbān → dōu (add, Fànwéi fùcí) zài chéngshì, suǒyǐ nóngcūn de qīngniánrén → niánqīngrén (replace, Míngcí) xīwàng qù chéngshì shòu → jiēshòu (replace, Dòngcí) liánghǎo de jiàoyù，zhǎo → zhǎodào (add, Jiéguǒ bǔyǔ) hǎo (add, Xíngróngcí) gōngzuò.

b. For this reason, people naturally gather in cities. Cities have a lot of opportunities to find a job and famous universities are located in cities. That's why young people hope to go a city, get a good education and find work. (TUFS_CH_056：Tokyo University of Foreign Studies 2nd year. Length of study: 36 months, Chinese Proficiency Test Level 3 (2012.3))

(7) a. 我现在在东京→(delete move, 补语 介词短语补语)一个人住→在东京(add move, 补语 介词短语补语)，所以不能跟我父母在一起，也不能直接孝敬他们。我唯一能做的事→就(add, 助词 语气助词)→是(add, 动词 关系动词)多给他们打电话，让他们**听 → 听 到(add, 结果补语)**我精神饱满的声音。

Wǒ xiànzài zài Dōngjīng → (delete move, Bǔyǔ, Jiècí duǎnyǔ bǔyǔ) yí ge rén zhù → zài Dōngjīng (add move, Bǔyǔ, Jiècí duǎnyǔ bǔyǔ), suǒyǐ bù néng gēn wǒ fùmǔ zài yìqǐ, yě bù néng zhíjiē xiàojìng tāmen. Wǒ wéiyī néng zuò de shì → jiù (add, Zhùcí, Yǔqì zhùcí) → shì (add, Dòngcí, Guānxi dòngcí) duō gěi tāmen dǎ diànhuà, ràng tāmen tīng → tīngdào (add, Jiéguǒ bǔyǔ) wǒ jīngshén bǎomǎn de shēngyīn.

b. Now I'm living on my own in Tokyo, so I can't be with my parents or look after them either. The only thing I can do is call frequently and let them hear I'm okay. (TUFS_CH_013: Tokyo University of Foreign Studies 3rd year. Length of study: 38 months)

The errors in (4)–(7) are ungrammatical because <-到 dào> is missing. Activity verbs in Chinese (or in any language) are atelic and cannot express change of state (i.e., telicity), and in Chinese activity predicates can convey telicity with the addition of the resultative complement (e.g., Li & Shirai, 2000; Li, 1990; Mochizuki, 2007), in this case with the addition of <-到dào> combined with a preceding activity verb turns it into an achievement verb.

Second, Japanese native speakers show notable omission of 〈–到 dào〉. The examples above show that the acquisition of the aspectual process in Chinese of "attaching <-到dào> to a verb to form an achievement verb expressing telicity" is difficult for Japanese native speakers.

Why is the acquisition of <-到dào> difficult for Japanese native speakers? The reason why it is difficult for Japanese L1 learners to acquire is that in Japanese, telicity is expressed in a different form. In Japanese, finality is expressed by an intransitive verb expression among its intransitive and transitive verb counterparts. The examples in (8)–(10) show that in many cases, Japanese forms corresponding to Chinese resultative complements are intransitive verbs forming transitive and intransitive pairs (for the details, see Mochizuki, 2004).

(8)	a. transitive	「見る mi-φ-ru」	看 kàn
	b. intransitive	「見える mi-**e**-ru」	看**到** kàn **dào** /看**见** kàn **jiàn**
(9)	a. transitive	「貯める tam-**e**-ru」	存 cún
	b. intransitive	「貯まる tam-**ar**-u」	存**满** cún **mǎn**
(10)	a. transitive	「直す nao-**s**-u」	修 xiū
	b. intransitive	「直る nao-**r**-u」	修**好** xiū **hǎo**

Transitivity alternations in Japanese are expressed by semi-productive verbal affixes (Jacobsen, 1992; Mochizuki, 2004). Transitive verbs in (8)–(10) are activity verbs and their corresponding intransitive verbs are achievement verbs, which are telic.

In contrast, in Chinese an "adjective expression resultativity" is attached to an activity verb and guarantees its telicity (Tai, 1984, 1985). In other words, telicity is expressed by verbal affixes in Japanese and by adjectives following a verb (in the form of an RVC) in Chinese. This difference between the forms for expressing telicity—an affix forming one part of a word vs. an adjective—may be exerting a cognitive burden on Japanese native speakers and making acquisition more difficult.

A similar phenomenon is observed in the second language acquisition of Japanese. Regardless of learners' native language, errors involving transitivity are frequently observed and acquisition is considered to be difficult.

First, comparing transitivity alternations in Japanese to Chinese, transitive verbs and their intransitive pairs often correspond to "activity verbs" and "activity verb + resultative complement" respectively.

Mochizuki (2009: 91) observed the following phenomenon in a writing class of 2nd year Japanese majors (advanced B2 level learners who have passed level 1 of the Japanese Language Proficiency Test) in the foreign languages department at Tokyo University of Foreign Studies. Advanced level Japanese learners whose native language is Chinese frequently used the intransitive variant of a pair when the transitive variant was appropriate. This observation led to the hypothesis below.

(11) "In Chinese native speakers' interlanguage system transitive/intransitive pairs of Japanese, the transitive form expresses an atelic action verb and the intransitive form expresses a telic action verb".

The hypothesis in (11) suggests that Chinese native speakers construct the interlanguage system such that "Japanese intransitive verbs are telic, so choose an intransitive variant to express a result state". This is indeed what is found in the high-level learners' corpus. Examples of errors cited in Mochizuki (2009: 91–92) are shown in (12).

(12)

a. 現在 夏 に も ビジネスマン が 変わらない スーツ を 着て,
Genzai natsu ni mo bijinesu-man ga kawaranai suutsu o kite ,
now summer DAT also businessman NOM change-NEG suit ACC wear

仕事 を する が, 熱い 季節 が 人 の 気持 を
shigoto o suru ga , atui kisetsu ga hito no kimochi o
work ACC do but hot season NOM person GEN feeling ACC

変わって (→変えて) しまう。 (CC604_2)
kaw- **a**- tte (→ka- **e**- te) shi- ma- u .
change-ITS:TE (→change-TRS:TE) do-perfective-NONPAST

b. 最近, 夏季 軽装
Saikin , kaki-keesoo
in these days summer-casual wear

に 変わる (→変える)
ni kaw-**ar**-u (→ka -**e**- ru)
DAT change-ITS -NONPAST (→change-TRS -NONPAST)

こと は 環境 に も よい 影響 を あたえる。 (CC604_2)
koto wa kankyo ni mo yoi ei-kyoo o atae - ru .
NML TOP circumstance DAT also good effect ACC give-NONPAST

c. 話し方 に 断続性 を 感じ,
Hanashikata ni danzoku-sei o kanji,
the way ofspeaking DAT intermittency ACC feel

つながってあって (→つなげて) ほしいな と。(CC702)
tsunag-**a**-tte-atte (→tsunag-**e**-te) hoshii-na to .
connect-ITS:TE-ASP:TE (→ connect-TRS:TE) want to NML

d. 聴衆　　を　　見ていて　　伝わりたい　　（→伝えたい）
Chooshuu　o　　miteite　　tsutaw-**ar**i-tai　　(→tsuta-**e**-tai)
audience　ACC　look at　　tell-ITS -DES　　(→ tell-TRS -DES)

と　いう　　感じ　　が　　ありました。　　(HC 703)
to -yuu　　kanji　　ga　　arimasita.
QT　　feeling　NOM　be-POL-PST

e. ゆめ　を　かなう　　（→かなえる）　　こと　は,
Yume　o　kana-**u**　　(→kana-**e**-ru)　　koto　wa
dream　ACC　come true-NONPAST　(→come true-TRS -NONPAST)　NML　TOP

人間　　に　　とって,　　これ以上　　の
ningen　　ni　　totte,　　koreijoo　　no
human being　DAT　regarding　no more　GEN

喜び　　は　　ないでしょう。(CC703_1)
yorokobi　　wa　　nai-deshoo.
happiness　TOP　NEG-POL-SFX

Learners also use intransitive rather than the appropriate transitive verbs as V2 in compound verbs, producing nonexistent compound verbs.

(13)

a. 主人公 は, 2 秒 ほど 新聞 を
Shujinkoo wa , ni byoo hodo shinbun o
main character TOP 2 second about newspaper ACC

読み続いている (→続けている) とき,
yomi-tsuzu-**i**-tei-ru (→tsuzu-**ke**-tei-ru) toki,
read-continue: TE-ASP-NONPAST(→continue-TRS:TE-ASP-NONPAST) when

読売 新聞 の ロゴ が 浮かび 上がる。(CC808_4)
Yomiuri shinbun no rogo ga ukabi- aga- ru .
The Yomiuri Shinbun GEN logo NOM float-up-NONPAST

b. 大手, 小手 の メーカー は ずいぶん 工夫し,
Oo te , shoo te no meekaa wa zuibun kufuu-shi ,
big company small company GEN manufacture TOP very ingenuity-do

コマーシャル を 作りつづいている (→続けている)。 (CC808_4)
komaasharu o tsukuri-tsuzu-**i** - teiru (→tsuzuk -**e**- tei ru)
commercial ACC make-carry on: TE-ASP-NONPAST (→carry on-TRS:TE-ASP-NONPAST)

c. 私 は 十八 年 ずっと この 町 の
Watashi wa juuhachi nen zutto kono machi no
I TOP 18 year every time this town GEN
発展 を 見届いて (→見届けて) きた。(CC804_6)
hattenn o mi-todo-**i**te (→mi-todok-**e**-te) kita .
development ACC see through:TE (→see through-TRS:TE) come-PAST

d. 多様化 した 言語 の 学習 と 使用 が
Tayoo-ka-shita gengo no gakushuu to shiyoo ga
diversification-do-PAST language GEN study and use NOM

新しい アイデンティティー を 生まれだす (→ 生み出す) こと に
atarashii aidenthithii o um-**are**-dasu (→um-**i**-dasu) koto ni
new identity ACC born-ITS-out (→born-out) NML DAT

つながる のである。 (CC811_5)
tsunag-ar-u no -de -aru .
link SE- NONPAST

What is the cause of this phenomena, whereby learners use an intransitive variant instead of a transitive variant? It is connected with word formation rules in Chinese verbs expressing resultativity shown in (14).

(14)

a. Event structure and predicate combination in Chinese resultative compound verbs

	V1	V2
1) event	causing event or preceding event	result event
2) verb	transitive verb / unergative verb / unaccusative verb	unergative verb / adjective

b. 打破 dǎ-pò （apply force and break something / and it breaks）
摔坏 shuāi-huài （drop something at high speed and break it/ and it breaks）
喊哑 hǎn-yǎ （shout too much and make yourself hoarse / and become hoarse）
累坏 lèi-huài （get tired and harm your health / and your health is harmed）

As shown in (14a), in Chinese there is no "transitivity harmony principle" (Kageyama, 1993) like that which restricts combinations of transitive and intransitive verbs in Japanese. As shown in (14b), many examples of Chinese compound verbs contain unergative verbs with no morphological distinction between transitive and intransitive. The rules for Chinese compound verb formation in (14a) and the interlanguage rule "choose an intransitive variant to express resultativity" may have produced erroneous Japanese compound verbs like "*-o yomi-tsuzuku*","*-o tukuri-tsuzuku*", "*-o mi-todoku*" and "*-o omoi-ukabu*".

As demonstrated above, Chinese native speakers erroneously select intransitive expressions to express "resultativity" in transitive/intransitive pairs of Japanese verbs. The root of this interlanguage is the effect of native language—the cognitive salience of resultative complements in Chinese.

Regarding Japanese native speakers' difficulty acquiring Chinese expressions for telicity/atelicity, it appears that learners have not fully acquired the grammaticalized aspectual meaning of <-到 dào>.

Instead, they memorize expressions with resultative and potential complements they have already learned and as a result, frequently select semantically inappropriate complements. Chinese resultative complements should not be taught to native speakers of Japanese as vocabulary items to memorize.

Instead, it would be effective to teach using "the concept-based language instruction method" Lantolf et al. (2021) propose. It is necessary to teach learners the basic functions of resultative complements, then how each complements combine productively with V1.

In Sect. 3, we will focus on aspectual compound verbs in Japanese and Chinese, comparing the two systems and their acquisition.

4 Aspectual Compound Verbs in Japanese and Chinese

Kageyama (2013: 11) defines aspectual compound verbs as follows:

(15) The argument structure of the sentence is basically determined by V1. V2 expresses lexical aspect in a wide sense, describing the unfolding situation expressed by V1.

a. complement type:

1) *ageru/agaru* (completion): *utai-ageru* "sing to the end"
migaki-ageru "finish by polishing"

2) *nogasu* (incompletion): *mi-nogasu* "miss a chance to see" or "turn a blind eye"

b. adverbial type:

3) *wataru* (all over): *hibiki-wataru* "reverberate to every corner"
hare-wataru "the entire sky becomes clear"

Chinese also has aspectual compound verbs that correspond to Japanese aspectual compound verbs.

(16) a. 〈−好 hǎo〉: perfective + resultant product "*-ageru* / *-agaru*"

1) 写好论文 xiě hǎo lùn wén
論文を書き上げた ronbun-wo kaki-ag-e-ta
wrote up a report
2) 蛋糕烤好了 dàn gāo kǎo hǎo le
ケーキが焼き上がった kēki-ga yaki-ag-a-tta
the cake is ready

b. 〈−上 $_1$ shàng〉: inchoative "*-dasu*"

1) 会议还没开大家就议论上了。Huì yì hái méi kāi dà jiā jiù yì lùn shàng le.
会議開始前に皆は討論しだした。
Kaigi kaishi-mae-ni mina-wa touron-shi-dash-i-ta
Everyone started the discussion before the meeting began.
2) 最近又忙上了。Zuì jìn yòu máng shàng le.
最近また忙しくなり出した。
Saikin mata isogashi-ku nari-dash-i-ta
I've got busy again recently.

c. 〈−光 guāng〉: perfective + disappearance of the Theme "*-kiru*"

1) 吃光 chī guāng 食べ切る tabe-kiru eat everything
2) 花光 huā guāng お金を使い切る okane-o tsukai-kiru use up
3) 卖光 mài guāng 売り切る/売り切れる uri-kiru / uri-kir-e-ru sell out

d. 〈−住 zhù〉: perfective + fixation "*-tomeru*", "*-komu*"

1) 叫住了一辆出租车 jiào zhù le yī liàng chū zū chē
タクシーを呼び止める Takushī-o yobi-tom-e-ru flag down a taxi
2) 记住 jì zhù
覚えこむ oboe-komu memorize

e. 〈−上$_2$ shàng〉: perfective + attachment "*-tsukeru* / *-tsuku*$_1$"

1) 贴上 tiē shàng 貼り付ける hari-tsuk-e-ru stick to
2) 装上 zhuāng shàng 取り付ける tori-tsuk-e-ru attach

f. 〈−着 zháo〉: perfective + achievement of goal: "*-tsuku*$_2$", "*-dasu*", "*-ateru*"

1) 睡着了 shuì zháo le 寝つく ne-tsuku get to sleep
2) 找着了 zhǎo zháo le 探し出す sagashi-dasu ; 探し当てる sagashi-ateru find

Chinese also shows prominent meaning expansion from directional complements expressing movement to resultative complements expressing aspect, as shown in (17).

(17) a. 〈−起来 qǐ lái〉 "upward movement" to "inchoative" "*-dasu*"

b. 〈−下去 xià qù〉 "downward movement" to "durative" "*–tsuzuku*, / *-tsuzukeru*"

5 Acquisition of Japanese Compound Verbs: *-kakaru/-kakeru*

Tamaoka and Chu (2013: 415–426) carried out a survey with learners of Japanese at a Chinese university concerning the acquisition of the compound verbs "*–agaru/-ageru*, *-kakaru/-kakeru*" and "*- hairu/-ireru"*. They concluded that lexical compound verbs expressing abstract meaning and with a V1 whose meaning was difficult to understand were most difficult to acquire. This was the case regardless of transitivity or of learners' length of study. Table 3 below shows an interesting phenomenon that appears in a table in Tamaoka and Chu (2013: 416) titled "percentage of correct answers for lexical compound verbs by learners with 1 year of study and learners with 2 years of study". The percentage of the compound verbs "*–kakaru/-kakeru*" expressing inchoative meaning is the lowest.

The polysemous compound verbs *-kakaru/-kakeru* have long been said to be particularly difficult for learners of Japanese. Particularly interesting is the observation by Tamaoka and Chu (2013: 416) that for Chinese learners of Japanese, percentages of correct answers for aspectual uses expressing inchoative meaning, like the examples in Table 2, are much lower than for those expressing concrete meaning, like *furi-kakaru* "pour down", *osoi-kakaru* "attack", *fuki-kakeru* "blow on" and *kise-kakeru* "drape over". What is the cause of this phenomenon?

We suggest that Chinese compound verbs "V1 + complement" do not express inchoative meaning, so there are no Chinese compound verbs equivalent to *–kakaru/-kakeru.* For example, *shini-kakaru* "nearly die" is expressed as 快 kuài 要 yào 死 sǐ > "nearly die". As the meaning is atelic, it cannot be expressed by resultative complements, i.e., aspectual compound verbs in Chinese.

Aspectual verbs like *yomi-kakeru* "start reading", *yomi-sasu* "read halfway" and *yomi-tsuzukeru* "continue reading", which do not express telicity, cannot be expressed by Chinese compound verbs and can only be expressed using verb phrases expressing inchoative, incomplete or continuative meaning. This is because Chinese aspectual compound verbs are typically resultative compound verbs and operate under a "telicity principle" stating that V2 must be a predicate expressing a result.

Table 3 Percentage of correct answers by learners for the inchoative aspectual compound verbs "-kakaru/-kakeru"(Tamaoka & Chu, 2013: 416)

	Learners with 1 year of study Percentage of correct answers (%)	Learners with 2 years of study Percentage of correct answers (%)
(1) V1 easy: *shini-kakaru* "nearly die"	49.2	70.9
(2) V1 difficult: *kuzure-kakaru* "nearly collapse"	60.0	58.3
(3) V1 easy: *nomi-kakeru* "about to start drinking"	64.6	73.8
(4) V1 difficult: *taore-kakeru* "nearly fall over"	38.5	57.3

6 Acquisition of Japanese Compound Verbs by Chinese Learners: -agaru/-ageru

While Chinese possesses resultative complement <-上 shàng> using the same Chinese character as *-agaru*/*-ageru*, the uses of the two differ, making acquisition difficult. The Japanese learner corpus (1) contains a Chinese-to-Japanese translation task[1] performed by L1 Chinese learners, collected with the cooperation of Shanghai International Studies University (SISU). The task displays the following non- use of aspectual compound verbs.

(18) レストランでちょうど蒸しまった→（蒸し上がった **mushi-aga-tta**）八宝飯を見たら、必ず注目します。 (SISU_010_Task_01)
When I see the freshly steamed "babaofan" in a restaurant, I always pay attention to it.

(19) 胡先生一家は、突然に出現した私に「ちょうど蒸した→（蒸し上がった **mushi-aga-tta**）八宝飯がありますので、食べた後帰りましょうね。」と言いました。 (SISU_043_Task_01)
Professor Hu's family said to me, the sudden visitor, "there's some freshly steamed "babaofan", so have some before you go".

The corrected expressions in (18) and (19) correspond to the aspectual compound verb < V1 ＋好 hǎo > shown in (20).

(20) < -好 hǎo > completion of action ＋ favorable result state/creation of result product: *-agaru/-ageru*
写好论文 (論文を書いて 仕上げた) finish writing a paper
蛋糕烤好了 (ケーキが焼けた) a cake finishes baking

When teaching Japanese compound verbs to native speakers of Chinese, it may be effective to compare compound verbs with resultative complements, like the correspondence between *–agaru*/- *ageru* and < -好 hǎo > in (20).

7 Acquisition of Chinese Compound Verbs by L1 Japanese Learners: <V1+ 上 shàng>

Chinese compound verbs of the form <V1＋上 shàng> are highly polysemous and their acquisition is difficult for Japanese native speakers. Similarly to the instruction of English prepositions and phrasal verbs, explanation with the concepts UP, ON, and WITH in a cognitive linguistics framework, as shown in (21) and (22), may be effective.

[1] As for the translation task from Chinese to Japanese, refer to the appendix in Chap. 4.

(21) <-上1 shàng > : result of action + reach {higher location/goal/level} UP concept:

a. movement upward:

登上山顶（山頂に登る）climb to the top of a mountain,

爬上八楼（８階まで登る）climb to the eighth floor

b. achievement of goal:

买上房子（家を買って手に入れる）

住上新房子（新しい家に住めるようになる） buy/get a house

c. achievement of level expressed by quantifier:

最近失眠，每天只能睡上三四个小时。

（最近は不眠で、毎日 3, 4 時間しか眠れない）

I have insomnia recently and can only sleep 3 or 4 hours a day.

只要中午睡上一刻钟，下午工作就特别有精神。

（お昼に 15 分眠れば、午後は仕事の効率がよい）

If you take a 15-minute nap around midday, you will work more efficiently in the afternoon.

(22) <-上２ shàng > : result of an action on “a flat surface + object ” comes into existence on the surface concept:

a. Fixation: *-tsukeru* / *-tsuku*:

1) 贴上 （貼り付ける）stick on
2) 写上姓名（氏名を書き込む/書き入れる）fill in your name
3) 穿上衣服（衣服をつける）put on clothes
4) 戴上{眼镜/帽子/手套}（{眼鏡/帽子/手袋}を身に着ける）
put on glasses/a hat/gloves

b. Two objects joined/touching:

WITH concept:

1) 关上门 （ドアを閉める）close the door
2) 锁上 （鍵をかける） lock
3) 接触上了 （連絡がついた/接触ができた） make contact with

c. Change of state, inchoative + durative: ～し始める/～になる

Inchoative ON concept:

1) 会议还没开大家就议论上了。
（会議開始前に皆はすでに討論し始めた）
Everyone started the discussion before the meeting began.
2) 最近又忙上了。（最近また忙しくなった）I’ve got busy again recently.
3) 爱上一个女演员（女優を好きになる）become infatuated with an actress

The Tokyo University of Foreign Studies International Center of Japan Studies “Online Dictionary of Misused Chinese” (https://corpus.icjs.jp/) referred to above contains instances of omission of <V1＋上 shàng> by native speakers of Japanese. Errors below are underlined.

(23) 但是农村的人体验→了(add，态助词)城市→的生活(add，宾语)的话→之后(replace，方位词)，他们一定→会(add，能愿动词)喜欢了→喜欢上(replace，时态助词 趋向补语)这个→种(replace，量词)方便→的(add，“的”字短语 结构助词)生活。

(TUFS_CH_053/ Tokyo University of Foreign Studies 2nd year. Length of study: 13 months)

(24) 因为在那里土地和房租很便宜，我们花很少→的(add，短语 助词 “的”字短语 结构助词)钱→就(add，副词 关联副词)可以住在→住上(replace，补语 介词短语补语)良→很(replace，副词 程度副词)好的房子。

(TUFS_CH_074/ Tokyo University of Foreign Studies 2nd year. Length of study: 13 months)

(25) 我开始缝纫的→(delete，结构助词)开端→(delete，名词)是→因为(add，连词)我想给迪斯尼的布制玩偶穿→穿上(add，补语 趋向补语)衣服，但是它的衣服很贵，所以我→就(add，副词 关联副词)挑战自己做，给它做浴衣。

(TUFS_CH_083/ Tokyo University of Foreign Studies 3rd year. B1 level Length of study: 26 months)

Instruction of <V1 ＋上 shàng> will likely benefit from reference to the spatial concepts expressed by English prepositions, as are used in the instruction of prepositions and phrasal verbs.

8 Underuse of the Atelic Auxiliary Verb <会 huì> by Japanese Learners of Chinese

While native speakers of Japanese have difficulty acquiring resultative complements to express the realization of events, we also observe the phenomenon of omission of the atelic auxiliary verb <会 huì> (shown as [φ→会]). This is prevalent even among learners of the highest proficiency, suggesting the difficulty of acquisition for native speakers of Japanese.

(26) 买东西的时候，我不(φ→会 huì)货比三家。
私は買い物するとき、あちこち値段を見比べたりしない。
When I go shopping, I don't go around comparing prices.
(TUFS_CH_027/ Tokyo University of Foreign Studies 4th year. Length of study: 37months)

(27) 我跟妈妈一起去公园的时候，妈妈（φ→会 huì）坐在公园的椅子上等着我练习。
私が母と公園に行くときは、母はいつもベンチに座って、練習が終わるまで付き合ってくれる。
When I go to the park with my mum, she always sits on a bench and waits for me until practice is over.
(TUFS_CH_027/ Tokyo University of Foreign Studies 4th year.
Length of study: 48 months)

The learners who wrote (26) and (27) are both 4th year Chinese majors with study abroad experience and over 4 years of Chinese language study. Omission of the atelic probability auxiliary verb <会 huì> still occurs for these and other learners at the highest of proficiency levels.

Zhang (2017: 22) reports that among instances of the atelic probability auxiliary verb <会 huì> in the Tokyo University of Foreign Studies Japanese Learners of Chinese Corpus, 4.1% are correct and 95.9% are errors (83.6% omission, 11.3% replace, 1.9% word order). In contrast, in the English Learners of Chinese Corpus, 93.1% of instances are correct, and errors (3.7% omission, 2.4% replace, 0.8% word order) account for only 6.3% of uses (Zhang 2017: 20). Errors of omission, which account for 83.6% of errors by Japanese native speakers, do not appear in the data of English native speakers.

Native speakers of English and Japanese thus display contrasting trends in the production of the auxiliary verb <会 huì>. Again, the influence of Japanese is expected to be the cause. The tense system in Japanese consists only of the non-past marker "*–ru*" and the past marker "*–ta*". The non-past marker "*–ru*" is also used to express atelic events, as there is no tense or aspectual form dedicated to atelic meaning. Such differences between the tense and aspect systems in Japanese and Chinese are presumed to be the cause of the difficulty in the acquisition of the atelic probability auxiliary verb <会 huì>.

9 Differences Between Forms in Chinese and Japanese Expressing Telic and Atelic Events

The expression of telic events in Chinese often requires an aspectual compound verb (resultative complement), and the expression of atelic events often requires the atelic auxiliary verb <会 huì>, but native speakers of Japanese have difficulty acquiring both. This is because Japanese uses the non-past form "*–ru*" for atelic events, which is an unmarked form. In Japanese, the prominent conceptual distinction is between past and non-past, and as shown in (26) and (27), "general events" are marked with "*–ru*".

The prominent conceptual distinction in Chinese is not tense but between "telic and atelic", which are connected with aspect and modality. Resultative complements are often used to express telic events. For negation of telic events, the negative form <不 bù> is inserted between V1 and V2 of a resultative compound verb in a so-called "potential complement", such as <找不到> "search for but unable to find", <写不出来> "try to write, but unable to do so" and <睡不着> "try to sleep but unable to fall asleep".

In contrast, Japanese uses transitive alternations (e.g., verb pairs like *ueru 'plant.tr'* and *uwaru 'be planted'*, *kiru 'cut'* and *kireru 'get cut'*, *naosu 'fix* and *naoru 'get fixed'*) and aspectual compound verbs (e.g., *yomi-komu 'read to the full'*, *kaki-dasu* 'start writing' and *kaki-ageru 'complete writing'*) to express the telicity of events.

Next to consider is why, despite aspectual compound verbs existing in both Japanese and Chinese, native speakers of one language struggle to acquire aspectual compound verbs when learning the other. The reason can be judged to stem from the syntactic differences between Chinese, an SVO language, and Japanese, an SOV language. These differences are reflected in the lexical structure of aspectual compound verbs. The next section will examine how syntactic differences between Japanese and Chinese lead to differences in aspectual compound verbs in the two languages.

10 Differences in the Structure of Japanese and Chinese Aspectual Compound Verbs

Aoki (2013) suggests that Japanese aspectual compound verbs developed in the form $_{VP}$[$_{VP}$[argument + V1]V2]], with the object complement of V2 compounding with V1. This historical change corroborates the fact that the SOV syntactic structure of Japanese is reflected in the internal structure of compound verbs and the prominence of compound verbs with an OV structure, where V1 functions as the object of V2.

Mochizuki and Shen (2011) state that in Chinese, with an SVO word order, verb compounding with an object complement does not occur. In Chinese, SVOC resultatives changed diachronically to "S[VC]O" constructions, resulting in the prominence of resultative compound verbs. Therefore, in Chinese atelic events like "inception" (-*kakeru*, -*dasu*, -*hajimeru*), "continuation" (-*makuru*), and "incompletion" (- *sokonau*, -*sonjiru*, -*wasureru*) cannot be expressed using compound verbs.

The internal structure of aspectual compound verbs in both Japanese and Chinese reflects the syntactic structure of each language. Japanese compound verbs are formed from compounding of the type "V1 inside object complement of V2 + V2", whereas Chinese compound verbs are formed from the VC (=V1 + resultative complement) structure.

In Chinese, "object complement" type compound verbs expressing inception (*-kakeru*, *-dasu*, *- hajimeru*), continuation (*-makuru*, *-tsuzukeru*), incompletion (*-sokonau*, *-sonjiru*, *-sobireru*, *-shikaneru*, *- okureru*, *-wasureru*), excessive action (*-sugiru*), repetition (*-naosu*), and reciprocal action (*-au*) do not exist. Unlike Japanese, Chinese does not allow compounding with an object clause.

The examples in (28) show how Japanese object complement type compound verbs are expressed in Chinese. For example, object complement type compound verbs expressing inception, continuation, incompletion, and repetition are expressed in Chinese using a verb phrase with an object clause of the structure [$_{VP}$V2 + [$_{IP}$ …V1…]], or with transitive sentences expressing the impossibility of past events with the structure [$_{IP}$ 没能méi néng [$_{VP}$…V1…]]. Aspectual compound verbs cannot be used to express these meanings.

(28) a.	「～始める」-hajimeru	([$_{VP}$ 开始 kāishǐ [$_{IP}$…V1 …]])	"start V1~ing"
b.	「～続ける」-tuzukeru	([$_{VP}$ 继续 jìxù [$_{IP}$…V1 …]])	"continue V1~ing"
c.	「～損なう/損ねる」-skonau / -sokoneru		
		([$_{IP}$ 没能 méi néng [$_{IP}$…V1 …]])	"fail to V1"
d.	「～忘れる」-wasureru	([$_{VP}$ 忘 wàng [$_{IP}$…V1 …]])	"forget to V1"
e.	「～直す」 -naosu	([$_{VP}$ 重新 chóng xīn [$_{IP}$…V1 …]])	"V1 again"

Why, then, do object complement type compound verbs not exist in Chinese? For example, why is the compound verb 忘写 wang-xie, forget-write > , corresponding to "*kaki-wasureru"* "forget to write" in Japanese, impossible in Chinese? The reason is that in Chinese "resultative" or "antecedent-result" type compound verbs, which follow the "temporal order" of events, are most favored.

There are compound verbs like <-上 shàng>, <-起 qǐ > and <-开 kāi > expressing inception that initially appear to be exceptions to the rule. These are aspectual compound verbs expressing "inchoative + durative", but in the field of Chinese linguistics are categorized as resultative complements.

However, these "inchoative + durative" type examples <下雨了 xià yǔ le > <哭起来了 kū qǐ lái le > are regarded as aspectually "bounded" changes of state in Chinese. This is connected to the fact that the perfective aspectual marker <-了le> is also used to express prospective aspect as a "perfect of persistent situation" (Comrie, 1976: 19–20, Mochizuki, 1997: 63–64). In other words, as shown in Fig. 2, in Chinese the aspectual marker <-了 le> <-起 qǐ> are used to express a "bounded" event and its result.

In Japanese too, past tense form "*-ta*" can be used to express a bounded change of state such as *hashit-ta*! "(someone) runs!" in instantaneous present uses during live sports coverage and *at-ta*! "there it is!" uttered on the discovery of an item being searched for.

The differences between Japanese and Chinese compound verbs can be summarized in Table 4, which is based on Mochizuki and Shen (2011). The reason native speakers of Japanese and Chinese have difficulty acquiring aspectual compound verbs in their respective target languages, we argue, is because syntactic differences between Japanese and Chinese are reflected in the structure of compound verbs.

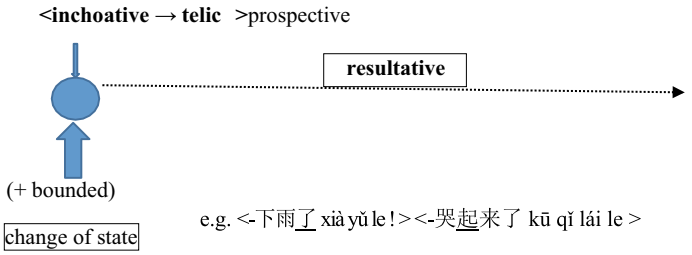


Fig. 2 Inchoative aspect in Chinese

Table 4 Differences in phrase structure, compound verb structure, and word order in Japanese and Chinese

	Japanese	Chinese
1. Structure of verb phrase	1. OV word order 2. verb phrase structure: Verb phrase head: right-sided Resultative predicates are not permitted after the verb	1. VO word order 2. verb phrase structure: Verb phrase head: left-sided Resultative predicates are placed after the verb
2. Preferred word order in compound verb	1. Matching phrase structure and Word structure principle →prominence of **object complement type compound verbs** e.g. (1) "forget to write" *kaki-wasureru* write-forget (2) "too early" *haya-sugiru* early-exceed the standard 2. Temporal Order Principle e.g. relation of causation: *obore-shinu* "drown" antecedent-result relation: *tabe-nokosu* "leave some food left uneaten" *ure-nokoru* "remain unsold"	1. "as a rule, word structure matches the temporal order of events" →prominence of **resultative compound verbs** e.g. <吃腻chī nì > <穿惯 chuān guàn> → complement type compound verbs do not exist <*腻吃> "get sick of eating" <*惯穿> "get used to wearing" 2. Matching Phrase Structure and Word Structure Principle → while object complement type compound verbs do not exist, subject complement type compound verbs do exist e.g. <-完>, <-上>, <-错>, <-多>, <-少>, <-遍>

11 Differences in the "Boundedness" of Temporal Cognition in Japanese and Chinese and Their Effects on Acquisition

As a rule, Chinese aspectual compound verbs are resultative compound verbs expressing telic events.

In contrast, the basic structure of Japanese aspectual compound verbs is the object complement type, and there are compound verbs like *–kakeru* and *–kakaru* that express atelic prospective aspect. As Tamaoka and Chu (2013: 416) have shown, the aspectual compound verbs *–kakeru* and *–kakaru* expressing atelic prospective aspect are difficult for Chinese learners of Japanese to acquire.

In contrast, for Japanese learners of Chinese, production of the resultative complements <-到 dao >, <-成 cheng> and <-完 wan>, which make the distinction in "boundedness" between telic and atelic, is much lower than native English learners of Chinese and native speakers of Chinese, as discussed earlier in Sect. 2 in this chapter. In addition, there is a strong possibility that even among highly proficient learners, omission of the auxiliary verb <会 huì> expressing probability with atelic events will remain as an "eternal error" through fossilization.

The difficulties of acquiring tense and aspect observed in the learner corpora of Japanese and Chinese are presumably influenced by the typological differences concerning the "boundedness of temporal cognition" in Japanese and Chinese. These differences can be summarized as below.

As shown in Table 5, Chinese shows a "prominence of bounded cognition". As a result, grammatical forms like the telic aspectual marker "-了le", resultative complements, and the atelic auxiliary verb <会 huì> mark whether an event is telic or atelic.

In contrast, Japanese shows a "lack of prominence of bounded cognition". While there are forms to distinguish past and non-past, grammatical forms marking telicity or atelicity are not prominent. The *"–te iru"* form is used for both atelic progressive aspect and telic resultative aspect. Japanese, with its "unbounded cognition", is unbounded in temporal cognition. The typological difference between Japanese and Chinese, whether their temporal cognition is bounded or unbounded, leads to difficulties in the acquisition of aspectual compound verbs and errors by learners of each language.

Japanese and Chinese differ in the "boundedness of temporal cognition". In other words, in Chinese there is a prominent boundedness distinction between telic and atelic events, whereas Japanese possesses "unbounded type cognition", for while there is a distinction between past and non-past, there is not a prominent distinction between telic and atelic events. The typological difference between "bounded" cognition in Chinese and "unbounded" cognition in Japanese makes acquisition of aspectual compound verbs difficult for native speakers of one language learning the other, leading to errors and non-use.

Japanese compound verbs often do not correspond to complex predicates in English or Chinese, so they present a challenge in Japanese language education. However, Japanese aspectual compound verbs are high frequency and occupy an important position in Japanese language education. Rather than memorizing them as new vocabulary items, it is necessary to teach them as a productive system of vocabulary, attempting comparison with English or Chinese by employing concept oriented instruction (Lantolf & Xi, 2021). Similar methods should be employed when teaching English phrasal verbs and Chinese complements.

Table 5 Boundedness of temporal cognition in Japanese and Chinese and influence on second language acquisition

Structural boundedness	Chinese	Japanese
1. Forms expressing boundedness in verb structure	**Strong** **1. Importance of resultative complements** **2. Second language acquisition of Japanese** ⇒ Native speakers of Chinese frequently select the intransitive variant expressing resultativity from verb pairs	**Weak** **1. a. Prominence of complement type compound verbs** **1. b. Cognition of resultativity occurs through transitivity alternations** **2. Second language acquisition of Chinese** ⇒ Omission of resultative complements by native speakers of Japanese
2. Forms expressing boundedness in sentence structure	**Strong** **1.** Telic/Atelic Distinction Telic a. "-了le" expressing completion b. resultative complements Atelic Auxiliary verb "会huì" **2. Second language acquisition of Japanese** ⇒ (1) **Overuse** of PAST **"–ta"** in telic events (2) Difficulty of acquisition of **Resultative Aspect**, " – *te iru*" even at high proficiency level learners of Japanese	**Weak** **1.** a. Past/Non-Past Distinction ・past "*-ta*" ・non-past "*-ru*" (used also in atelic events) b. Morphological ・telic/atelic distinction ・*-te iru* for completive and non-completive 1) atelic progressive aspect 2) telic resultative aspect **2. Second language acquisition of Chinese** ⇒ **Omission** of the auxiliary verb "会huì"

12 Conclusion

The differences in boundedness between Chinese and Japanese share similarities with the differences pointed out by comparative research into English and Japanese including Ikegami (1981) and Kageyama (2002). It will be effective not only for the study of English but also for the study of Chinese if native speakers of Japanese are made aware of the unbounded characteristics of their native language. The first step is to "understand Japanese before acquiring a foreign language".

Acknowledgements This research received the following funding.
[1] Grant-in-Aid for Scientific Research JSPS KAKEN (JP16H01934) "Japanese Learner Corpus Development through International Collaboration and Applied Research into Language Acquisition and Education" (2016–2020, PI: Kumiko Sakoda).
[2] Tokyo University of Foreign Studies International Center for Japan Studies "Japanese Learner Corpus Project" (2016–2019, PI: Keiko Mochizuki).

[3] Grant-in-Aid for Scientific Research JSPS KEKEN (JP17H0235) “Research on cross-referential learners’ corpora of English, Chinese and Japanese though international educational collaboration at secondary and tertiary levels” (2017–2020, PI: Keiko Mochizuki).
[4] Grant-in-Aid for Scientific Research JSPS KAKEN (JP 20H01278) “Research on cross-referential learners’ corpora of English, Chinese and Japanese though international educational collaboration at secondary and tertiary levels” (2020–2022, PI: Keiko Mochizuki).

Notes

(1) Data collected from 165 3rd and 4th year Chinese majors (CEFR B1-B2) at Tokyo University of Foreign Studies in 2013 and 2014. The data consists of 425 essays written as class homework and comprises 179, 940 words.
(2) This corpus of English native learners of Chinese, which is not publicly available, has been error tagged by the author’s research group. The data is comprised of 691 essays by 691 learners, with a total of 136, 212 characters. This includes 225 essays by applicants for a CEFR A2 level test (35, 518 characters), 344 essays by applicants for a CEFR B1 level test (33,490 characters) and 122 essays by applicants for a CEFR B2 level test (67,204 characters).
(3) Frequencies for native speakers of Chinese were obtained from〈现代汉语语料库检索〉available at http://www.cncorpus.org/index.aspx, produced by the Chinese Ministry of Education Institute of Applied Linguistics.

References

Aoki, H. (2013). Fukugodoshi no Rekishitekihenka [Historical Change in Japanese Compound Verbs]. In T. Kageyama (Ed.), *New explorations into the mysteries of compound verbs* (pp. 215–241). Hitsuji Shobo.
Comrie, B. (1976). *Aspect: an introduction to the study of verbal aspect and related problems.* Cambridge University Press.
Eduardo, N., & James P. L. (2006). Concept-based instruction and the acquisition of L2 Spanish. In *The art of teaching Spanish: Second language acquisition from research to praxi* (pp. 79–102).
Huang, J. C. T. (2006). Resultative and unaccusatives: A parametric view. *Bulletin of the Chinese Linguistic Society of Japan, 234*, 1–43.
Ikegami, Y. (1981). ‘Suru’ to ‘Naru’ no Gengogaku: Gengo to Bunka no Taipolozi eno Shiron [The Linguistics of ‘Do’ and ‘Become’: Comparative analysis on the typology of language and culture]. Taishukan Publishing.
Jacobsen, W. M. (1992). *The transitive structure of events in Japanese.* Kurosio Publishers.
Kageyama, T. (1993). Bunpoo to Gokeisei [Grammar and word formation]. Hitsuji Shobo.
Kageyama, T. (2002). Kejime no Nai Nihongo [Japanese as an unbounded language]. Iwanami Publishers.
Kageyama, T. (2013). Goiteki Fukugodoshi no Shintaikei: Sono Rironteki Oyoteki Imiai [A new system of lexical compound verbs: Theoretical and applied implications]. In T. Kageyama (Ed.), *New Explorations into the Mysteries of Compound Verbs* (pp. 3–46). Hitsuji Shobo.
Kageyama, T. (2016a). Verb-compounding and verb-incorporation. In T. Kageyama & H. Kishimoto (Eds.), *Handbook of Japanese Lexicon and word formation* (pp. 273–310). De Gruyter Mouton.

Kageyama, T. (2016b). Agents in anticausative and decausative compound verbs. In T. Kageyama & W. M. Jacobsen (Eds.), *Transitivity and valency alternations: studies on Japanese and beyond* (pp. 89–124). De Gruyter Mouton.

Kageyama, T. (2018). *Compound and complex predicates in Japanese.* Oxford University Press.

Kageyama, T., & Kanzaki, K. (2014). *Compound verb Lexicon (online database).* National Institute for Japanese Language and Linguistics, Retrieved August 23, 2020, from https://db4.ninjal.ac.jp/vvlexicon/db/.

Kageyama, T., & Li, S. (2018). Resultative constructions in Japanese from a typological perspective. In P. Pardeshi & T. Kageyama (Eds.), *Handbook of Japanese contrastive linguistics* (pp. 193–226). De Gruyter Mouton.

Lantolf, J. P., Xi, J., & Minakova, V. (2021). Sociocultural theory and concept-based language instruction. *Language Teaching, 54*(3), 327–342.

Li, P. (1990). *Aspect and aktionsart in child Mandarin.* Doctoral dissertation, Leiden University, the Netherlands.

Li, Y. (1993). Structural head and aspectuality. *Language, 69*, 480–504.

Li, P. & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect.* De Gruyter Mouton.

Matsumoto, Y. (2021). The semantic differentiation of verb-te verb complexes and verb-verb compounds in Japanese. In K. Taro, P. E. Hook, & P. Pardeshi (Eds.), *Verb-Verb complexes in Asian languages.* Oxford Scholarship Online.

Mochizuki, K. (1997). On perfect aspect in Chinese. *Area and Culture Studies*, *55*, 55–71.

Mochizuki, K. (2004). *Causative and inchoative alternation: comparative studies on Verbs in Chinese.* Ph.D. thesis. National Tsing Hua University.

Mochizuki, K. (2007). Patient-Orientedness in resultative compound verbs in Chinese. In Y. Kawaguchi, T. Takagaki, N. Tomimori, & Y. Tsuruga (Eds.), *Corpus-based perspectives in linguistics* (pp. 287–300). John Benjamins.

Mochizuki, K. (2009). Error analysis in voice by advanced-level Chinese learners of Japanese: Comparative analysis with Chinese. *Area and Culture Studies*, *78*, 85–106.

Mochizuki, K., & Shen, Y. (2011). Word formation in compound verbs in Japanese and Chinese. *Comparative Linguistics in Chinese and Japanese, 2*, 46–72.

Mochizuki, K., & Shen, Y. (2012). Inheritance of argument structure and compounding constraints of resultative compound verbs in Chinese and Japanese. In L. E. Clemens & C.-M. Louis Liu (Eds.), *Proceedings of the 22rd North American Conference on Chinese Linguistics (NACCL-22) and the 18th International Conference on Chinese Linguistics (IACL-18)* (Vol. 2, pp. 341–355). Harvard University.

望月圭子，申亚敏，福田翔，游韦伦，张学博，张正．2016. 以英日语为母语的汉语学习者在学习量词短语与结果补语时的偏误差异． 齐沪杨主编《现代汉语虚词研究与对外汉语教学:第六辑》pp.421-446. 上海译文出版社．[Mochizuki, K., Shen, Y., Fukuda, S., You, W., Zhang, X., & Zhang, Z. (2016). The acquisition of the Chinese numeral classifier resultative complements by English L1 and Japanese L1 learners. In Q. Huyang (Ed.), Modern Chinese function research and teaching Chinese (6th series) (pp. 421–446). Shanghai Translation Publishing House.]

Mochizuki, K. (2018). Second language acquisition of Japanese compound verbs: From comparative perspectives with english phrasal verbs and Chinese compound verbs. *Area and Culture Studies*, *96*, 183–204.

Newbery-Payton, L., & Mochizuki, K. (2020). L1 influence on use of tense/aspect by Chinese and Japanese learners of English learner corpus. *Learner Corpus Studies in Asia and the World, 4*, 67–93.

Nishiyama, K. (1998). V-V compounds as serialization. *Journal of East Asian Linguistics, 7*(3), 175–217.

Rocca, S. (2002). Lexical aspect in child second language acquisition of temporal morphology: A bi-directional study. In M. R. Salaberry & Y. Shirai (Eds.), *Tense-aspect morphology in L2 acquisition* (pp. 249–284). John Benjamins.

Shen, Y. (2007). Argument structure and lexical conceptual structure of resultative verb compounds in Chinese. In T. Kageyama (Ed.), *Lexicon Foram 3* (pp. 195–229). Tokyo: Hitusji Shobo.

Shen, Y. (2009). *Resultative compound verbs in Chinese: From a viewpoint of comparative analyses with resultative compound verbs in Japanese and English resultative constructions.* Ph.D. dissertation, Tokyo University of Foreign Studies.

沈家煊. (1995). “有界”与“无界”.《中国语文》第5期. pp.367-380.pp. 367–380. [Shen, J. (1995). Bounded and unbounded. *Chinese Literatures*, *5*, 367–380.]

Tai, J.H.-Y. (1984). *Verbs and times in Chinese: Vendler's four categories.* Chicago Linguistic Society.

Tai, J.H.-Y. (1985). Temporal sequence and Chinese word order. In J. Haiman (Ed.), *Iconicity in syntax* (pp. 49–72). John Benjamins.

Tamaoka, K., & Chu, X. (2013). Chugokujin Nihongogakushusha no Goiteki Fukugodoshi no Shutoku ni Eikyosuru Yoin. In T. Kageyama (Ed.), *New explorations into the mysteries of compound verbs* (pp. 413–430). Hitsuji Shobo.

湯廷池・許淑慎 (2015)《對比分析研究入門》上冊，致良出版社.[Tang, T., & Xu, S. (2015a). *An introductory study of contrastive linguistics* (Vol. I). Jillion Publishing Co.]

湯廷池・許淑慎 (2015)《對比分析研究入門》下冊， 致良出版社.[Tang, T., & Xu, S. (2015b). *An introductory study of contrastive linguistics* (Vol. II). Jillion Publishing Co.]

Vendler, Z. (1957). Verbs and times. *The Philosophical Review, 66*(2), 143–160.

Zhang, J. (2017). Tense, aspect and modality in Chinese and Japanese viewed from the Chinese auxiliary Verb “hui”, empirical research based on learners corpora. Master's Thesis, Tokyo University of Foreign studies.

Zhang, Z. (2019). Acquisition of Chinese resultative complements by Japanese native speakers and the influence of native language: A learner corpus based analysis. *Language, Area and Culture Studies, 25*, 51–59.

Zhao, H., & Shirai, Y. (2018). Arabic learners' acquisition of English past tense morphology: Lexical aspect and phonological saliency. *International Journal of Learner Corpus Research, 4*(2), 253–276.

The (Non-)acquisition of the Chinese Definiteness Effect: A Usage-Based Account

Ludovica Lena

Abstract This chapter investigates the acquisition by French L1 learners of the Definiteness Effect (DE) that characterize Chinese existential–presentational constructions (EPC). Building upon a video-retelling task, oral elicited productions of 15 French advanced learners of L2 Chinese are analysed. In contrast to previous research on L2 DE, mainly conducted within generative approaches to second language acquisition, the present study adopts a functional, usage-based framework and reports on non-target-like performance at advanced levels of acquisition. It is argued that learners are aware of the DE that characterize the EPC in the target language, which is shown by the marginal use of definite pivots in referent-introducing EPCs. By contrast, what they seem not to be aware of is that the EPC should not be used in reintroduction contexts. As a consequence, learners use the EPC format when discourse-old referents are concerned. Strictly speaking, however, they do not 'violate' the definiteness restriction, since a different form, with a different function, is operating in the interlanguage.

L. Lena (✉)
Xiamen university, Xiamen, China
e-mail: ludolena@xmu.edu.cn

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_12

1 Introduction

This paper investigates the acquisition of the Definiteness Effect (DE) that characterizes Mandarin Chinese (hereafter: Chinese) *you* existential–presentational constructions[1] (EPC), focusing on the production of French learners of Chinese as a second language (L2). The DE in French-speaking Chinese L2 learners is studied here within the framework of functional approaches to language learning, by way of using an elicited production task. The (non-)acquisition of the DE is considered in a holistic perspective, that is, linking the phrasal rules (i.e. DE) to the rules of contextualization (Lenart & Perdue, 2004), and considering the (in)definite expressions occurring in learners' EPCs with respect to their general role in the interlanguage system.

DE is defined as a constraint on the occurrence of definite NPs in certain contexts, with existential sentences being the most prominent context for DE to appear (Milsark, 1977). Across languages, EPCs show a certain sensitivity to definiteness, defined as a tendency for definite NPs not to appear in these constructions (Leonetti, 2008). A large body of the relevant literature has focused on DE in English *there* existential sentences (1).

(1) There is {a man/??the man} in the garden.

To account for DE, several explanations have been offered (see McNally, 2019 for a recent review). Lyons (1999: 46) claimed that the DE 'is more likely to be a semantic or pragmatic constraint than a syntactic one. It has much in common with the constraint on indefinite subjects or topics […]; that too is a strong cross-linguistic tendency, stricter in some languages than in others, and also involving something broader than grammatical definiteness'. In McNally (2019: 1839) terms, 'no analysis that appeals exclusively to the form of the pivot will account for the definiteness restriction'. Hence, there is general consensus on the role of the information-structure (IS) articulation over the manifestation of DE. As Leonetti (2016) puts it, '[t]hat IS is relevant for the appearance of DEs in existential sentences and related constructions is hardly a novel insight. It has been repeatedly pointed out that pivot DPs [determiner phrases] are typically focal and existential sentences are central cases of thetic constructions, i.e. constructions lacking an aboutness topic, typically receiving an 'all-focus'/'all-new' interpretation' (2016: 80). In this view, DE thus results from the incompatibility of definite pivotal NPs with the primary function of EPCs, which is that of introducing a novel referent into the discourse (Lambrecht, 2000).

[1] The label "existential-presentational construction" (EPC) is borrowed from Li (2014) ("existential-presentative construction" in her terminology, see also Givón, 1988). While the label "existential" points to some semantic property characterizing the construction, i.e. that of asserting the "existence" of an entity (or rather its location from a situation-based perspective, see Creissels, 2019), "presentational" refers to its pragmatic function which is that of introducing ("present") an entity into discourse. In the literature, "existential construction" generally defines monoclausal forms (e.g. *there is someone in the house*), while biclausal constructions (e.g. *there is someone looking for you*) are more often named "presentational" (see Sarda and Lena, Forthcoming, for a recent review). In this study, the term EPC is used to denote both monoclausal and biclausal *you*-constructions.

Along with the DE, its counterexamples have long been acknowledged as well (see Rando & Napoli, 1978; Lumsden, 1988; Abbott, 1993; Ward & Birner, 1995, among others). For instance, in (2), the second occurrence of a *there* construction involves a demonstrative NP in the pivotal position, which refers anaphorically to the expression *one flight.*

(2) I think there was one flight where
we had one problem. It wasn't ours,

but *there was that one flight*. (Ward & Birner, 1995: 727, reported in McNally, 2019: 1834)

To account for sentences like (2), Abbott (1993) and Ward and Birner (1995) have argued that existential sentences can also serve to reintroduce or focalize referents that have already been evoked in discourse. The deictic and the so-called 'list-reading' existential sentences are also known to accept definite pivots (Rando & Napoli, 1978).

The literature on the existential sentences also provides extensive discussion of exceptions to the DE in different languages (Leonetti, 2008; Bentley, 2013, Bentley and colleagues, 2013, 2015). Beaver and colleagues (2006) pointed out that the cross-linguistic variation in the definiteness restriction is not to be understood as an absolute value; rather, languages will select different intervals in the continuum of 'definiteness', with the DE on the pivot correlating with subject canonicality, which also is a gradient phenomenon, showing relative cross-linguistic variation itself (Bentley, 2013).

This brings us back to the languages considered for the present study, i.e. French (learners' source language) and Chinese (learners' target language). First, French and Chinese differ with respect to the morphological marking of definiteness, the former being an article language, while the latter is a language lacking articles. As far as DE is concerned, the two languages show a contrasting behaviour as well, with Chinese EPCs manifesting a stronger DE than French (see Lena, Forthcoming (a)).[2] Chinese and French both have an EPC governed by the existential predicator (Creissels, 2019), *yǒu* 'have, exist' (3a) and *il y a* 'there is [lit: there has]' (3b), respectively:

(3) a. Yǒu rén gěi nǐ dǎ diànhuà.
Exist person to you make phone.call
'Lit. There's someone who telephoned you.' (Li & Thompson, 1981: 131)

b. Il y a quelqu'un qui t' as téléphoné.
It$_{EXPL}$ there has someone who to.you has telephoned
'Lit. There's someone who telephoned you.'

[2] For the purpose of this study, it is sufficient to note that French monocausal ('existential') sentences sometimes include definite pivots, as in *Tiens, il y a Jean* 'Hey, there is Jean!' (Creissels, 2019). At the same time, the literature offers numerous examples of biclausal ('presentational') constructions with definite NPs (e.g. *il y a le chat qui miaule* 'there is the cat meowing'; see Lambrecht, 1988 and Karssenberg, 2017).

While in French strong expressions can appear within the EPC (4a), the Chinese *you*-construction manifests DE (4b):

(4) a. Il y a Jean qui a téléphoné.
It$_{EXPL}$ there has John who has telephoned
Lit: 'There is John who telephoned.' (Lambrecht, 1988: 136)

b. *Yǒu Yuēhàn dǎ-le diànhuà.
Exist John make-PFV phone.call
Intended meaning: 'There is John who telephoned.'

The sentence (4b) is considered a violation of the DE because of the use of the proper noun *Yuēhàn*—proper nouns being definite by definition, since they are direct labels for particular referents (Chafe, 1976: 39).

Most research has focused on the acquisition of the English DE, and to date, no studies have addressed the L2 acquisition of the Chinese DE—although some DE-relevant findings are reported in Yang and colleagues' (2007) study on the L2 acquisition of Chinese existential patterns, which I return to in a moment. However, it should be noted that the *there* construction—the typical context showing DE in English—do not represent a major IS strategy in English, given that intonational devices are more often used in this language to express focus articulations (Vallduví, 1991, Lambrecht, 1994: 318, Sasse, 2006). By contrast, French speakers learning Chinese are facing an L2 which makes extensive use of EPCs as an IS device, as their source language does (Klein, 2012; Lambrecht, 1988).

Counterexamples to the DE have been documented in Chinese as well (see Lena, 2020c: 208–213, Forthcoming (a) for a review). Huang (1987) showed that DE is not observed when a full lexical NP appears in the topic position (e.g. *túshūguǎn yǒu nèi ben shū* [library exist that CL book] 'there is that book at the library'). Li (1996) argued that two types of *you*-constructions should be identified, one introducing a new entity and the other asserting the existence of an event, with only the former manifesting DE. Hu and Pan (2007) discussed the role of adverbs such as *hái* 'also' or *zhǐ* 'only': 'what is unnoticed in the literature is that, although it is generally excluded from the post-*you* position, a definite NP can occur there if a focus particle is introduced into the relevant sentences' (Hu & Pan, 2007). Cai (2000) and Xia (2009) report on the contexts that allow definite NPs within the *you*-construction, showing that these are subject to strong constraints and rarely found in main clauses; when this is the case, definite-pivot EPCs in Chinese are accepted in some restricted contexts, and cannot stand in isolation.

Despite these exceptions to Chinese DE, it is generally assumed that in the great majority of contexts only weak expressions such as indefinite NPs can occupy the pivotal position (3a). Therefore, native speakers (NS) of Chinese are not likely to accept EPCs including a strong pivot such as (4b). While sentences like (4a) are perfectly acceptable in French, from a statistical point of view the *il y a* construction does tend to include indefinite NPs, as Karssenberg's (2018) corpus inquiry showed.[3]

[3] Karssenberg's (2018) study focuses on *il y a* clefts, that is, 'biclausal constructions' following the terminology adopted in the present article.

This feature is congruent with their main pragmatic function of introducing non-topical referents into discourse, which are typically unidentifiable ('new') referents.[4] There is, however, a different type of *il y a* construction, which is used to express a focus-background articulation, being more closely linked to the expression of strong NPs (Lambrecht, 1988: 154, Karssenberg, 2018: 63). These forms are often triggered in questions-answers contexts:

(5) (Context: 'What's your favourite TV show right now?')

"How I Met Your Mother" c'est génial, *y'a aussi "Lost" qui est bien.*

'"How I Met Your Mother" is great, there's also "Lost" that is good.' (Karssenberg, 2018: 63).

Building on a story retelling task (Sect. 5.2), the current study only reports on the first EPC type, used to introduce—and eventually reintroduce, as we shall see it later—referents into discourse and is therefore not concerned with constructions having an IS articulation like (5). It is nonetheless useful to keep in mind that EPCs in learners' source language (i.e. French) are highly multifunctional, with the focus-background type being strongly connected to the expression of definite NPs.

In what follows, I begin by discussing how (in)definiteness is marked in Chinese.

2 (In)definiteness in Chinese

As an article-less language, Chinese lacks the grammatical category of (in)definiteness, which of course does not mean that such a distinction cannot be expressed in this language.

LaPolla (1995) describes the possible encoding of referents in Chinese on the basis of their identifiability and accessibility status (Ariel, 1990; Chafe, 1994; Gundel et al., 1993; Lambrecht, 1994: 78, 106; Prince, 1981). The [(numeral +) classifier] sequence, leaving aside its non-referential uses (Chen, 2003), is mainly used to express unanchored non-identifiable referents, that is, prototypical (quantified) new referents—the ones marked by the indefinite article in languages such English and French (e.g. *yí ge lièrén* 'a hunter' in [6]). By contrast, nouns modified by demonstrative determiners *zhè* 'this' and *nà* 'that' are used to encode identifiable referents, that can be in one of the three activation states—active, accessible or inactive (Lambrecht, 1994: 109)—according to LaPolla (1995). The NP *zhè zhi gǒu* 'this dog' in the sentence below (from Chen, 2004 : 1153) points to an activated referent thus offering an example of the anaphoric use of the demonstrative determiner:

(6) Yǒu yí ge lièrén... yǎng-zhe yì zhi gǒu. **Zhè zhi gǒu** hěn dǒngshì.
Exist one CL hunter keep-DUR one CL dog this CL dog very intelligent
'There was a hunter who had a dog. The dog was very intelligent.'

[4] As Karssenberg (2018: 98) notes, however, the use of proper nouns such as *Jean* in (4a) is motivated by the familiarity between the interlocutors, which is virtually absent from the corpora she consulted.

While acknowledging that the demonstrative determiners in Chinese are 'the closest to definite articles in other languages' (*ibid.*: 1151), Chen illustrates the differences between the two. *Zhè* 'this' and *nà* 'that' serve all the typical functions of demonstratives, as they are used in situational, discourse deictic and contrastive anaphoric contexts. In addition, they can be used where demonstratives (say, in languages like English) are not allowed: non-contrastive anaphoric, shared knowledge and frame-based association (*ibid.*: 1151–1153). However, uses characteristics of fully grammaticalized definite articles, like those marking shared specific and general knowledge, and frame-based association, 'are exceptional rather than the norm' with Chinese demonstratives (*ibid.*: 1156). Similarly, Crosthwaite et al. (2018) show that Chinese speakers more often use bare nouns (BNs) when establishing bridging reference (i.e. frame-based anaphora).

Indeed, besides nouns that are marked (either by a numeral or by a demonstrative), Chinese speakers can make use of BNs as referring expressions (e.g. *gǒu* 'dog[s]'). Their interpretation with respect to definiteness is not straightforward. Indeed, BNs as referring expressions present the highest degree of ambiguity, given that they can designate unidentifiable, inactive, accessible and active referents (LaPolla, 1995: 305, see also Lena, Forthcoming (b)). Between accessible referents, BNs can denote deictically identifiable referents and uniquely identifiable referents (Cheng & Sybesma, 1999: 510, Chen, 2004: 1165. Finally, BNs achieve an indefinite interpretation in some positional contexts (Chao, 1968: 76; Li & Thompson, 1981: 510; Xu, 1995; Hole, 2012: 61): they are generally interpreted as indefinite when appearing as postverbal subjects and within a *you*-construction (besides the generic-partitive readings that are not relevant for the purpose of the present study, see Lena, Forthcoming (b), for details).

In sum, even if most of their functions overlap, demonstratives and numerals in Chinese do not cover the same range of functions as articles in article languages. As discussed in the following section, the acquisition of the article system is documented to be a hard task for the L2 learner—particularly if the L1 lacks articles—who is inevitably faced to the multifunctionality of the forms involved. Conversely, speakers whose L1 is an article language learning an article-less language such as Chinese will be confronted to the repertory of referring expressions available in the target language to express (in)definiteness, most of which are multifunctional as well. Chinese BNs are expected to be particularly challenging for the L2 user, not only for their morphologically unmarked form but also because the overall effect of the sentence pattern, and the discourse context, over the (in)definiteness interpretation.

3 Previous L2 Research on the DE

Previous studies have mostly focused on the acquisition of the L2 article system, where a great amount of research has been conducted on the acquisition of articles in English (e.g. Berry, 1991; Grannis, 1972; Master, 1997; Zobl, 1984). The misuse

of definite and indefinite articles in a second language is often analysed as a positive or negative transfer from the source language. That is, the presence of a similar article system in the L1 should facilitate the acquisition of the article system in the target language, but when the mother tongue has no equivalents, this can result in an increased difficulty for the learner. For instance, Zobl (1984, reported in Towell & Hawkins, 1994: 9) notes that the acquisition of *a* and *the* in English L2 is faster for speakers whose first language distinguishes between definite and indefinite determiners (e.g. French, Spanish), compared to learners of article-less L1 backgrounds (e.g. Chinese, Russian). Based on an oral elicited picture description task, Sleeman (2004) compared the acquisition of definiteness distinctions by Dutch- and Japanese-guided L2 learners of French and shows that Dutch (an article language) speakers performed better in oral speech than Japanese (an article-less language) L1 speakers.

A few L2 studies have focused on the DE in relation to existential constructions. The case studies conducted by White (2003) and Lardiere (2005) analysed the production data from one advanced Turkish speaker of L2 English and one advanced Mandarin speaker of L2 English, respectively. White's (2003) study built on spontaneous production data collected through a series of interviews and reported no DE violations, 'even though the subject did make errors in article suppliance, in the form of omission' (White, 2003). Lardiere (2005), in her study of a Chinese learner of English, similarly reports no DE violations. Within her dataset, she obtains 37 contexts for existential *there*-constructions, and no definite articles were produced in any of them, despite the fact that the speaker 'tends to overuse definite more than indefinite articles overall' (Lardiere, 2005).

Building on larger data, King et al. (2006) found that Chinese speakers of low intermediate proficiency in English did not distinguish between DE violations and equivalent grammatical sentences in a grammaticality judgment task (GJT). More advanced subjects, on the other hand, showed target-like sensibility to DE.

Yang et al. (2007) studied the acquisition of the Chinese existential patterns by low to intermediate learners of different L1 backgrounds (English, Japanese and Korean). To collect the data, a GJT (Fig. 1), a guided composition task (Sasaki, 1990) and a free composition task were used.

While Yang et al.'s (2007) study was not focusing on the DE, several items included in the GJT present interesting findings in this regard. *You*-constructions are generally rejected by Chinese NS when including a definite pivot. However, the acceptance rate is higher when a preverbal locative is added. 'This shows that although Chinese speakers can accept definite nouns in existential sentences, the acceptance rate of definite nouns is lower than indefinite nouns, because the main

桌子上有两本书。(On the table there are two books)
1 非常合语法 (very grammatical) 2 可能合语法 (possibly grammatical) 3 可能不合语法 (possibly ungrammatical) 4 非常不合语法 (very ungrammatical)

Fig. 1 Example of (natural) target item used in Yang et al. (2007) GJT (my translation)

function of existential sentences is to introduce new information'. (Yang et al., 2007, my translation). Interestingly enough, English and Japanese learners are less prone to accept definite nouns appearing in EPCs than Chinese NS, thus not accepting sentences like *zhuō-zi = shang yǒu nà-ben shū* 'on the table there is that book' or *fángjiā = li yǒu Xiǎo Mín* 'in the room there is Xiaomin'.[5] While Yang et al.'s (2007) study is informative with respect to the acquisition of different features of the existential patterns in Chinese, a satisfying explanation regarding learners' treatment of the DE is not really provided. They seem to take into account a possible L1 functional transfer, though their argument is essentially limited to English *there*-sentences. I suggest that uncontextualized target sentences might not be the most useful way to investigate this issue: given that the prototypical function of EPCs in Chinese is that of introducing new referents, learners might not be able to imagine contexts in which a definite pivot is allowed (see Sect. 6). Their study shows, nonetheless, that low to intermediate learners of L2 Chinese do not manifest DE violations; on the opposite, they fail to identify DE exceptions.

White (2008a) makes use of an elicited production task (based on pictures description) from 18 Turkish speakers and 15 Mandarin speakers of various proficiency levels. No DE violations are reported in her study. In another study (also based on a picture's description elicited production task), White (2008b) show that intermediate and advanced L1 Chinese speakers of L2 English are sensitive to the DE, 'treating *there*-insertion constructions in a native-like way despite non-native performance on articles in general' (White, 2008b).

Similar conclusions are reached by Yu and Su (2011), who carried an online contextualized GJT on 50 Chinese intermediate and advanced learners of L2 English, and reported no DE violations. The study by White et al. (2012) was also based on a GJT where each test sentence was preceded by a short context (see Fig. 1 for an example), and found that advanced Russian and Turkish learners of English could perform just as well as NS on judgments of DE. The authors argue that the participants' ability to judge English sentences appropriately cannot be explained in terms of L1 transfer.

Replicating White et al. (2012)—with minor modifications—Snape and Sekigami (2016) found that Japanese L2 learners of English at advanced levels of proficiency are able to differentiate between grammatical and ungrammatical EPCs, while the intermediate learners can correctly identify grammatical items but cannot detect DE violations. Hence, sensitivity to the English DE seems to improve along with the learner's proficiency level. The authors claim that the presence of the DE in Japanese is likely to aid acquisition of affirmative EPCs in L2 English, despite the absence of articles in Japanese. Note, however, that their study, as White et al. (2012) did, also included negative EPCs, which can include weak or strong expressions in Japanese, but only weak expressions in English. They conclude that advanced learners' sensibility to DE in English negative EPCs cannot be explained in terms of a facility prompted by their source language in this case.

[5] The Korean group does not seem to make distinctions on the basis of the definiteness of the pivot, as several grammatical target items including indefinite pivots are also rejected.

In general, the results from previous longitudinal studies suggest that second language learners become sensitive to definiteness restriction in English as their linguistic competence becomes more target-like over time. No violations of the DE are observed at advanced levels of the acquisition. By contrast, the current research reports on DE 'violations' produced by adult French advanced learners of L2 Chinese. These are analysed as the results of a functional transfer of the pragmatic functions that the EPC can serve in the source language (i.e. French). Note that all the studies on the DE presented in this section—with the exception of Yang et al. (2011)—are conducted within generative approaches to second language acquisition (SLA).

Before turning to the data and the results of the current study (Sect. 5), what follows presents previous research on learners' acquisition of the referring expressions in the target language, conducted within the framework of functional approaches to SLA.

4 Definiteness Effect in the Light of Previous L2 Research on the Acquisition of Reference

While not directly addressing learners' 'sensitivity to DE', various studies conducted in a functionalist framework analysed the L2 acquisition of the linguistic forms that enable discourse construction and anaphoric linkage (Ahrenholz, 2005; Chini, 2005; Leclercq & Lenart, 2013, Ryan, 2015, Lenart, 2020). By highlighting the complexity of the form–function relationship, past research has studied the different linguistic means used to encode the noun (e.g. the use of determinants) and their relations with the expression of different referential values (e.g. introduction vs. maintenance of reference) in the productions of learners with various levels of competence (Lenart, 2006, Lenart & Perdue, 2004, Watorek, 2004, Watorek et al., 2014). Longitudinal studies have shown an early preference for distinguishing between definiteness and indefiniteness by marking only definiteness[6] (Chaudron & Parker, 1990).

In a functionalist approach, the acquisition of the L2 article system is considered in a holistic perspective, assuming that the information expressed morpho-syntactically by the articles can be conveyed by other linguistic devices, be them lexical, morphological or positional. Past studies have shown that, even if the target forms are not mastered, nonnatives can implement distinct ways of manipulating NP forms for different discourse contexts: 'even with less than perfect mastery of the target system, learners still use different forms of the system to make discourse distinctions' (Chaudron & Parker, 1990). For instance, Lenart and Perdue (2004) studied the oral narrative of Polish (an article-less language) adult basic learners of L2 French and reported several problems concerning the omission of nominal determiners. Yet, these learners were fully capable of organizing their narrative according

[6] Interestingly enough, lower learners tend to use BNs for indefinite reference, while using a definite NP for definite reference. As the proficiency increases, indefiniteness is then encoded formally (Chaudron & Parker, 1990).

to pragmatic organizational principles, where referential continuity was ensured by nominal and pronominal anaphora or by the absence of any marking (zeros).

The complex operation of building a coherent narrative involves the interaction between the discourse level and the utterance level. Two sets of rules—phrasal rules and rules of contextualization—interact in determining the learners' use of a particular form in a particular context (Lenart & Perdue, 2004). It has been shown that the interaction between the two may vary across languages (Lambert et al., 2008) and also within the same language depending on the speech type (Watorek et al., 2014). As Ryan (2020) puts it, '[t]his dual processing demand, lying as it does at the confluence between pragmatics and grammar, has proved an intriguing site for SLA research'. In the area of noun reference, learners have not only to master the internal structure of the NP but also the appropriate use of its realizations in a given situation (Lenart & Perdue, 2004).

When producing EPCs, speakers are effectuating two related yet distinct operations. The selection of a sentence pattern goes along with the choice of a referring expression which appropriately designate the NP referent—both operations being dependent on the pragmatic context. That is, the use of an EPC is motivated by the IS articulation, while the NP form is related to the discourse status of the referent involved. In Chinese, for instance, EPCs are strongly linked to the referent-introducing function, and the referents denoted by the pivot NPs tend to be brand new (Li & Thompson, 1981: 612). In French, the EPC is also frequently used to introduce referents into the discourse, but these can have an accessible status, which is reflected in the definiteness of the pivot (e.g. *Jean* in [4a] which belongs to the common ground but is discourse-new).

In this study, the DE is considered in relation to the discourse status of the denotatum and the function(s) that the EPC is assigned in the (inter)language. The (in)definiteness of pivotal NPs is not analysed in a strict binary fashion (i.e. non-violations vs. violations) but the frequency of the forms involved is also taken into account, as well as their status in the inventory of referring expressions used to mark (in)definiteness oppositions in learners' productions.

5 The Current Study

5.1 Participants

The data analysed in this chapter were originally collected for a bi-directional project exploring referent (re)introductions in the narratives of both L1 French L2 Chinese and L1 Chinese L2 French learners. The oral productions of 15 NS of French, 15 NS of Chinese (the control groups) and 15 L1 French L2 Chinese learners and 15 L1 Chinese L2 French learners (the L2 groups) were collected (see Lena, 2017, 2020b). For this study, only the data produced by French learners of L2 Chinese, and the control group of Chinese NS, are considered.

The French learners' group is composed of eight men and seven women. They have studied Chinese at the university for at least 3 years (5 years on average) and speak a third language at minimum (i.e. English). They belong to Bartning's (1997) definition of 'learners with a high level of education', that is, learners that have studied the L2 in higher education and have a strong metalinguistic knowledge of the target language.

In addition, the participants in this study lived in a country where the target language is spoken (i.e. China) for at least 1,5 years (3 years on average), while 14 of them had settled in China permanently at the time of the experience, and were recorded there.

5.2 Materials and Procedure

The stimulus used to elicit film retellings is an extract from Charlie Chaplin's silent film *Modern Times* developed as part of a European Science Foundation project (see Klein & Perdue, 1992). Following Turco (2008) and Sun (2008) among others, the short sequence comprising the bread robbery scene was selected, to which a final scene was added in order to study the expression of reintroduced referents which leave the stimulus for a period before returning later. The storyline can be described as follows:

> Sequence one: *A hungry girl steals a loaf of bread from a bakery shop. A lady nearby sees the robbery and promptly informs the baker. The baker runs after the girl. The girl bumps into Charlie Chaplin, and the two fall to the ground. A police officer arrives with the baker. Chaplin takes the blame for the stolen bread. The police officer brings him away.*
>
> Sequence two: *Chaplin is freed from the police station. He founds the girl outside waiting for him. They hug.*

The researcher met each participant individually in a quiet room. The subjects were instructed to watch the 2-min video on a computer screen. They were told that they could watch the video as many times as they needed, and that they could take time before starting their retelling. When they felt ready, they were then asked to retell the story to a fictional naïve listener—sharing no mutual knowledge—while the experimenter (the author of this paper) was recording them. The subjects were also told that they should try to speak in a spontaneous way, as they would tell the story to a close friend. To emphasize this point, the researcher highlighted that there was not a 'correct' or 'wrong' way of saying things, and that they should just try to speak as they would normally do.

The author of this paper then transcribed the recordings collected and coded them according to the types of referring expressions (e.g. quantified nouns, BNs, DEM + N, etc.—see Tables 3 and 4). Further, EPCs were identified in each dataset (Tables 1 and 2).

Table 1 The overall use of EPCs in Chinese NS' and French learners' narratives

	Chinese NS (n = 15)	Chinese L2 (n = 15)
Instances of EPCs	42 (4.8%)	60 (7.2%)
Total acts of reference	860 (100%)	822 (100%)

Table 2 Nominal expressions occurring within the EPC in the narratives of Chinese NS and French learners of L2 Chinese

Referring expressions	Examples	Occurrences in the corpora	
		Chinese NS (n = 15)	Chinese L2 (n = 15)
Bare nouns	*jǐngchá* 'police'	11 (26.1%)	4 (6.6%)
Quantified nouns	*yí ge rén* 'a person'	28 (66.6%)	40 (66.6%)
Bare genitive nouns	*miànbāodiàn (de) lǎobǎn* 'bakery owner'	2 (4.7%)	2 (3.3%)
DEM + N	*nà ge miànbāodiàn (de) lǎobǎn* 'that bakery owner'	1 (2.3%)	12 (20%)
Proper noun	*Zhuóbiélín* 'Chaplin'	– (0%)	2 (3.3%)
Total EPCs		42 (100%)	60 (100%)

Table 3 The overall use of bare nouns and demonstrative phrases in the narratives of Chinese NS and French learners of L2 Chinese

	Chinese NS (n = 15)	Chinese L2 (n = 15)
Bare nouns	173 (20.1%)	106 (12.8%)
DEM + N	174 (20.2%)	173 (21%)
Total acts of reference	860 (100%)	822 (100%)

The bread robbery scene includes the introductions of five main characters (the thief, the baker, the snitch, Chaplin, and the police officer), offering typical contexts for EPCs to be used. While in a previous study (Lena, 2020b), I have focused on the encoding of those characters, for the purpose of the current research the syntactic unit (i.e. the EPC) is taken as the starting point for the analysis. That is, the introduction of peripheral individuals, inanimate entities as well as introductions in the reported speech were all considered, since any occurrence of EPC provided useful data for the analysis of learners' sensitivity to Chinese DE and possible DE 'violations'. Hence, all the percentages are calculated over the total acts of reference (Ryan, 2015) observed in each corpus (Chinese L1 and Chinese L2).

Table 4 The overall use of lexical referring expressions in the narratives of Chinese NS and French learners of L2 Chinese

	Chinese NS (n = 15)	Chinese L2 (n = 15)
Unmarked lexical NPs		
Bare nouns	173 (20.1%)	106 (12.8%)
Bare-head genitive phrases	46 (5.3%)	32 (3.8%)
Bare-head relative phrases	12 (1.3%)	5 (0.6%)
Subtotal	231 (26.8%)	143 (17.3%)
Marked lexical NPs		
Quantified nouns	112 (13%)	113 (13.7%)
DEM + N	174 (20.2%)	173 (21%)
Subtotal	286 (33.2%)	286 (34.7%)
Others[a]	343 (39.8%)	393 (47.8%)
Total acts of reference	860 (100%)	822 (100%)

[a]This category includes non-lexical referring expressions (zeros and pronouns) which were not considered for this study

5.3 Results

The total number of EPCs found in the oral narratives produced by the two groups of speakers (French learners of Chinese and Chinese NS) are presented in Table 1.

Note that, even if frequently used for referents introductions in Chinese, *you*-constructions are marked forms, since the introductions of new referents are statistically rare when compared to the operation of establishing anaphoric relations in discourse.[7] Overall, the relatively increased use of *you*-constructions in the French learners' narratives when compared to the NS group (7.2% contra 4.8%) is discussed in detail elsewhere (see Lena, 2020b, 2020c: 427–429) as the result of a 'unicity of functions' effect (see Bartning & Kirchmeyer, 2003). That is, learners lean on a smaller inventory of presentational constructions to introduce new referents into the discourse while underusing bridging operations.

Table 2 presents in detail the nominal expressions appearing in the EPCs produced by Chinese NS and French learners of Chinese.

A note regarding the (in)definiteness characterization of the referring expressions presented in Table 2: quantified nouns are classified as 'indefinite', and so do BNs that typically acquire an indefinite reading in this context (see Sect. 2). Nouns modified by a demonstrative determiner, as well as proper nouns, are considered 'definite'. Bare genitive nouns such as *miànbāodiàn (de) lǎobǎn* '[the] bakery owner' stand somewhere in the middle. According to Chen (1986: 16–17, cited in LaPolla, 1995:

[7] As Li (2014) puts it, "PCs [presentative constructions] are used sparingly and never in series, especially those with a foregrounding function. This is reasonable considering the fact that foregrounding PCs introduce important participants into discourse. Such participants are deemed to be small in number".

307), all NPs marked by a genitive phrase or a relative clause are 'definite'.[8] Consider the examples below, from an NS (7) and a learner (8):

(7) CH1

Yǒu	yí	ge	nǚde			
Exist	one	CL	woman			
tā	jīngguò	yì	jia	miànbāodiàn	[…]	
3SG	pass	one	CL	bakery		
Ránhòu	tā	kàndào	##			
then	3SG	see				
pángbiān	yǒu	**miànbāodiàn**		**de**	**shīfu**	
near	exist	bakery		SUB	master.worker	
cóng	tā	chē	lǐmiàn	bān	huòwù	#
from	3SG	car	inside	move	goods	
jìn	diàn=li.					
enter	shop=in					

'There's a lady who passes by a bakery shop […] then she sees that nearby there is the owner of the bakery who is taking the goods from his car and enters the shop'

(8) CH2_FR1

Kāishǐ	de	shíhòu						
Begin	SUB	moment						
yǒu	yí	ge	nǚhái	[…]				
exist	one	CL	girl					
tā	lùguò	yì	jia	miànbāodiàn	#			
3SG	pass	one	CL	bakery				
ránhòu	jiùshì	zhè	ge	shíhòu	tā	kàndào		
after	then	this	CL	moment	3SG	see		
yǒu	**miànbāodiàn**	**de**	**lǎobǎn**	#				
exist	bakery	SUB	owner					
tā	cóng	tā	de	chē	bǎ	nà	ge	miànbāo
3SG	from	3SG	SUB	car	ACC	that	CL	bread
##	ná=dào	##	diàn	lǐmiàn				
	bring=to		shop	inside				

Lit: 'At the beginning there's a girl […] she passes by a bakery shop, then at this moment she sees that there is the owner of the bakery, he takes the bread from his car into the shop'.

In both extracts, the NPs included in the *you*-constructions present a new referent which is connected to a discourse-old entity (i.e. *miànbāodiàn* 'bakery') that has been introduced as an indefinite quantified object a few propositions earlier. Such nouns can be considered pragmatically 'anchored' (Prince, 1981) since the nominal head is indefinite (i.e. discourse-new), while the genitive phrase is definite (i.e. discourse-old or inferable, depending on the context). From a morphological point of view, such forms are unmarked since they commute with quantified genitive phrases (e.g. ***yí ge*** *miànbāodiàn de lǎobǎn* 'a bakery owner')—i.e. 'indefinite'—and with genitive phrases modified by a demonstrative determiner (e.g. ***nà ge*** *miànbāodiàn de lǎobǎn* 'that bakery owner')—i.e. 'definite'. Bare genitive NPs are indeed marginal in our corpora, but they do raise the question of *what* precisely should be considered as

[8] Note that bare-head NPs can also denote unique reference:
(#Na/#Zhe ge) **Taiwan** (**de**) **zongtong** hen shengqi.
[#that/#this CL Taiwan SUB president very angry]
'The president of Taiwan is very angry.' (Jenks, 2018).

a DE violation in Chinese. Given that such forms are [-marked] (i.e. not modified either by a quantifying expression or by a demonstrative, as just said), bare genitive NPs will be classified as 'indefinite' (just as BNs) for the purpose of this study.

With all this in mind, let us now consider the DE in the data (Table 2). To begin with, the *you*-constructions in the two groups tend to include indefinite quantified nouns, in equal proportions (66.6%). Note that learners' absolute number of indefinite quantified nouns in EPCs is not sufficiently informative per se, as it might result from an overuse of nouns modified by an indefinite marker (notably *yí-ge* 'one-CL') in their corpus. As just said, bare genitive nouns are marginal in both groups. The use of a proper noun within the EPC is observed only in French learners' productions but remains marginal as well (2 occurrences). Overall, two main tendencies set apart the use of EPCs observed in learners' narratives from the ones produced by NS: the reduced use of BNs and the increased use of nouns modified by a demonstrative phrase. Examples of each case are provided below[9]:

(9) CH2_FR1 Ránhòu yě yǒu **jǐngchá** dào-le
Then also exist police arrive-PFV
Lit: 'Then there's also [the] police who arrives'

(10) CH2_FR1 Zhè ge miànbāodiàn yǒu **zhè ge shāngdiàn de rén**
This CL bakery exist this CL shop SUB person
Lit: '(In) this bakery there is this vendor (lit. this shop person)'

Note that, in (10), the first *zhè* is used in a typical anaphoric context (*miànbāodiàn* 'bakery' has been evoked in previous discourse), while, in its second occurrence, the proximal demonstrative is used in a way that could be considered cataphoric, since it modifies a new entity of high thematic importance with continuing presence in the following discourse.

In order to link these facts—i.e. the underuse of BNs and the overuse of DEM + N in learners' EPCs—to the role that these forms play as referring expressions in the interlanguage, Table 3 shows the relative frequency of BNs and DEM + N in the narratives of Chinese NS and French learners of L2 Chinese. The percentages are calculated over the total acts of reference observed in each corpus.

The only notable difference between the two groups of speakers is a decreased use of BNs in French learners' productions (12.8% contra NS 20.1%) (see also Liu & Huang, 2015). Table 4 offers a more detailed view of the overall use of lexical

[9] Sentences of this kind (which are found in the L1 corpus as well) question the 'indefinite' reading typically associated to BNs when appearing within EPCs. While the definiteness distinction is clear for prototypical examples such as *Rén lái-le* 'the (expected) person(s) came' versus *Yǒu rén lái-le* 'there's someone who came (= someone came)', it is not unproblematic to assume that the BN *jǐngchá* 'police' changes its referential interpretation between the two sentence patterns *Jǐngchá dào-le* '[the] police arrives' and *Yǒu jǐngchá dào-le* 'there's [the] police who arrives' (9). In fact, the noun *jǐngchá* 'police' seems to refer in both cases to an inherently uniquely identifiable referent (see also Lena, Forth. (a)). Out of the 4 instances of *you*-construction including a BN in learners' productions, 3 are used to introduce the referent of the police officer. A look at the lexical nature of BNs in their corpus show a similar tendency: learners seem to use BNs to denote uniquely identifiable referents while less prone to use it in (purely) anaphoric contexts (e.g. *nǚhái* '[the] girl').

referring expressions in the oral narratives of the two groups, by distinguishing between marked and unmarked forms.

Several observations are noteworthy. First, NS' and learners' overall use of indefinite quantified nouns is roughly the same (13% and 13.7%, respectively). In addition, no inappropriate use of the indefinite marker (i.e. cases where a definite should be expected) is observed in learners' corpus. The (unlikely) hypothesis of overgeneralization of indefinite quantified NPs in learners' productions influencing the use of such forms within their EPCs (Table 2) thus can be excluded.

Second, French learners use BNs to a lesser extent (12.8%) when compared to the proportion observed in the NS' productions (20.1%). The difference is slightly higher when one considers any bare or bare-head NP, that is including bare-head genitive phrases and bare-head relative phrases[10] (17.3 vs. 26.8%).

As previously noted, the acquisition of BNs can be triggering for the L2 learner whose L1 is an article language, given the absence of pre-nominal modifiers in these forms by definition. Thus, the relative under-representation of BNs in French learners' narratives could be seen as the influence of the L1's system (an article language) preventing the use of an unmarked form (i.e. BNs) in the IL. The low frequency of BNs in learners' EPCs (Table 2) can be linked to the reduced use of these forms more generally observed in their narratives.

Table 4 also shows that learners do not overuse DEM + N in general, since these forms represent 21% of the referring expressions observed in their productions, contra 20.2% in NS' narratives. Hence, the under-representation of BNs does not result in an increased use of DEM + N in learners' discourse.

It has been shown that article-language learners tend to introduce definite new referents (i.e. bridging reference) with DEM + N, instead of 'simply' using felicitous BNs (Crosthwaite et al., 2018, Lena 2020b). It could be the case for such a marked strategy to be mobilized for referent introductions, while in acts of reference tracking learners do not overuse deictic demonstratives (even though not relying on BNs as much as natives do). To test this hypothesis, one should analyse the distribution of referring expressions with regard to the activation state of the referents involved, that is, including pronouns and zero anaphora.[11]

[10] In general, learners produce significantly less embedded relative clauses, be them modified by a demonstrative or not. This is noteworthy given that such forms in Chinese serve not only to identify a referent but also to encode backgrounded portions of the narrative. The link between entity referring and event grounding, and how do learners achieve the same distinctions in their narratives is an interesting one that will be left aside for further research. For the purpose of the current study, the label "DEM + N" embraces any NP modified by a demonstrative determiner, thus including relative phrases whose head is modified by a DEM, as *dǎjiù tā de nà ge xiānsheng* [rescue 3SG SUB that-CL gentleman] 'the [lit. that] man who rescued her'. All those forms virtually commute with the unmarked forms (e.g. *dǎjiù tā de xiānsheng* [rescue 3SG SUB gentleman] '[the] man who rescued her').

[11] When compared to NS' discourse, learners' productions differ in their overuse of pronominal reference (namely, third person pronoun *tā*) not only over zeros but also with respect to BNs (see also Ryan, 2015).

For the purpose of the current study, it is sufficient to note that French learners of L2 Chinese do not overuse demonstratives in order to 'replace' the definite article available in their source language. What they do, however, is produce *you*-constructions including definite NPs, namely nouns marked by the demonstrative determiner, thus 'violating' the Chinese DE. Why is it so?

As it has been pointed out, in most cases, *you*-constructions in French learners' productions include an indefinite quantified NP pragmatically denoting a brand-new unanchored referent (in Lambrecht's 1994: 165 terms):

(11) CH2_FR1 Yǒu **yí** **ge** **nǚhái** zǒu=zài lù=shang
Exist one CL girl walk=at road=on
'There is a girl walking on the street'

(12) CH2_FR1 Gānghǎo nà ge shíhòu yǒu **yí** **ge** **chē** ##
just that CL moment exist one CL car
yí **liang** **chē** kāi-guo-lai.
one CL car drive-pass-come
Jiùshì miànbāodiàn de chē
be.precisely bakery SUB car
'Right then there is a car driving over. It is the bakery truck.'

When French learners' *you*-constructions include bare nouns (9) and bare genitive nouns (8), such forms point to referents that are uniquely identifiable or anchored, respectively. As opposed to cases like (11–12), these are accessible referents. Yet the EPC is used to introduce them into discourse for the first time.

As far as the maintenance and the reintroduction of referents are concerned, the *you*-construction should not be used if speakers understand the function of this device, which is to introduce (brand-)new referents into discourse. Indeed, only one occurrence of *you*-construction including a maintained referent is observed in the Chinese NS' data (whereas the existential verb *you* is preceded by the adverb *hai* 'also' triggering a 'list' interpretation—see Hu & Pan, 2007), and no occurrences of *you*-construction including a reactivated referent are found in the L1 corpus.

With such issues in mind, Table 5 considers now French learners' 'violations' of the target language DE in the light of the pragmatic status of the referent denoted by the pivot NP.

In French learners' narratives, only four occurrences (6.6%) of a *you*-construction including a strong determiner are used to introduce new referents into the discourse. In other words, referent-introducing *you*-constructions in most cases include indefinite pivots.

Table 5 Distribution of definite and indefinite NPs within French learners' *you*-constructions in relation to the pragmatic status of the pivot referent

	Indefinite NPs	Definite NPs	Total
Introduced referents	46 (76.6%)	4 (6.6%)	50 (83.3%)
Reintroduced referents	– (0%)	10 (16.6%)	10 (16.6%)
Total	46 (76.6%)	14 (23.3%)	60 (100%)

When EPCs include a DEM + N in the context of referent introductions, they are used to encode the characters of the baker (13) and the police officer (14), that is, the same entities that are likely to be introduced into discourse as definite preverbal subjects, by virtue of a frame-based association and unique reference, respectively (Lena, 2020b).

(13) CH2_FR1 Yǒu **nà ge miànbāodiàn de lǎobǎn** tíng chē le
Exist that CL bakery SUB owner stop car CRS
'There is that owner of the bakery who parked his car'

(14) CH2_FR1 Nà ge shíhòu ne #
That CL moment PAU
yǒu **nà ge jǐngchá** ##
exist that CL police
yě ## dào # yě dào-le
also arrive also arrive-PFV
Lit: 'At that moment, there is that police who also arrives'

EPCs like (13–14) are still used to put forward discourse-new referents, just as the corresponding forms including bare nominals discussed above (8–9). These sentences, however, represent DE violations. Note that the issue here is two-fold: at the utterance level, learners produce an EPC where the canonical sentence structure (bridging) is an option; at the NP level, they select a demonstrative ([+anaphoric]) to mark referents with a discourse-old status. In Chinese, BNs can—and more often do—felicitously encode such accessible referents.

In (15-16) the context is radically different. The EPC is used to reintroduce referents into the narrative and the entities involved are by consequence discourse-old. It is not the nominal marking with a demonstrative [+anaphoric] which is inappropriate, but the selection of an EPC in these contexts. (Note that in Chinese L1, such reintroduced referents are encoded by nouns modified by a demonstrative, but not appearing within EPCs.)

(15) CH2_FR1 Ránhòu tā cóng jǐngchájú chū-qu de shíhòu
Then 3SG from police.station exit-go SUB moment
yǒu **zhè ge nǚrén** děng tā.
exist this CL woman wait 3SG
Ránhòu zhè ge nǚrén jiù gěi tā yǒngbào
Then this CL woman then give 3SG hug
'Then, when he gets out of the police station, there is this woman waiting for him. Then this woman gave him a hug.'

(16) CH2_FR1 Tā cóng nà ge jǐngchájú chū-lai de shíhòu #
3SG from that CL police.station exit-come SUB moment
jiùshì yǒu **nà ge nǚhái** zài wàimiàn děng-zhe tā
then exist that CL girl be.at outside wait-DUR 3SG
'When he gets out of the police station then there is that girl outside who was waiting for him.'

In sum, DE violations are not frequent in general, and are quite marginal very rare in the case of referent introductions. Recall that in Chinese L1, the DE does not result in a categorical constraint, given that *you*-including a strong determiner are marginally possible (Sect. 1). One could speculate that Chinese L2 French learners

display a native-like behaviour, where the *you*-construction is mainly linked to the expression of prototypical unidentifiable referents but can marginally include identifiable referents. However, granted that EPC patterns *per se* are marked forms, definite-pivot EPCs are even more marked in terms of their frequency. It is very unlikely that learners are exposed to sufficient input to integrate those exceptions into their treatment of definiteness in EPCs, assuming that 'a large and representative sample of language is required for the learner to abstract a rational model that is a good fit to the language data' (Wulff & Ellis, 2018: 75).

At the same time, if French EPCs generally do not manifest the DE (4a), it has been shown that the *il y a* construction statistically tends to include indefinite NPs (Sect. 1). But in French L1, the *il y a* construction can also serve the function of reintroducing referents (as also noted by Lenart & Perdue, 2004). Below is an example from a retelling produced by a French NS produced in the same context as (15–16):

(17) FR1 Il y a la fille qui l' attend à la sortie
It$_{EXPL}$ there has the girl who him waits at the exit
Lit. 'There is the girl waiting for him at the exit'

Overall, French learners' producing in Chinese L2 adopt a pragmatic organization of the information typical of the interlanguage, where the source language influence seems to be playing only a minor role (as discussed in Lena, 2020b). However, the DE 'violations' observed here reasonably seem to follow the IS principles operating in learners' L1. Crucially, such an L1 influence is related to the *inventory of functions* that the EPC can convey in the source language, and which are transposed in the interlanguage.

The presence in the target language of an EPC which has a functional equivalent in one's L1 can trigger its use in the L2, but may also lead to an overgeneralization: French learners are known to overuse EPCs in the L2 (see Turco, 2008; Leclercq, 2008). The proximity between French and Chinese EPCs (Sect. 1) might have the paradoxical effect of making the need for readjustments less critical (a similar speculation is reported in Lambert and colleagues, 2008).

6 The Need for an Integrated Approach to Study the DE Acquisition

Snape and Sekigami (2016), reporting White and colleagues' (2012) concern, note that '[p]ast findings were dependent on spontaneous production data and as a result there were rather infrequent productions of *there* constructions. There may not have been many contexts where a *there* construction was required, thus spontaneous spoken production may not be the most appropriate method to look for DE violations'. (Snape & Sekigami, 2016). To date, researches on DE seems to have been conducted mostly on the basis of GJTs (King et al., 2006; Snape & Sekigami, 2016; White et al., 2012; Yang et al., 2007; Yu & Su, 2011; Zielke, 2016). That being said, I do agree with Snape and Sekigami (2016) that the number of EPCs in elicited

production data is indeed small. Instead of 'dismiss' spoken productions data altogether, the solution is perhaps to integrate different kinds of datasets. While GJTs can provide additional information, they can be misleading when considered alone. It is delicate, for instance, to leave the learner with the task of figuring out the right discourse context for a sentence, even if indications are provided. To illustrate my point, the following are attempts to create informative judgment task items on the basis of White and colleagues' (2012) model (see Fig. 2). Figure 3 presents a natural (occurring in L1 data) target item. Such sentences typically occur at the beginning of the story, that is, without context. Note that the canonical subject–verb (SV) equivalent is possible in Chinese, thus without using an EPC (as shown in Fig. 4). For instance, it might be the case that subjects situate the target sentence as the beginning of a novel instead of an oral narrative (also prompted by the written nature of the stimulus), which could influence their judgment. Though not affecting DE sensitivity altogether, subjects' correction of the target item in favour of an SV sentence would not be informative with regards to their treatment of definiteness in EPCs.

Figure 4 is an example of a not-target-like use of the *you*-construction (from learners' data) which includes a definite NP used to initiate a cataphoric chain. The following context (in brackets) is necessary here to specify the salience of the referent denoted by the pivot, and the referring expression selected to encode the target entity in the contextualizing text might bias the subject's response.

Figure 5 provides an example of non-native-like use of the *you*-construction (also from learners' data) which includes a definite NP in a reintroduction context. In order to provide a sufficient background, the contextualizing text should be as precise as

Anne is feeling sick, so she makes an appointment to see Dr. Salter. She arrives early and the nurse tells her to go right in, saying: *There's the doctor here already.*
How natural is this sentence in this context? If you choose 'unnatural', please correct the sentence. natural not sure unnatural Correction:

Fig. 2 Example of (unnatural) target item used in White et al. (2012) GJT on DE

故事这样开始: 'The story begins like this:' 有一个女孩路过一家面包店。 'There is a girl who passes by a bakery shop'
How natural is this sentence in this context? If you choose 'unnatural', please correct the sentence. natural not sure unnatural Correction:

Fig. 3 Possible natural target item in GJT on Chinese DE

一个很穷的女孩路过一家面包店 。
'A poor girl passes by a bakery shop'

这个面包店有这个商店的人。（商店的人不在的时候，女孩就偷了一个面包。）
'In this bakery shop there is this vendor. (When the vendor is not there, the girl steals a bread)'

How natural is this sentence in this context? If you choose 'unnatural', please correct the sentence.
natural not sure unnatural
Correction:

Fig. 4 Possible unnatural target item (cataphoric definite NP) in GJT on Chinese DE

（有）一个很穷的女孩偷了一家面包店的一条面包。(一个)警察到的时候，(有)一个男孩说是他自己偷（那个）面包，所以（那位）警察把他带到警察局。过了几天，（这位）男孩被放了出来，然后他发现…
'(There is) a poor girl [who] steals a bread from a bakery shop. When (a) police officer arrives, (there is) a boy [who] say that he was the one who stole (that) bread, so (that) police takes him to the police station. A few days later, (this) boy is freed, and found [that]…'

警察局外面有那个女孩一直等他。
'outside [the] police station there is that girl waiting for him'

How natural is this sentence in this context? If you choose 'unnatural', please correct the sentence.
natural not sure unnatural
Correction:

Fig. 5 Possible unnatural target item (reintroduced referent) in GJT on Chinese DE

possible—facing space constraints. It can be questioned whether a short-text context can possibly supply the extra-linguistic background like the one provided by the filmed stimulus used in this study. Then, the question arises as to which forms should be selected to denote discourse-new and discourse-old referents (in brackets), to avoid bias. Plus, as presented in Fig. 5, the context might not be enough centred on the character of the girl to justify its reactivation by using a marked syntactic pattern (i.e. the *you*-construction). This being the case, French learners could reject the unnatural target item even if the same forms are found in their spoken productions.

Finally, in some cases, learners could identify a natural target item correctly (e.g. Fig. 3) without providing any information about their *preference* (that is, one form is accepted but another one would be preferred, or the opposite situation around). Multiple choice items such as Fig. 6 could perhaps provide this kind of information.[12]

While agreeing with White et al. (2012) that traditional GJTs with uncontextualized sentence items are misleading when one analyses speakers' sensitivity to the DE—and IS-motivated phenomena in general—in what precedes I raised some doubts about the possibility of recreating natural occurring contexts by using this

[12] Note, however, that even Chinese L1 speakers' judgment may vary in this respect. That is, *you*-constructions and canonical SV order might commute according to various factors including the discourse register (Liu & Zhang, 2004, Zhou & Shen, 2016: 113). In addition, it is not clear from my data which factors determine NS' choice between quantified nouns and BNs as EPC pivots, in contexts such as the one presented in Fig. 6.

（有）一个很穷的女孩走在路上。她路过一家面包店的时候，她很想吃东西，但是没有钱买。突然偷了一条面包就跑。但是，这时候…
'(There is) a poor girl walking on the road. When she passes by a bakery shop, she really wants to eat something, but she doesn't have any money. Suddenly, she steals a bread and runs away. But, at this moment…'

(a) 有一位警察到了 'There is a police [officer] who arrived'
(b) 有警察到了 'There is [the] police who arrived'
(c) 一位警察到了 'A police [officer] arrived'
(d) 警察到了 '[The] police arrived'

Based on the context provided, please give each sentence a score ranging from 0 (very unnatural) to 5 (very natural).

Fig. 6 Possible multiple-choice item in GJT on Chinese DE

kind of tasks. My intention is by no way that of discredit GJTs as a method to collect learners' data. Rather, what I wish to highlight is the amount of information that might be lost due to IS values that cannot be fully evoked by means of this elicitation technique, and the need for an integrated methodological approach to study the DE acquisition. Though my remarks specifically concern French L1 learners of L2 Chinese, it might be the case for similar problems to arise when considering different L1/L2 combinations.

7 Discussion and Conclusions

In contrast to previous findings on the acquisition of definiteness restrictions in a second language (Lardiere, 2005; Snape & Sekigami, 2016; White et al., 2012; White, 2003, 2008a, 2008b), the current study does report on DE 'violations', that is, French learners of L2 Chinese participating in this study produce definite pivots, contrasting with the DE existing in the EPCs of the target language (i.e. Chinese). This is all the more interesting because the learners that produced the retellings were advanced learners.

Facing an article-less target language, French learners' 'violation' of the Chinese DE manifests itself in the production of *you*-constructions including nouns marked by a demonstrative determiner and marginally proper nouns. Two main tendencies were identified in the learners' data. The increased use of nouns modified by a demonstrative determiner goes along with the reduced use of BNs in the *you*-constructions produced by French learners. Further, an investigation of the overall distribution of the lexical referring expressions in their narratives showed that learners do not produce more nouns modified by a demonstrative determiner in general, nor do they use quantified nouns to a greater extent. Thus, the hypothesis of a general overgeneralization of the indefinite marker and demonstratives—which could bias the occurrence of natural and unnatural EPCs, respectively—is not proven true. That is, the frequency in which these forms are found in EPCs appears to be correlated

with this syntactic context. BNs, on the other hand, are indeed less represented in French learners' overall productions: in EPCs, the reduced use of BNs thus does not solely correlate with the sentence pattern. A study of the general use of Chinese BNs with respect to the other—not exclusively lexical—referring expressions in French learners' interlanguage was beyond the scope of the present contribution. As far as EPCs are concerned, the few instances including a BN are used by learners to introduce uniquely identifiable and inferable referents (i.e. NS' sentences including brand-new indefinite-referent BNs such as *yǒu* ***rén*** *tōu-le nǐde miànbāo* 'there's someone who stole your bread', *yǒu* ***miànbāo*** *zài chē = shang* 'there's [some] bread on the truck' are not found in the L2 corpus).

Turning to DE violations *stricto sensu,* it has been shown that these are marginal for referent introductions. In these cases, the demonstrative [+anaphoric] is used to introduce new—though inferable—referents into discourse. The NP-level marking (i.e. the demonstrative determiner) is inappropriate to denote these referents and incompatible with the selection of a marked sentence structure (i.e. the EPC pattern). Though not preferred by natives, the EPC is still possible in these contexts, as far as new referents are introduced into discourse. In most cases, however, French learners produce definite pivots for referent reintroductions. Here, the nominal marking with a demonstrative [+anaphoric] is appropriate, yet incompatible with the selection of an EPC pattern. This can be explained by considering that in the learners' interlanguage, the *you*-construction is assigned a function that is not available in Chinese L1. In other words, learners are *aware* of the DE that characterize the EPC in the target language, which is shown by the marginal use of definite expressions in referent-introducing EPCs. By contrast, what they seem not to be aware of is that the EPC should not be used in reintroduction contexts. As a consequence, they use the EPC format when discourse-old referents are concerned. Strictly speaking, however, they do not 'violate' the definiteness restriction of the target language, since a different form, with a different function, is operating in the interlanguage.

Observations show that the acquisition of the DE in a second language cannot be acknowledged by having recourse uniquely to a sentence-level approach. Nor it can be accounted for by a single-level analysis. For instance, a reduced number of BNs within learners' EPCs is observed. Even if this results in no DE violation, it has been shown that frequency is informative on how definiteness is treated by learners. In a functionalist perspective of L2 acquisition, it seems more useful to connect the sentence level ('learners' sensitivity to DE') to both the system of referring expressions available in the L2 and the pragmatic function(s) that the sentence pattern can convey. Learners give (at least) two functions to the *you*-construction: that of introducing and reintroducing referents. In Chinese, the *you*-construction cannot be associated to the expression of reactivated referents. In the interlanguage, learners do assign this function to the *you*-construction, which is reasonably interpreted as the result of a negative transfer from the L1. Given the additional function acquired by the *you*-construction, *as a consequence,* learners produce definite pivots, since reactivated referents are definite (i.e. discourse-old) by definition.

8 Limitations and Suggestions for Future Research

Given the nature of the elicited task used to collect data, speakers were not 'forced' to choose one particular linguistic form (i.e. the EPC) to encode (new) reference. As a consequence, referents could be introduced—or reintroduced—by means of other devices, which accounts for the reduced number of EPCs found in the data. It is therefore necessary to point out the limitations of the current study, which included a limited set of EPCs available for analysis and cross-linguistic comparison. Hence, the hypotheses formulated though this paper should be further confirmed by larger samples of data.

As said earlier, the elicitation task was first conceived in order to explore more generally the linguistic strategies for referent (re)introductions in Chinese and French as second languages (Lena, 2017, 2020b, 2020c: 126–137). In hindsight, it would have been useful to clearly space out the scene including the reintroductions of referents by a sequence of distractors, instead of relying on the speakers' understanding of the logical progression of the narrative. That is, the avoidance of an EPC to encode reactivated referents might also originate from the missed perception of an interruption from the preceding point of the story. In other words, if speakers do not conceive a break in the narrative, the referents will be treated as maintained referents, not as reactivated ones. However, sentences such as *yǒu zhè ge nǚhái děng tā* 'there is this girl waiting for him', as in (15), were submitted to three Chinese NS, giving them the appropriate context, and were systematically rejected as unnatural. Finally, in the oral narratives collected using the same stimulus from L1 Chinese L2 French speakers (Lena, 2020b)—which were not considered for the current study—no occurrences of EPCs in the context of reintroduced referents were found. These issues nonetheless demand for a study based on a stimulus that address systematically the referent reintroducing function in L2 Chinese.

Finally, it is worth emphasizing that in production tests triggered by a video sequence like the one adopted in this study, the cognitive load is important (Chini, 2005). Speakers have to introduce and track referents in discourse while managing the narrative cohesion and avoiding ambiguity, all of this in a short time. Many introductions (and reintroductions) are condensed in the two-minute film used as a stimulus. As Ryan (2020) recently suggested, learners' referent tracking in a second language can be influenced by the extra-linguistic context, with informal and unpressured contexts leading to more target-like performances. While elicited production tasks have the undeniable benefit of providing comparable data, they should be integrated with more diversified sources of learners' productions. The combination of different elicitation techniques with spontaneous corpus data is promising, especially when studying IS aspects of the acquisition process.

References

Abbott, B. (1993). A pragmatic account of the definiteness effect in existential sentences. *Journal of Pragmatics, 19*(1), 39–55.

Ahrenholz, B. (2005). Reference to persons and objects in the function of subject in learner varieties. In H. Hendriks (Ed.), *The structure of learner varieties* (pp. 19–64). De Gruyter.

Ariel, M. (1990). *Accessing noun-phrase antecedents.* Routledge.

Bartning, I. (1997). L'apprenant dit avancé et son acquisition d'une langue étrangère. Tour d'horizon et esquisse d'une caractérisation de la variété avancée. *Acquisition et Interaction en Langue Étrangère (AILE), 9*, 9–50.

Bartning, I., & Kirchmeyer, N. (2003). Le développement de la compétence textuelle à travers les stades acquisitionnels en français L2. *Acquisition et Interaction en Langue Étrangère (AILE), 19*, 9–39.

Beaver, D., Francez, I., & Levinson, D. (2006). Bad subject: (Non)-canonicality and NP distribution in existentials. In E. Georgala, & J. Howell (Eds.), *Proceedings of SALT 15* (pp. 19–43.). CLC Publications.

Bentley, D. (2013). Subject canonicality and definiteness effects in Romance *there*-sentences. *Language, 89*(4), 675–712.

Bentley, D., Ciconte, F. M., & Cruschina, S. (2013). Existential constructions in crosslinguistic perspective. *Italian Journal of Linguistics, 25*(1), 15–43.

Bentley, D., Ciconte, F. M., & Cruschina, S. (2015). *Existentials and locatives in romance dialects of Italy*. Oxford University Press.

Berry, R. (1991). Re-articulating the articles. *ELT Journal, 45*(3), 252–259.

Cai, W. (2000). Dai dingzhi jianyu de "you" ziju [*You* sentences with a definite pivot]. *Journal of Zhenjiang Teachers College, 2*, 97–98.

Carroll, M., & Lambert, M. (2006). Reorganizing principles of information structure in advanced L2s. In H. Byrnes, H. Weger-Guntharp, & K. A. Sprang (Eds.), *Educating for advanced foreign language capacities: Constructs, curriculum, instruction, assessment* (pp. 54–73). Georgetown University Press.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and topic* (pp. 25–55). Associated Press.

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* University of Chicago Press.

Chao, Y. R. (1968). *A grammar of spoken Chinese.* University of California Press.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition, 12*(1), 43–64.

Chen, P. (1986). *Referent introducing and referent tracking in Chinese narratives.* Ph.D. dissertation, UCLA.

Chen, P. (2003). Indefinite determiner introducing definite referent: A special use of 'yi "one"+classifier' in Chinese. *Lingua, 113*(12), 1169–1184.

Chen, P. (2004). Identifiability and definiteness in Chinese. *Linguistics, 42(6),* 1129–1184.

Cheng, L.L.-S., & Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry, 30*(4), 509–542.

Chini, M. (2005). Reference to person in learner discourse. In H. Hendriks (Ed.), *The structure of learner varieties* (pp. 65–110). De Gruyter.

Creissels, D. (2019). Inverse-locational predication in typological perspective. *Italian Journal of Linguistics, 31*(2), 37–106.

Crosthwaite, P. (2014). Definite discourse-new reference in L1 and L2: A study of bridging in Mandarin, Korean, and English. *Language Learning, 64*(3), 456–492.

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages: A learner corpus study. *International Journal of Learner Corpus Research, 2*(1), 68–100.

Crosthwaite, P., Yeung, Y., Bai, X., Lu, L., & Bae, Y. (2018). Definite discourse-new reference in L1 and L2: The case of L2 Mandarin. *Studies in Second Language Acquisition, 40*(3), 625–649.

Feng, S. (2019). The acquisition of English definite noun phrases by mandarin Chinese speakers. *Studies in Second Language Acquisition, 41*(4), 881–896.

Givón, T. (1988). The pragmatics of word order: Predictability, importance and attention. In M. Hammond, E. A. Moravcsik, & J. Wirth (Eds.), *Typological Studies in Language* (Vol. 17, pp. 243–284). John Benjamins.

Grannis, O. C. (1972). The definite article conspiracy in English. *Language Learning, 22*(2), 275–289.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*(2), 274–307.

Hole, D. (2012). The information structure of Chinese. In M. Krifka & R. Musan (Eds.), *The expression of information structure* (pp. 45–70). De Gruyter.

Hu, J., & Pan, H. (2007). Focus and the basic function of chinese existential you-sentences. In I. Comorovski & K. von Heusinger (Eds.), *Existence: semantics and syntax* (pp. 133–145). Springer.

Huang, C.-T.J. (1987). Existential sentences in Chinese and (in)definiteness. In E.J. Reuland & A.G.B. Ter Meulen (Eds.), *The representation of (in)definiteness* (pp. 226–253). MIT Press.

Jenks, P. (2018). Articulated definiteness without articles. *Linguistic Inquiry, 49*(3), 501–536.

Karssenberg, L. (2017). French *il y a* clefts, existential sentences and the focus-marking hypothesis. *Journal of French Language Studies, 27*(3), 405–430.

Karssenberg, L. (2018). *Non-prototypical Clefts in French: A Corpus Analysis of "il ya" Clefts*. De Gruyter.

King, E., Steinhauer, K., & White, L. (2006). The definiteness effect in L2 acquisition: What can event-related brain potentials tell us. Paper presented at Generative Approaches to Second Language Acquisition 8, Banff.

Klein, W. (2012). The information structure of French. In M. Krifka & R. Musan (Eds.), *The expression of information structure* (Vol. 5, pp. 95–126). De Gruyter Mouton.

Klein, W., & Perdue, C. (1992). *Utterance structure: Developing grammars again*. John Benjamins.

Lambert, M., Carroll, M., & von Stutterheim, C. (2008). Acquisition en L2 des principes d'organisation de récits spécifiques aux langues. *Acquisition et Interaction en Langue Étrangère (AILE), 26*, 5–10.

Lambrecht, K. (1988). Presentational cleft constructions in spoken French. In J. Haiman & S. A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 135–179). John Benjamins.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge University Press.

Lambrecht, K. (2000). When subjects behave like objects: An analysis of the merging of S and O in Sentence-Focus constructions across languages. *Studies in Language, 24*(3), 611–682.

LaPolla, R. J. (1995). Pragmatic relations and word order in Chinese. In P. Dowing, & M. Nooman (Eds.), *Word order in discourse* (pp. 297–329). John Benjamins.

Lardiere, D. (2004). Knowledge of definiteness despite variable article omission. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *BUCLD 28 Proceedings* (pp. 328–339). Cascadilla Press.

Lardiere, D. (2005). On morphological competence. In L. Dekydspotter, R.A. Sprouse, & A. Liljestrand (Eds.), *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)*. Cascadilla Proceedings Project.

Leclercq, P. (2008). L'influence de la langue maternelle chez les apprenants adultes quasi-bilingues dans une tâche contrainte de verbalisation. Etude de l'expression du déroulement en français et en anglais. *Acquisition et Interaction en Langue Étrangère (AILE), 26*, 51–69.

Leclercq, P., & Lenart, E. (2013). Discourse cohesion and accessibility of referents in oral narratives: A comparison of L1 and L2 acquisition of French and English. *Discours, 12*.

Lena, L. (2017). Les énoncés «présentatifs» chez les apprenants sinophones de français L2. *SHS Web of Conferences* (Vol. 38).

Lena, L. (2020a). Presentational sentences: A study on the information structure of path verbs in spoken discourse. In D. Chen & D. Bell (Eds.), *Explorations of Chinese theoretical and applied linguistics* (pp. 45–81). Cambridge Scholars Publishing.

Lena, L. (2020b). Referent introducing strategies in advanced L2 usage: A bi-directional study on French learners of Chinese and Chinese learners of French. In J. Ryan, & P. R. Crosthwaite (Eds.), *Referring in a second language: Studies on reference to person in a multilingual world* (pp. 164–183). Routledge.

Lena, L. (2020c). *L'introduction des entités dans le récit en français et en chinois: des usages natifs aux variétés d'apprenants*. Ph.D. dissertation, INALCO, Paris & La Sapienza University, Rome. https://hal.archives-ouvertes.fr/tel-03262068/.

Lena, L. (Forthcoming-a). An unbridgeable gap? The treatment of definiteness restrictions in French and Chinese presentational constructions. *Discours, 31*.

Lena, L. (Forthcoming-b). Partition and existence: The case of *you ren* 'there's someone, there are people' in Chinese. In L. Sarda, & L. Lena (Eds.), *Existential constructions across languages: Forms, meanings and functions*. John Benjamins.

Lenart, E. (2006). *Acquisition des procédures de détermination nominale dans le récit en français et polonais L1, et en français L2. Étude comparative de deux types d'apprenant: Enfant et adulte*. Ph.D. dissertation, Université Paris VIII Vincennes-Saint Denis.

Lenart, E., & Perdue, C. (2004). L'approche fonctionnaliste: Structure interne et mise en œuvre du syntagme nominal. *Acquisition et Interaction en Langue Étrangère (AILE), 21*, 85–121.

Lenart, E., & Perdue, C. (2006). Acquisition des procédures de détermination nominale dans le récit: Une analyse fonctionnaliste et comparative. *Zeitschrift Für Literaturwissenschaft Und Linguistik, 36*(3), 70–94.

Leonetti, M. (2008). Definiteness effects and the role of the coda in existential constructions. In A. Klinge, & H. Hoeg-Müller (Eds.), *Essays on determination* (pp. 131–162). John Benjamins.

Leonetti, M. (2016). Definiteness effects: The interplay of information structure and pragmatics. In S. Fischer, T. Kupisch, & E. Rinke (Eds.), *Definiteness effects: Bilingual, typological and diachronic variation* (pp. 66–119). Cambridge Scholars Publishing.

Li, W. (2014). The pragmatics of existential-presentative constructions in Chinese: A discourse-based study. *International Journal of Chinese Linguistics, 1*(2), 244–274.

Li, Y. A. (1996). Definite and indefinite existential constructions. *Studies in the Linguistic Sciences, 26*(1/2), 175–191.

Li, C., & Thompson, S. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Liu, D., & Huang, F. (2015). Muyu wei Fayu, Yingyu zhi huayu xuexizhe de xianding chengfen xide [On the acquisition of determiner elements by English and French learners of Mandarin]. *Journal of Chinese Language Teaching, 12*(1), 83–117.

Liu, A., & Zhang, B. (2004). Pianzhang zhong de wuding mingci zhuyu ju ji xiangguan jushi [The discourse function of indefinite subject sentences and related constructions]. *Special Issue of Journal of Chinese Language and Computing, 14*(2), 97–105.

Lumsden, M. (1988). *Existential sentences: Their structure and meaning*. Routledge.

Lyons, C. (1999). *Definiteness*. Cambridge University Press.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System, 25*(2), 215–232.

McNally, L. (2019). Existential sentences. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics—sentence and information structure* (pp. 281–305). De Gruyter.

Milsark, G. L. (1977). Toward an explanation of certain peculiarities of the existential construction in English. *Linguistic Analysis, 3*(2), 1–29.

Prince, E. F. (1981). Towards a taxonomy of given-new information. In P. Cole (Ed.), *Radical pragmatics* (pp. 233–255). Academic Press.

Rahimi, M., & Youhanaee, M. (2013). Definiteness effect (DE) in English as a second language. *Theory & Practice in Language Studies, 3*(8), 1417–1423.

Rando, E., & Napoli, D. J. (1978). Definites in *there*-sentences. *Language, 54*(2), 300–313.

Ryan, J. (2015). Overexplicit referent tracking in L2 English: Strategy, avoidance, or myth? *Language Learning, 65*(4), 824–859.

Ryan, J. (2020). Under-explicit and minimally explicit reference: Evidence from a longitudinal case study. In J. Ryan, & P. Crosthwaite (Eds.), *Referring in a second language: Studies on reference to person in a multilingual world* (pp. 100–118). Routledge.

Sacks, H., & Schegloff, E. A. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 15–21). Irvington Publishers.

Sarda, L., & Lena, L. (Forthcoming). Existential constructions: In search of a definition. In L. Sarda, & L. Lena (Eds.), *Existential constructions across languages: Forms, meanings and functions*. John Benjamins.

Sasaki, M. (1990). Topic prominence in Japanese EFL students' existential constructions. *Language Learning, 40*(3), 337–367.

Sasse, H. J. (2006). Theticity. In G. Bernini, & M. L. Schwartz (Eds.), *Pragmatic organization of discourse in the languages of Europe* (pp. 257–308). De Gruyter.

Sleeman, P. (2004). The acquisition of definiteness distinctions by L2 learners of French. *Linguistics in the Netherlands, 21*(1), 158–168.

Snape, N. (2009). Exploring Mandarin Chinese speakers' L2 article use. In N. Snape, Y.-K. I. Leung, & M. Sharwood Smith (Eds.), *Representational deficits in SLA: Studies in honor of Roger Hawkins* (pp. 27–51). John Benjamins.

Snape, N., & Sekigami, S. (2016). Japanese speakers' L2 acquisition of the English definiteness effect. In S. Fischer, T. Kupisch, & E. Rinke, (Eds.), *Definiteness effects: Bilingual, typological and diachronic variation*. Cambridge Scholars Publishing.

Sun, J.-L. (2008). Conceptualisation étendue du temps topique dans les narrations des apprenants sinophones en français langue étrangère. *Acquisition et Interaction en Langue Étrangère (AILE)*, *26*, 71–88.

Towell, R., & Hawkins, R. D. (1994). *Approaches to second language acquisition*. Multilingual Matters.

Turco, G. (2008). Introduction et identification d'un référent chez les apprenants francophones de l'italien L2. *Acquisition et Interaction en Langue Étrangère (AILE), 26*, 211–237.

Vallduví, E. (1991). The role of plasticity in the association of focus and prominence. In Y. No, & M. Libucha (Eds.), *ESCOL '90: Proceedings of the Eastern States Conference on Linguistics* (pp. 295–306). Ohio State University Press.

Ward, G., & Birner, B. (1995). Definiteness and the English existential. *Language*, 722–742.

Watorek, M. (2004). Construction du discours par des apprenants de langues, enfants et adultes. *Acquisition et Interaction en Langue Étrangère (AILE), 20*, 129–172.

Watorek, M., Lenart, E., & Trévisiol-Okamura, P. (2014). The impact of discourse types on the acquisition of nominal determination in L2 French. *Linguistik Online, 63*(1), 87–117.

White, L. (2003). Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. *Bilingualism, 6*(2), 129–141.

White, L. (2008a). Definiteness effects in the L2 English of Mandarin and Turkish speakers. In H. Chan, H. Jacob, & E. Kapia (Eds.), *Proceedings of the 32nd Annual Boston University Conference on Language Development* (pp. 550–561). Cascadilla Press.

White, L. (2008b). Different? Yes. Fundamentally? No. Definiteness effects in the L2 English of Mandarin speakers. In L. Dekydspotter, R. A. Sprouse, & A. Liljestrand (Eds.), *Proceedings of the 9th Generative Approaches to Second Language Acquisition Conference (GASLA 2007)* (pp. 251–261). Cascadilla Proceedings Project.

White, L., Belikova, A., Hagstrom, P., Kupisch, T., & Özçelik, Ö. (2012). Restrictions on definiteness in second language acquisition: Affirmative and negative existentials in the L2 English of Turkish and Russian speakers. *Linguistic Approaches to Bilingualism, 2*(1), 54–89.

Wulff, S., & Ellis, N. C. (2018). Usage-based approaches to second language acquisition. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice. (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (pp. 37–56). John Benjamins.

Xia, X. (2009). Dai youding jianyu de "you" zi ju jufa, yuyi ji yuyong fenxi [Syntactic, semantic and pragmatic analysis of *you* sentences with a definite pivot]. *Haiwai Huawen Jiaoyu, 1*, 24–29.

Xu, L. (1995). Definiteness effects on Chinese word order. *Cahiers De Linguistique Asie Orientale, 24*(1), 29–48.

Yang, S., Huang, Y., Gao, L., & Cui, X. (2007). Hanyu zuowei dier yuyan cunxianju xide yanjiu [A study on the acquisition of existential sentences in Chinese as a second language]. *Hanyu Xuexi, 1*, 59–70.

Yu, S., & Su, J. (2011). *There be* cunzaiju xide zhong de dingzhi xiaoying yanjiu [The definiteness effect in the L2 English of Chinese learners]. *Waiyu Jiaoxue Yu Yanjiu, 43*(5), 712–725.

Zielke, M. (2016). The acquisition of the definiteness effect in the L2 European Spanish of L1 German and L1 Turkish speakers. In S. Fischer, T. Kupisch, & E. Rinke (Eds.), *Definiteness effects: Bilingual, typological and diachronic variation* (pp. 447–474). Cambridge Scholars Publishing.

Zhang, X. (2016). A corpus-based study on Chinese EFL learners' acquisition of English existential construction. *Journal of Language Teaching and Research, 7*(4), 709–715.

Zobl, H. (1984). Cross-language generalisations and the contrastive dimension of the interlanguage hypothesis. In A. Davies, C. Criper, & A. Howatt (Eds.), *Interlanguage*. Edinburgh University Press.

Zhou, S., & Shen, L. (2016). Hanyu zhong de "wuding NP zhuyu ju" ji xiangguan de "you" zi chengxian ju [Indefinite subject sentences and their relation with presentational *you* in Chinese]. *Liyun Yuyanxue Kan, 3,* 105–120

Typological and Comparative Approaches

Acquisition of the Chinese Auxiliaries: Insights from Cross-Referential Learners' Corpora of Chinese, English, and Japanese

Zhang Zheng, Sho Fukuda, Laurence Newbery-Payton, Tomohito Ishida, and YaMing Shen

Abstract Chinese epistemic modal verbs "huì" "yào" are difficult to Japanese L1 learners, even C1 level Japanese L1 learners tend to lack "realis/irrealis" modality markers. This tendency is related to Japanese modal/tense/aspect system, namely, non-past tense marker "-ru" covers all irrealis situations. On the other hand, English L1 learners of Chinese use epistemic modal auxiliary verbs "huì" "yào" properly, this tendency might be related to the English auxiliary system which is quite similar to the Chinese auxiliary system. We also discuss overgeneralization by L1 English learners of Chinese. For example, overuse of "huì" by English L1 learners might be caused by the overgeneralization that "huì" is the same as English auxiliary "will". In "TUFS-Shanghai International Studies University learner corpus of English", Chinese learners of English display characteristic overuse of the modal verbs "would" and "will" to express habituality, reflecting overgeneralization of the Chinese modal verb "huì" while Japanese learners of English tend to omit "will" in future contexts.

Z. Zheng (✉)
Institute of Global Studies, Tokyo University of Foreign Studies, Evergreen 201, 3-53-16 Momijigaoka, Fuchu City, Tokyo 183-0004, Japan
e-mail: zhangzheng.apple@icloud.com

S. Fukuda
The University of Toyama, 1535-21, Shimookubo, Toyama-shi, Toyama, Japan

L. Newbery-Payton
Institute of Global Studies, Tokyo University of Foreign Studies, Flat 404, Nakagawa 2-9-9, Tsuzuki Ward, Yokohama, Kanagawa Prefecture 224-0001, Japan

T. Ishida
National Tsing Hua University, Gaocui Rd. East Dist, No. 35, Aly. 4, Ln. 9, Hsinchu City 300, Taiwan

Y. Shen
Institute of Global Studies, Tokyo University of Foreign Studies, 1-27-3 Momijigaoka, Fuchu City, Tokyo 183-0004, Japan

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_13

1 Introduction

Modal verbs have been highlighted as problematic forms for learners to acquire (会 huì, , 要 yào, 能 néng). This paper focuses on the use of modal verbs by learners of Chinese. Analysis reveals distinct trends which can be considered to reflect learners' native languages. First, we show that the influence of L1 on L2 modal verb use is observed in the written production of Japanese and English learners of Chinese (3.1 & 3.2). 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 (3.3). Taken together, the data presented provides evidence for difficulty for native speakers of Chinese, English, and Japanese learning each other's languages.

First, in Chinese epistemic modality, even B1 level Japanese L1 learners of Chinese omit modality markers, i.e., the epistemic modal auxiliary verbs 会 huì, 要 yào, 能 néng. We argue that this is related to the Japanese modal/tense/aspect system, namely, the non-past basic verb form "-ru" covers all irrealis meanings, this leads to Japanese learners fail to mark epistemic meanings.

On the other hand, English L1 learners of Chinese use the epistemic modal auxiliary 会 huì, 要 yào, 能 néng properly. This might be related to the English auxiliary system which is quite similar to the Chinese auxiliary system as shown below. This study focuses on (2), namely, the epistemic uses of these auxiliary verbs.

(1) "can" type:

   a. Ability: 会 huì / 能 néng / 可以 kěyǐ
   b. Probability: 会 huì / 能 néng

(2) "will" type

   a. Volitional: 要 yào/想 xiǎng
   b. Probability: 要 yào

(3) "must/should" type

   a. Obligation: 要 yào/应该 yīnggāi/必须 bìxū
   b. Probability: 要 yào

We also discuss overgeneralization by L1 English learners of Chinese. For example, overuse of "huì" might be caused by the overgeneralization that "huì" is the same as English auxiliary "will". As Newbery-Payton and Mochizuki (2020) shows, in TUFS-Shanghai International Studies University learners corpus of English, Chinese learners of English display characteristic misuse of the modal verbs "would" and "will" expressing habituality, reflecting overgeneralization of the Chinese modal verb huì. In contrast, Japanese learners of English omit will in future contexts. Errors involving "huì" "will" are thus in complementary distribution. In both case studies, we, therefore, observe the overgeneralization by both English L1 learners of Chinese and Chinese L1 learners of English that "huì" = "will".

2 The Study

2.1 Methods

Our study uses cross-referential learners' corpora. By comparing L1 Japanese/English learners of Chinese, we can find the differences in the acquisition of Chinese. We suggest that the linguistic typology of L1 affects second language acquisition.

Figure 1 shows the frequency of top 10 modal auxiliary verbs in National Language Committee Modern Chinese Corpus (http://www.cncorpus.org/). According to previous research on first language acquisition, it suggests that deontic and dynamic meaning is acquired earlier than epistemic meaning (Wells, 1979). Also, studies on second language acquisition have concluded that learners tend to acquire deontic and dynamic meaning earlier than epistemic meaning. Since the acquisition of epistemic meaning is delayed in comparison of deontic and dynamic meaning in both L1 and L2 acquisition, this paper focus on 会 huì, 要 yào and 能 néng, which are most frequently used by Chinese native speakers and each of them has the epistemic meaning and deontic or dynamic meaning. This paper presents an empirical study on the difficulties of using those modal auxiliary verbs in L2 Chinese.

Data is extracted from learners' corpora written by native English speakers and native Japanese speakers at CEFR-based B1 levels. We focus on a significant difference in the production of 会 huì, 要 yào and 能 néng between the corpora of native English speakers and native Japanese speakers. The corpus of Japanese native speakers displays an underuse of the modal auxiliary verbs 会 huì, 要 yào and 能 néng. On the other hand, the corpus of English native speakers does not underuse 会 huì, 要 yào and 能 néng as frequently as native Japanese speakers and shows an overuse of 会 huì, 要 yào and 能 néng. This striking contrast is due to differences in the means of expressing modality in each language.

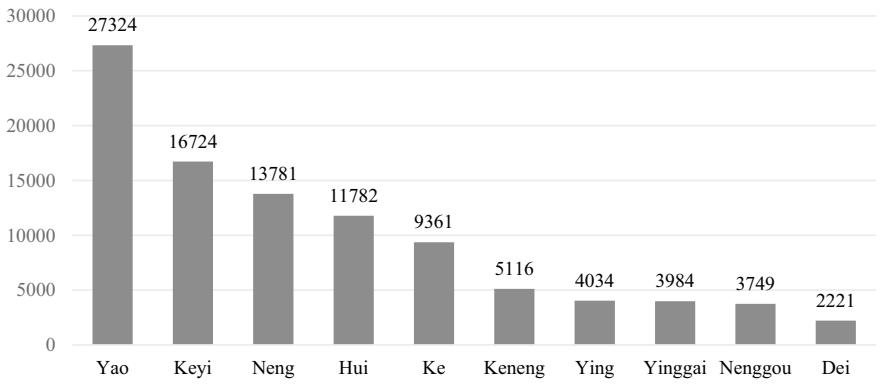


Fig. 1 Frequency of modal auxiliary verbs in CN corpus

Table 1 Data summary

Learner group	Files	Total characters
JLC	286	110,419
ELC	344	33,490
Total	620	143,909

Palmer (2001) points out that there are two ways in which languages deal grammatically with the overall category of modality: the modal system and mood. Both may occur within a single language. In most languages, however, only one of these devices seems to occur or, at least, one is much more salient than the other. Under this classification, both Chinese and English mainly use modal verbs, whereas Japanese mainly uses adhesive verbs and morphology and auxiliary words to express modality (Wen, 2019). Thus, we suggest that the modal systems of Chinese and English are expected to be easier to acquire for speakers of these languages. On the other hand, Chinese and Japanese are expected to be more difficult to acquire for each other.

This study discusses the differences in the acquisition of Chinese by Japanese L1 and English L1, and how the language typology, in this case the different means of expressing modality, affects the acquisition in L2 learning.

2.2 Data

The data used in this study consists of essays by L1 Japanese learners of Chinese (JLC) and L1 English learners of Chinese (ELC). JLC's group was undergraduate students majoring in Chinese at Tokyo University of Foreign Studies. ELC's data was obtained from a TOCFL (Test of Chinese as a Foreign Language) writing pretest provided by National Taiwan Normal University. Both groups of learners are roughly B1 level, which is an intermediate level in the CEFR framework. Initially, 286 JLC essays and 344 ELC essays were collected. A summary is provided in Table 1.

2.3 Data Processing

The essays were proofread by Chinese native speakers with an MA. or Ph.D. in linguistics/language education and sufficient experience in teaching Chinese at university level. Errors and the corresponding corrections were added to the essay texts using a program developed by Yu Kang. Proofread essays clearly indicate errors and corrections so that the errors can be identified within the respective sentences. Results are reported in the following sections.

3 Results and Discussions

3.1 Quantitative Analysis

Tables 2, 3 and 4 display the major categories of errors related to the modal auxiliary verbs 会 huì, 能 néng and 要 yào observed in each group of learners. "Underuse" refers to instances where learners incorrectly omitted a modal auxiliary verb. "Overuse" indicates that deleting a modal auxiliary verb will lead to a correct expression.

Table 2 Frequency of correct and error in 会 huì, 要 yào, 能 néng: Chinese learner corpus (Japanese L1)

		会 huì		要 yào		能 néng	
		Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
Correct		183	35.10%	142	55.00%	123	50.40%
Error	Underuse	*(291)*	*(55.70%)*	*(81)*	*(31.40%)*	*(86)*	*(35.20%)*
	Overuse	*(3)*	*(0.60%)*	*(1)*	*(0.40%)*	*(0)*	*(0.00%)*
	Replace	*(23)*	*(4.40%)*	*(31)*	*(12.00%)*	*(34)*	*(13.90%)*
	Move	*(2)*	*(0.40%)*	*3)*	*(1.20%)*	*(1)*	*(0.40%)*
	Total of errors	319	64.90%	116	45.00%	121	(50.00%)
Total		522	100.00%	258	100.00%	244	100.00%

Table 3 Frequency of correct and error in 会 huì, 要 yào, 能 néng: Chinese learner corpus (English L1)

		会 huì		要 yào		能 néng	
		Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
Correct		667	81.04%	550	89.58%	129	80.12%
Error	Underuse	*(64)*	*(7.78%)*	*(12)*	*(1.95%)*	*(16)*	*(9.94%)*
	Overuse	*(79)*	*(9.60%)*	*(41)*	*(6.68%)*	*(8)*	*(4.97%)*
	Replace	*(12)*	*(1.46%)*	*(11)*	*(1.79%)*	*(7)*	*4.35%)*
	Move	*(1)*	*(0.12%)*	*(0)*	*(0.00%)*	*(1)*	*(0.62%)*
	Total of errors	156	18.96%	64	10.42%	32	19.88%
Total		823	100.00%	614	100.00%	161	100.00%

Table 4 Frequency of correct and error in 会 huì, 要 yào, 能 néng: Chinese learner corpus (Korean L1)

		会 huì		要 yào		能 néng	
		Frequency	Proportion	Frequency	Proportion	Frequency	Proportion
Correct		94	56.60%	89	86.40%	89	81.70%
Error	Underuse	*(61)*	*(36.70%)*	*14*	*(13.60%)*	*13*	*(11.90%)*
	Overuse	*(0)*	*(0.00%)*	*0*	*(0.00%)*	*1*	*(0.90%)*
	Replace	*(10)*	*(6.00%)*	*0*	*(0.00%)*	*6*	*(5.50%)*
	Move	*(1)*	*(0.60%)*	*0*	*(0.00%)*	*0*	*(0.00%)*
	Total of errors	72	43.30%	14	13.60%	20	18.30%
Total		166	100.00%	103	100.00%	109	100.00%

As shown in Tables 2 and 3, the proportion of underuse errors caused by JLC was considerably higher than ELC data. In contrast, the proportion of overuse errors caused by ELC was higher than JLC data. It may ultimately reflect the effects from L1 (Japanese) linguistic typology in the L2 (Chinese) acquisition. For ELC L1 (English) and L2 (Chinese) alike possess modal auxiliary verbs, whereas for JLC L1 (Japanese) and L2 (Chinese) have different forms of expression.

First of all, among the Chinese auxiliary verbs, Japanese L1 learners most commonly make mistakes with 会 huì, 要 yào and 能 néng. Therefore, we will take up these three auxiliary verbs and analyze them.

Table 2 shows that Japanese L1 learners misuse 会 huì the most in comparison to the correct use (319; 64.9%). Furthermore, underuse accounts for over half of the total at 55.7%. In other words, underuse of 会 huì is remarkably common. On the other hand, in 要 yào and 能 néng, the proportion of correct use is the higher, with 55% of all uses deemed correct for 要 yào and 50.4% for 能 néng.

Turning to the Chinese compositions of English L1 learners, unlike the Chinese compositions of Japanese L1 learners, the proportion of correct use of 会 huì (81.04%) is higher than that of incorrect use, and the proportion of underuse is not significantly higher than the other error categories. Another aspect in which English L1 learners differ from Japanese L1 learners is in the overuse of 会 huì. Overuse of hui accounts for 9.60% of errors by ELC, whereas the figure is close to zero for JLC.

To summarize the above discussion, the proportion of misuse 会 huì, 要 yào and 能 néng caused by Japanese L1 learners was larger than English L1 learners. Furthermore Japanese L1 learners, conspicuously underuse 会 huì, while overuse is rarely seen. English L1 learners, on the contrary, show overuse of 会 huì and relatively few examples of underuse. This may be related to the linguistic form of the L1; whether or not there is a form equivalent to Chinese 会 huì in Japanese (no) or English (yes).

To further strengthen this argument, we now turn to the misuse of Chinese 会 huì by Korean L1 learners.

Korean, which is the same agglutinative language as Japanese, behaves in the same tendency as Japanese data: the proportion of errors involving underuse of 会 huì is high, and overuse is low. The Korean L1 leaners are likely at a higher proficiency level than the JLC learners, as they belong to the Chinese department of a university in China. However, the trends are still similar to those of Japanese L1 learners in terms of the tendency of misuse of the auxiliary verb 会 huì. In other words, it is thought that the typology of Japanese has influenced the language of study Chinese.

3.2 Errors in Chinese Modal Auxiliary Verbs in Japanese Learners' Writing

This section will discuss error trends in Chinese modal auxiliary verb 会 huì in Japanese learners' writing. Through the investigation in Sect. 3.1, the underuse showed by Japanese L1 leaners and the overuse showed by English L1 leaners are the most notable error types in our parallel corpus. The following section will specifically discuss those contrastive error types of the Chinese modal auxiliary verb 会 huì caused by Japanese L1 and English L1.

Before discussing data and findings, it is necessary to first establish a basic understanding of 会 huì's meaning and usage. In terms of semantic meaning, we subsumed the use of 会 huì into four classifications as followings.

A. Modality meaning

   (A1) Dynamic modality meaning
   (A2) Epistemic modality meaning

B. Tense/Aspect

   (B1) Future meaning
   (B2) Habitual meaning

"Dynamic" and "epistemic" are the two types of the usual three categories of modality. The conceptual "modality" has been given various definitions and described with different sets of terms by researchers. It is generally accepted that modality refers to "realis/irrealis" (Givón, 1994) and it is associated with three types of meaning: epistemic, deontic, and dynamic. Many Chinese linguists developed their Chinese modality theories from the semantic classifications advanced by Lyons (1977), such as Tsai (2015). For 会 huì, the (A1) dynamic modality meaning expresses the ability or volition of the subject, such as (4a). The (A2) epistemic modality meaning of 会 huì indicates a subjective conjecture about the irrealis event, it is intended to express the proposition that "I think…" or "It will be…" as illustrated in (4b). 会 huì can also be used as a future tense marker as (B1).

Table 5 Semantic classification for the underuse of 会 huì

		Underuse	Correct	Proportion of omission (%)
A. Modality	Dynamic	2	30	6.25
	Epistemic	104	66	61.18
B. Tense/aspect	Future	113	54	67.66
	Habitual	76	25	75.25

According to Yang (2015), the future tense in Chinese is an overt expression, which can be expressed by adverbs and modal auxiliary verbs, such as 会 huì, especially in conditional sentences, which means a possible behavior or state under certain circumstances, such as (4c). The future tense marker 会 huì is also used to express the habitual meaning as (B2), which describes the regular and repeated occurrence of a scene for a period of time (Yang, 2015: pp. 115–158), such as (4d).

(4)

a. 他 会说英语。(Dynamic modality meaning)
He can speak English.
b. 明天他 会来。(Epistemic modality meaning)
I think he will come tomorrow.
c. 如果明天有课，他 会来学校的。
Rúguǒ míngtiān yǒu kè, tā huì lái xuéxiào de. (Future meaning)
If they have a lesson tomorrow, he will come to school.
d. 他经常 会来学校。Tā jīngcháng huì lái xuéxiào. (Habitual meaning)
He often comes to school.

Referring to the semantic classification of 会 huì summarized above, we figured out frequency result for underuse and correct use of 会 huì caused by Japanese L1 learners shown in Table 5.

Table 5 shows that among the uses of 会 huì, there are many misuses of underuse in the "Epistemic" (proportion of misuse: 61.18%), "Future" (proportion of misuse: 67.66%), and "Habitual" (proportion of misuse: 75.25%) categories. In contrast, there are few underuse in the use of "Dynamic" 会 huì (proportion of misuse: 6.25%). In the next part, we will discuss the reasons that cause Japanese students to underuse Hui from the view of modality and tense/aspect.

It has been theorized that acquiring modality will be difficult for second language learners, particularly when the first language and second language use different ways of expressing modality. It has also been pointed out that acquisition of modal verbs is difficult in the order of "dynamic and denotic > epistemic". Through the above observation, we have also confirmed that it is easier to acquire dynamic and deontic meaning than the epistemic meaning for JCL since they omitted the epistemic 会 huì more frequently than the dynamic 会 huì.

In the case of acquisition for the dynamic 会 huì by JCL, based on the correspondence between Chinese and Japanese, we can see that in the meaning of dynamic,

there are corresponding forms in Japanese, such as "-tai" (volition) and "-dekiru, -(rar)eru" (ability). Furthermore, they are semantically clear.

As shown in the following examples the dynamic 会 huì has been omitted in some cases, but data show that such cases are very few, while in most cases, dynamic "会" is used correctly, as in (5a) and (5b).

(5) a. Dynamic: ability

Correct use by Japanese L1:
很少 有 外教 **会** 用 流利的日语，所以 外教 不能 仔细地
Hěnshǎo yǒu wàijiào huì yòng liúlì de Rìyǔ， suǒyǐ wàijiào bù néng zǐxì de
教 学生 英语。
jiāo xuéshēng Yīngyǔ.
Translated to Japanese:
外国人教員 は 日本語 を 流暢 に **話せる**
Gaikokugokyouin-wa Nihongo-o ryūchō-ni hanas-eru
Foreign teachers-NOM Japanese-ACC fluently **speak-can**
こと は ほとんどない
koto-wa hotondo-nai
fact-TOP nearly -Neg
Foreign teachers can rarely speak Japanese fluently.

b. Dynamic: volition

Correct use by Japanese L1:
可是 为了不后悔 的 购物，我 以后 也 **会** 继续 采用 这个 方式。
Kěshì wèile bú hòuhuǐ de gòuwù，wǒ yǐhòu yě huì jìxù cǎiyòng zhège fāngshì
Translated to Japanese:
…今後も この やり方 を 引き続き 採用したい。
...kongo-mo kono yarikata-o hikitsuzuki saiyōshi-tai
Future-also this method-ACC continue **adopt-want**.

We would like to continue to adopt this method in the future.

In contrast, Japanese also has auxiliary verb to express "inference", such as "-darou", just like the use of 会 huì 's epistemic meaning but its addition is not essential.

c. Epistemic

Error use by Japanese L1:
十 岁 以后 外语 学习 能力 <**φ→会**> 下降。(epistemic)
*Shí suì yǐhòu wàiyǔ xuéxí nénglì <**φ→huì**> xiàjiàng.*
Translated to Japanese:
10 歳を 過ぎると、 外国語 の
10sai-o sugiru-to, gaikokugo- no
10 years old-ACC pass-when foreign language-GEN
学習能力 が 下がって行く(**だろう**)。
gakushūnōryoku-ga sagatte-iku (darō).
learning ability-NOM decline-go (**probably-IRR**).

When you turned 10 years old, your ability to learn a foreign language will decline (probably).

"Future", and "habitual" uses of 会 huì have no obligatory corresponding form in Japanese. Epistemic, future, or habitual meaning is often expressed using the non-past basic verb form "**-ru/-u**" covers all irrealis meanings such as (5d) and (5e).

d. Future

Error use by Japanese L1:
有 时候 吃不完 的话，〈**φ→会**〉 浪费 食物。
Yǒushíhou chī-bù-wán de huà, 〈**φ→huì**〉*làngfèi shíwù.*
Translated to Japanese:
食べきれないと、 食べ物 の 無駄に なってしま**う**。
*Tabe-kir-e-nai-to, tabemono-no mudani nat-te-shima-**u**.*
Eat all-can-NEG-CONJ, food- GEN wasted be-PROG-end-**PRS**
If you cannot eat it all, the food will be wasted.

e. habitual

Error use by Japanese L1:
闲 时 我 <**φ→会**> 打开 电视机， 找找 有趣 的 节目。
Xiánshí wǒ <**φ→huì**> *dǎkāi diànshìjī，zhǎozhao yǒuqù de jiémù.*
Translated to Japanese:
暇な ときは テレビを つけて、
Himana-toki-wa terebi-o tsuke-te,
free time-when-TOP TV-ACC turn on- PRS
面 白い 番 組 を 探**す**。
*omoshiroi bangumi- o sagas-**u**.*
interesting program-ACC look for-**PRS**.
In my free time, I always turn on the TV and look for interesting programs.

Also in the misuse of auxiliary verbs “能” and “要”, when they are used in the epistemic meaning, they are omitted.

(6) *Error use by Japanese L1:*
从 这些 观点 来 判断，我 认为 在家里 自己 做饭 比 在 外 边 吃
Cóng zhèxiē guāndiǎn lái pànduàn,wǒ rènwéi zài jiāli zìjǐ zuòfàn bǐ zài wàibian chī
<**φ→要**> 好得 多。
<**φ→ yào** > *hǎo de duō*
Translated to Japanese:
…、家で ご飯を 作る 方 が 良いと 思 う。
Ie-de gohan-o tsukuru-hou-ga yoi- to- omo-u.
home-LOC food-ACC cook-rather-NOM good-QUOT-think-PRS
…I think it is better to cook at home.

(7) *Error use by Japanese L1:*
这个 房子 一定 <**φ→能**> 让 你们 生活 得 更 好。
Zhèige fángzi yídìng <**φ→ néng** > *ràng nǐmen shēnghuó de gèng hǎo.*
Translated to Japanese:
この家に 住むと、きっと あなたたち の 生活 が さらに 良く なる。
Kono-ie-ni sumu-to, kitto anatatachi -no seikatsu-ga sarani yoku-nar-u.
DEM-house-LOC live- CONJ surely you-GEN life-NOM even good-be-PRE
Living in this house will surely make your life even better.

In this section above, we provided evidence for the influence of L1 on use of modal auxiliaries in L2 Chinese. The lack of obligatory morphological forms in Japanese appeared to be related to underuse of modal auxiliaries in L2 Chinese. English native speakers, on the other hand, appeared to have fewer difficulties with 会 huì, presumably due to similarities with “will” and other modal verbs in L1. In this case then, (perceived) similarity between L1 and L2 was beneficial to learners. However, this is

not necessarily always the case. In the next section, we provide supporting evidence for our theory that modal auxiliary use is affected by L1 characteristics, although in this case perceived similarities also lead to errors in some cases.

3.3 Errors in English Modal Auxiliaries in Japanese/Chinese Learners' Writing

In this section we provide evidence of a similar phenomenon in L2 English. Newbery-Payton and Mochizuki (2020) analyzed an L1-to-English translation task conducted by Chinese and Japanese native speakers and discovered a number of distinct error trends. The most notable difference in relation to the previous sections was nonnative use of the modal verb "will". L1 Chinese learners exhibited overuse of "will", whereas L1 Japanese learners exhibited underuse of "will".

Qualitative analysis revealed that L1 Chinese learners overused "will" in situations where the original text expressed habitual meaning. This tendency to use "will" was despite the narrative largely being confined to the past. Examples from the translation task are shown below, together with the relevant sentence from the original Chinese text. The original sentences prominently feature 会 huì, suggesting that Chinese L1 learners associate "will" and "would" with 会 huì. As a result, in contexts where 会 huì is required or preferred, Chinese L1 learners tend to select "will" or "would". In (8) and (9), learners used "will" where no modal verb is required. In (10)–(12), learners used "will" where "would" is appropriate. Note that the learners' translations may not match the original sentences and any other errors have not been highlighted or corrected.

(8) Now whenever I think back that study life, I (will) φ always recall the scene of visiting my teacher. (Ch_57_2013)

如今每当我回想起当时的留学生活时，总是**会**想起每回到老师家里做客时的情景。

Rújīn měidāng wǒ huíxiǎng-qǐ dāngshí de liúxué shēnghuó shí，zǒngshì huìxiǎng-qǐ měihuí dào lǎoshī jiā li zuòkè shí de qíngjǐng.

(9) Ever since that time, whenever I find rice pudding in any Chinese restaurant, I (will) φ order it, just to retaste Professor Hu and his family's hospitality. (Ch_05_2013)

从那以后，每当在中国餐馆里看到八宝饭，我一定**会**点来品尝，不为别的，就只为想再回味一次胡老师和他家人的待客之道。

Cóng nà yǐhòu，měidāng zài Zhōngguó cānguǎn li kàndào Bābǎofàn，wǒ yídìng huì diǎn lái pǐncháng，bú wèi biéde，jiù zhǐ wèi xiǎng zài huíwèi yí cì Hú lǎoshī hé tā jiārén de dàikèzhīdào

(10) Even though, my professor and his families (will) would warmly welcome me into the house. (Ch_40_2013)

（也由于当时老师宿舍里还没有安装电话，所以常常都是无事先告知的突然造访，）但是尽管如此，老师及其家人每次也都一定**会**欣然开门迎客，我也从未尝过闭门之羹。

(Yě yóuyú dāngshí lǎoshī sùshè li hái méiyǒu ānzhuāng diànhuà, suǒyǐ chángchang dōu shì wú shìxiān gàozhī de tūrán zàofǎng,) dànshì jǐnguǎn rúcǐ, lǎoshī jí qí jiārén měicì yě dōu yídìng huì xīnrán kāi mén yíng kè, wǒ yě cóng wèicháng guò bìménzhīgēng.

(11) As soon as I sat down, my teacher (will) underline put some Long Jin tea in a Chinese traditional cup with a cover and then made tea for me using hot water. (Ch_34_2013)

我一坐定后，老师**会**先在一个传统中国式的、带盖子的茶杯里放入一小撮的龙井茶叶,然后从热水瓶里倒出热开水，为我沏上一杯热茶。

Wǒ yí zuòdìng hòu, lǎoshī huì xiān zài yí ge chuántǒng Zhōngguóshì de, dài gàizi de chábēi li fàngrù yìxiǎocuō de lóngjǐng cháyè, ránhòu cóng rèshuǐpíng li dǎo chū rè kāishuǐ, wèi wǒ qī shàng yìbēi rèchá.

(12) Each time on our class, the teacher (will) would roll the quilts for long strip like Western cabbage cake and then make up the bed like a long bench and told me to sit on it to enjoy class. (Ch_49_2013)

每次上课时，老师都**会**将棉被卷成像西式卷心蛋糕似的长条状，然后将床铺整理得如同一条长凳子，要我坐在上面上课。

Měicì shàngkè shí，lǎoshī dōu huì jiāng miánbèi juǎn chéng xiàng xīshì juǎnxīn dàngāo shìde chángtiáo zhuàng, ránhòu jiāng chuángpù zhěnglǐ de rútóng yì tiáo cháng dèngzi，yào wǒ zuò zài shàngmian shàngkè.

One reason Chinese L1 learners may have difficulty distinguishing “will” and “would” is the fact that 会 huì is used regardless of time reference. For example, 会 huì corresponds to “will” in (13a), but in reported speech in (13b), which exhibits so-called “tense shift”, the appropriate modal verb is “would”.

(13)

a. Wo hui hen mang.
   ‘I will be busy.’
b. Zhansan shuo ta hui hen mang
   ‘Zhangsan said that he would be busy.’ (Lin, 2006, p. 18)

Tsai (2015) lists the following uses of 会 huì. The use of 会 huì as a “future modal” (14d) corresponds to “will”; the habitual meaning in (14c) does not show the same correspondence.

(14)

a. yiqian waijiaoguan dou hui fayu. [verb]
   before diplomat all know French
   ‘In old time, all diplomats know French.’
b. yiqian waijiaoguan dou hui shuo fayu. [dynamic modal]
   before diplomat all can speak French
   ‘In old time, all diplomats can speak French.’

c. waijiaoguan changchang hui lai zheli. [deontic modal]
diplomat often tend.to come here
'Diplomats often tend to come here.'
d. waijiaoguan hui changchang lai zheli. [future modal]
diplomat will often come here
'Diplomats will come here often.'
e. shui hui wang dichu liu. [generic modal]
water HUI towards low.land flow
'Water flows to lower places.'
f. waijiaoguan dagai hui lai zheli. [epistemic modal]
diplomat probably Irr come here
'Diplomats will probably come here.'
(Tsai, 2015: 278)

Binnick (2005) offers the following examples of habitual "will", taken from online sources. However, compared to the use of "would" to mark habitual events in the past, habitual "will" is a marked form. Non-past habitual events are more commonly expressed without modal verbs, often with adverbials like "every now and then" or "from time to time" in the examples below. not typically expressed using modal verbs.

(15) The dress is kept in a bag but every now and then she will bring it out for review.
(Binnick, 2005: 339)
(16) From time to time he will yell that he doesn't "want to be managed," but overall, I am the one who is more frustrated. (Binnick, 2005: 341)
(17) Patch is very affectionate. She would prefer to be by your side all day. She will jump up and head butt you to get your attention. (Binnick, 2005: 358)

Habitual "will" is also in competition with other uses of the modal verb. In conditional sentences, "will" is typically interpreted as expressing future time reference. Note also that Binnick's examples are all in the third person. While a comprehensive study of "will" is beyond the scope of the present paper, it appears that volitional readings are favored over habitual readings in the first person. These two factors mean that "will" is disallowed in (8)–(9). Furthermore, as Carlson (2012: 834) states, habitual "will" does not appear with individual-level states. Overall, we can say that "will" is restricted in its habitual uses and thus does not match 会 huì in this regard. Note that in the translation examples habituality is explicitly expressed by adverbial expressions like "always" (12) or "every time" (13). It is possible that such adverbials act as a trigger for Chinese L1 learners' L1-like use of the modal verb.

Given the paucity of habitual "will" in L1 English, it seems reasonable to assume that learners have received minimal exposure to such forms. Therefore, the source of non-nativelike habitual use of "will" by Chinese native speakers more likely stems from the future time reference use, where "will" and 会 huì overlap.

Jarvis and Pavlenko (2008: 178–180) claim that subjective similarities between source and recipient languages (in this case, between Chinese and English) are a

major cause of crosslinguistic influence. Furthermore, when "perceived similarities are numerous enough, they lead the learner to assume a strong similarity between the languages a whole, which in turn leads them to assume additional specific similarities beyond the ones they have already encountered." In the current case study, the objective similarity between 会 huì and "will", namely, their use to mark future time reference, resulted in a perceived similarity between 会 huì and "will" in habitual meaning.

It would be oversimplistic to describe Chinese L1 learners' association of "will" with 会 huì as an example of negative transfer, because in other contexts this in fact leads them to select the correct modal verb. This is evident when comparing errors involving "will" made by L1 Japanese learners. The examples below show how Japanese L1 learners omit "will" when it is required to mark future time reference. Crucially, Chinese L1 learners do not exhibit similar errors. In other words, the presence of an auxiliary verb in L1 to mark future time reference appears to make this use of "will" easier for Chinese L1 learners. Japanese lacks obligatory morphological marking for both future time reference and habituality.

(17) I (φ) will never forget the tender, mild, blissful sweetness of fresh-steamed Babaofan. (Jp_07_2013)
(18) First, I (φ) will tell the memory during my study in Shanghai. (Jp_43_2013).

In summary, the presence of the auxiliary verb 会 huì in L1 appears to be a factor in erroneous use of "will" in habitual contexts, but helps Chinese L1 learners to avoid errors in marking future time reference. In contrast, Japanese lacks comparable forms to "will" in either habitual or future time reference senses. The complementary distribution of errors can be summarized in the below Table 6.

Odlin (2008: 317–318) reports a study by Sastry-Kuppa (1995) investigating the use of "will" as a marker of habitual aspect, including in past contexts, by native speakers of Tamil with a high level of English proficiency. Sastry-Kuppa suggests this distinct non-nativelike use of "will" stems from the extension of a similar future tense marker in Tamil to habitual meaning unrestrained by temporal reference. The similarities to 会 huì in Chinese are striking, as are the non-nativelike uses of "will" that appear.

In conclusion, considered together with the data reported above, these results provide support to the hypothesis that L1 affects the use of modal verbs in L2 (for distinct error trends related to tense and aspect, see Newbery-Payton & Mochizuki, 2020).

Table 6 Issues of overuse and omission involving 会 huì

	Chinese L1		Japanese L1	
Future time reference	会 huì	Nativelike use	(none)	Omission
Habituality	会 huì	Overuse	(none)	Nativelike use

4 Conclusion

This paper has provided evidence from a variety of phenomena that suggest that characteristics of L1 can effect use of modal auxiliaries in L2. This may have a beneficial effect, as in the case of ELC,'s use of 会 huì, or it may lead to overgeneralization, as in the case of CLE's use of "will". In contrast, the lack of an equivalent form in L1 lead to omission of 会 huì and other modal auxiliaries by JLC, but in the case of JLE it in fact prevented over generalization of "will" to habitual meaning. In all cases, analysis of corpus data has allowed us to identify potential areas of difficulty for native speakers of Chinese, English, and Japanese learning each other's languages. This paper's findings, therefore, have direct implications for the teaching of these forms in the language classroom.

References

Binnick, R. (2005). The markers of habitual aspect in English. *Journal of English Linguistics, 33*(4), 339–369.

Carlson, G. (2012). Habitual and generic aspect. In R. Binnick (Ed.), *The Oxford handbook of tense and aspect*. Oxford University Press.

Givón, T. (1994). Irrealis and the subjunctive. *Studies in Language*, *18*, 265–337.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition.* New York: Routledge.

Lin, J. (2006). Time in a language without tense: The case of Chinese. *Journal of Semantics, 23*, 1–53.

Lyons, J. (1977). *Semantics* (Vol. II). Cambridge: C.U.P.

Newbery-Payton, L., & Mochizuki, K. (2020). L1 influence on use of tense/aspect by Chinese and Japanese learners of English. School of Language & Communication, Kobe University. *Learner Corpus Studies in Asia and the World, 4*, 67–93.

Odlin, T. (2008). Conceptual transfer and meaning extensions. In P. Robinson, & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 306–340).

Palmer, F. R. (2001). *Mood and Modality* (2nd ed.). Cambridge: Cambridge University Press.

Sastry-Kuppa, S. (1995, March 3). That's why he will talking for English: the expression of habitual aspect in the English of untutored and low-level tutored Indian speakers. Presentation at the Ninth International Conference on Pragmatics and Language Learning, University of Illinois.

Tsai, W. (2015). On the topography of Chinese modals. In U. Shlonsky (Ed.), *Beyond functional sequence* (pp. 275–294). Oxford: Oxford University Press.

Wells, C. G. (1979). Learning and using the auxiliary verb in English. In V. Lee (Ed.), *Cognitive development: Language and thinking from birth to adolescence.* London: Croom Helm.

Wen, X. (2019). A Syntactic and semantic contrast study of modal expressions across languages. *International Journal of Languages, Literature and Linguistics, 5*(4), 263–268.

Yang, L. (2015). *The subjectivity and subjectification of modal auxiliaries in Chinese.* Doctor thesis, National University of Singapore.

Second Language Acquisition Studies Observed in "The International Corpus of Japanese as a Second Language" (I-JAS) by Chinese Speakers: From the Perspectives of Pragmatic Transfer

Kumiko Sakoda

Abstract This study examines the pragmatic transfer from L1 Chinese to L2 Japanese based on the learners' corpus. The research addresses the following two research questions: (1) What are the specific tendencies among native Chinese learners of Japanese in "request" expressions, compared with French, Spanish, and English learners? (2) Do Chinese speakers have specific tendencies to be affected by their native language? We analyzed the role play data of learners of Japanese whose native languages are Spanish, French, English, and Chinese. We discovered that "suspended clauses (incomplete sentences)", such as "I have a favor to ask you, but …" which are frequently used by Japanese native speakers, are rarely used by the learners of Japanese: Spanish, French, English, and Chinese. However, native Chinese learners use the confirmation expressions more often, such as "is it OK?" at the end of the sentence than other language speakers, which native Japanese hardly use. We then examined pairs of Chinese native speakers by having them work on the same tasks in Chinese. We found they use the confirmation expressions often in Chinese to show politeness. Here are the results of this study: (1) the learners rarely used the "suspended clauses", however, it was not specific to Chinese speakers; and (2) "the confirmation expressions" was observed more frequently among Chinese speakers compared with the other speakers, and it can be considered a negative transfer from learners' native language, Chinese.

1 Importance of Language Acquisition Studies Using Learner Corpora

Learners make errors in the process of acquiring foreign and second languages. Great numbers of researchers and linguists have investigated the mechanisms behind learners' errors (Schachter, 1974, Kellerman, 1979, Kellerman and Sharwood Smith, M. 1986, Odlin, 1989, 2003, Ellis, 2008). In the past, researchers used a contrastive

K. Sakoda (✉)
Morito Institute of Global Higher Education, Hiroshima University, Higashihiroshima, Japan
e-mail: sakodak@ninjal.ac.jp

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_14

analysis called the phenomena "interference" and "language transfer" but the more recent term is "cross linguistic influence" since L1 and L2 are both influenced through the language acquisition process.

Acquisition research on Japanese as L2 focused on language transfer also has been conducted since the 1970s, but no conclusions have been reached yet.

Research by Inaba (1991) targeting English native speakers, using true–false tests, concluded that differences between Japanese and English influenced the acquisition of conditionals. Ikoma and Shimura (1993) also reported pragmatic transfer found in their research on "refusal", which targeted English native speakers using discourse completion tests and showed the tendency of their "refusals" to differ from those of the Japanese native speakers' refusals. On the other hand, other research did not specify the influence of native languages, like Sakoda (1998), who tested the use of demonstratives by native speakers of Chinese and Korean, and Sugaya (2004), who tested grammar, targeting native speakers of German, Russian, English, and other languages.

According to Okuno (2000), results may differ if research targets native speakers of one language or multiple languages. She analyzed the influence of native language and indicated that while 83% of research targeting native speakers of one language considered whether native language transfer had occurred, only 46% of the research concluded there was a possibility of language transfer.

The above result suggests that in order to argue for the effect of native language, it is necessary to examine data of learners with a greater variety of native languages. Compared to English learners' data, there is little data from learners of Japanese, and it is mostly limited to native speakers of English, Chinese, and Korean. Therefore, this study examined whether or not the Chinese language has effects on the acquisition of Japanese, using the "International Corpus of Japanese as a Second Language: I-JAS" which was released in 2020. This research addresses the following two research questions:

(1) What are the specific tendencies among native Chinese learners of Japanese in "request" expressions in Japanese, compared with native French, Spanish, and English learners?
(2) Is there a strong possibility that the trends specific to Chinese speakers are affected by their native language?

2 I-JAS: International Corpus of Japanese as a Second Language

2.1 Conventional Japanese Learner Corpora

Corpora are language resources in which a large volume of texts and utterances (by learners) is compiled into a database. Japanese learner corpora started from the collection of error examples and compositions, as Japanese language education began

to spread. In the 1990s, corpora of compositions and utterances by Japanese learners began to appear.

The KY corpus includes data from the utterances of 90 learners, collected from the OPI (Oral Proficiency Interview), which is an oral ability assessment. The assessment is divided into 9 levels from novice to superior for 30 English, 30 Chinese, and 30 Korean speakers.

The conversational database (cross-sectional survey) of the National Institute for Japanese Language and Linguistics mainly contains 30 min of conversation by 339 Korean native speakers, as well as native speakers of other languages such as English, Chinese, and Indonesian. C-JAS (Corpus of Japanese as a second language) is a longitudinal corpus of Japanese learners. It is a collection of about one-hour utterance data taken over 3 years from 3 Chinese and 3 Korean speakers. There is another data set, LARP at SCU (Language Acquisition Research Project at Soochow University), which collected mainly composition data of 37 Taiwanese college students over 4 years.

However, there are some issues with conventional Japanese learner corpora. First, the number of learners is low. Second, learners' native languages are not balanced. (Most corpora contain data from English, Chinese, and Korean native speakers; data from other languages is insufficient.) Third, the proficiency levels of the learners in the data are not clear. Finally, information on the learners' backgrounds is unavailable.

2.2 Summary and Features of I-JAS

Considering the issues with conventional Japanese learner corpora, we released the "International corpus of Japanese as a second language (I-JAS)" in March 2020. This study analyzes the role play data included in I-JAS. This section gives a summary of the features of I-JAS.

I-JAS has been constructed for the purpose of elucidating the effects on language acquisition processes in different language environments, including differences in learners' native languages. It consists of data collected through 4 years of research, targeting learners in 20 educational institutions in 17 different countries and areas, including Japan. The following explanation summarizes the five main features of I-JAS:

1) I-JAS collects data from a total of 1050 research subjects. It is the largest corpus of its kind, with data collected from 1000 Japanese language learners and 50 native speakers of Japanese. There are 850 JFL learners with 12 different native language backgrounds, 100 Japanese learners studying in educational institutions within Japan, and 50 JSL Japanese learners who do not study at educational institutions, for example, those who are married to Japanese or who work in Japan and with their families. The 12 native languages of the learners in the corpus are English, Chinese, Korean, German, French, Vietnamese, Russian, Spanish,

Indonesian, Hungarian, Turkish, and Thai. These languages were chosen to provide balanced data from the viewpoint of linguistic typology.

The effect on the learning process of different learning environments, whether outside or within Japan, will emerge by comparing the data taken from JFL learners (outside Japan) and JSL learners (those who learn in a classroom environment in Japan). The effect of classroom teaching on the learning process can be clarified by comparing the data taken from 150 JSL learners, of which there are 100 classroom learners and 50 natural environment learners. Furthermore, the data of 50 Japanese native speakers who completed the same tasks is available in I-JAS, and can be compared to the data described above so that the differences between native speakers and learners in each category can be clarified. The role of the learning environment can also be revealed by comparing different learning environments: within or outside Japan, and classroom or natural setting.

2) The second feature is that all the targeted learners have taken two kinds of Japanese proficiency tests, and their results are published. The tests are J-CAT (Japanese Computerized Adaptive Test) and SPOT (Simple Performance-Oriented Test). J-CAT is an adaptive test, to which item response theory is applied (the question items are selected and presented based on the test taker's ability). J-CAT consists of 4 categories, "listening, reading, vocabulary, and grammar", which makes it similar to the Japanese Language Proficiency Test (JLPT). As J-CAT is available online, test takers are able to take the test anytime and anywhere by applying to register. They are also able to obtain the results when they complete the test (https://j-cat.jalesa.org/). SPOT is one of the tests within the TTBJ (Tsukuba Test Battery) developed by Tsukuba University, which consists of 90 questions. It is a fill-in-the-blanks test, where test takers identify one Hiragana within a sentence played to them (https://ttbj.cegloc.tsukuba.ac.jp/). I-JAS includes information regarding learners' scores on both tests, which makes it possible for researchers to objectively judge learners' Japanese language ability and compare data by native language or different learning environment using the data of learners who have equal Japanese ability based on their scores.
3) The third feature is that I-JAS contains data from seven different tasks, including utterances and compositions, which can be used according to the objectives and target of the research. The tasks shown in (3) are included in I-JAS.

(3) Tasks included in I-JAS:

a	Story Telling	oral data	Learner narrates a story based on a 4–5 frame comic
b	Interview	oral data	One-on-one 30-min structured conversation between researcher and learner
c	Role Play	oral data	Learner plays the role of a part-time worker, talking to his/her boss. The worker makes a certain request to the boss in one role play and refuses the boss's request in the other
d	Picture Description	oral data	Learner looks at a drawing and freely describes what appears in it

(continued)

(continued)

e	Story Writing	written data	Learner looks at the same 4–5 frame comic and types the story into a computer
f	E-mails	written data	Learner writes 3 different email responses, including "request" and "declining"
g	Essay	written data	Learner writes his/her own opinion on fast food and home cooking in an essay of approximately 600 characters

Data was collected from 1000 learners and 50 Japanese native speakers for tasks a. ~ e. For tasks f. and g., however, as it would take a long time to complete along with the other tasks within the same day, the tasks were given to learners who volunteered. Learners worked on these tasks at home, not with the researcher.

Two topics, "picnic" and "key", were prepared for tasks a. and e. At first, the pictures were shown to the learners to confirm that they understood the story before they completed the tasks. For story writing (task e.), learners were told they were allowed to write a different story to the one they had narrated first, and that they were allowed to write while they planned the story. These two tasks may be used to study the effect of planning time on language use.

In task c., the role plays were conducted for request and refusal, which are close to realistic communicative contexts.

Since tasks a., ~ c., and e. include many descriptions of movements, a picture description task (task d.) was added, in which " ~ *teiru* (~ing)" can be used to describe an ongoing situation. The same task topics were set for tasks a. (speaking) and e. (writing).

Tasks f. and g. are both composition tasks. The learners were given the tasks beforehand so that they could freely choose a place and time to work on them. Task f. consists of 3 email responses and task g. is a task to write approximately 600 characters based on the learner's opinion. Learners write the task as an entry to an essay contest on the theme, "Our diet: slow food and home cooking".

4) The fourth feature is that I-JAS contains various additional information. For example, for the utterance data, the actual audio data is provided together with a transcript, so that both can be used. The compositions were written only by the learners who volunteered, and the number of compositions data is not the same as that of the utterances. Note that there is only data from those learners who volunteered to complete the tasks. Information regarding the time taken and references used are published along with the compositions themselves.

   The second type of additional information is the learners' background information. Since learners' home country, environment, and other factors are different, we conducted a questionnaire on their background beforehand, and published the information related to their Japanese learning, such as their learning environment (classroom, natural setting, etc.), family (family structure, language used, etc.), part-time job (whether they use Japanese at work or not), and their Japanese learning style.

The third type of additional information is utterance and composition data collected from 50 Japanese native speakers who completed the same tasks. They are a well-balanced group for comparison with a mix of male and female speakers in their twenties to fifties.

5) The last feature is that morphological information is added to I-JAS. I-JAS uses the corpus search application, "CHUNAGON", which enables not only string searches, but also searches using morphological information. This makes it easier to conduct quantitative analysis of Japanese learner corpus data. It also enables analytical research from various fields, such as vocabulary and morphology studies.

This section has given an overview of I-JAS and its 5 main features. The next chapter will focus on analyzing the data from request expressions used in the role play in I-JAS, in order to examine whether or not Chinese learners of Japanese are influenced by their native language.

3 Previous Studies on Japanese Learners' "Request and Refusal"

Research on Japanese learners' "request and refusal" increased in the 1990s together with the growth in the number of language acquisition studies from the viewpoints of social linguistics or pragmatics. For instance, Ikoma and Shimura (1993) targeted English native speakers studying Japanese as a foreign language, comparing their refusal expressions in English and Japanese (L2) with that of Japanese native speakers. The influence of English was observed in the learners' Japanese, which indicated the possibility of pragmatic transfer from their native language. Specifically, it was reported that Japanese native speakers have a tendency to suggest an alternative idea when turning down a request and use incomplete sentences, such as "… *desuga* [however, but…]", when turning down their superior's request. While English native speakers do not suggest alternative ideas and directly turn down their superiors without using incomplete sentences. It was suggested that this tendency indicates a pragmatic transfer from English expressions.

Izaki (2000) researched deviation and unsuitability observed in requests made by French learners of Japanese. The study showed that Japanese native speakers provide a preliminary and introductory step before requesting, such as "In fact, I have something that I'd like to ask you …", whereas, in the case of the learners, such a preliminary step is almost never provided. Izaki observes that while "asking for changes" is regarded as a "request" in Japan, learners have a tendency to consider it as "negotiation", thus suggesting that request expressions are influenced by social and cultural differences between Japan and France.

Other studies on request expressions, such as Kashiwazaki (1992), Samejima (1998), and Lee (2008), suggest that Chinese learners of Japanese have a tendency to use "complete sentences" and "direct expressions".

However, there are some issues with the past studies. First, much of the data was collected from Japanese learners of a single native language. As pointed out in Sect. 1, it is necessary to target Japanese learners of several native languages to discuss the effect of the native language. The results of the previous studies may, therefore, not be accurate. Secondly, many of the studies used written data from conversation completion tests, which may not be the actual language that learners of Japanese use.

4 Data Analysis of I-JAS Based on "Request" Role Play

This chapter will introduce the two research studies (Sakoda et al., 2017) which use.
I-JAS data, taking into consideration the issues indicated in the previous chapter.

4.1 Suspended Clauses by L2 Learners of Japanese

The data for analysis was taken from a total of 60 targeted learners studying Japanese overseas: 15 French native speakers, 15 Spanish native speakers, 15 Chinese native speakers, and 15 English native speakers. As a comparison group, the data of 15 Japanese native speakers was used. The learners' Japanese levels were measured based on the results of Japanese language ability tests. The 15 learners in each native language group were chosen after their Japanese ability had been shown to be homogeneous, based on statistical tests.

(4) Extract from an I-JAS role play instruction card

> You are working as a part-time staff member at a Japanese restaurant. (…) Now you work three days a week. However, you want to change to working two days a week as you have got busier. Please tell the restaurant manager that you want to change the number of workdays from three days to two days and get his/her permission. (Please indicate to the researcher when you are ready to begin.)

The task is a role play in which a learner and a researcher (a Japanese native speaker) converse one on one. The content of the role play card is shown in (4). The card is written in the learner's native language. After having the learner read the card silently, we confirmed that the learner understood the content and then began the role play. The Japanese native speaker (researcher) played the role of the restaurant manager, and the learner played the role of the part-time staff member and student. The researchers in their roles as a manager would not immediately

accept the student's request and would come up with several reasons to dissuade the student from decreasing their shifts in order to continue a back and forth discussion. Conversations were recorded and transcribed.

The first half of the role play where the part-time member of staff begins to talk to the manager, brings up the topic, and mentions his/her request, was divided into three sections, "introduction, precondition, and request", for analysis. The following is the example as shown in (5).

(5) Introduction (A), e.g. *Anoo, go-soodan ga arundesu ga…* [Excuse me, well, I have something to discuss with you, and…]

Precondition (B), e.g. *Ima, shuu mikka hataraite irundesu kedo…* [Right now, I work three days per week, but …].

Request (C), e.g. *Shuu futsuka ni kaetaindesu kedo…* [I would like to change to two days per week, and …].

The sentences used for each utterance were categorized as in (6).

(6) Suspended clauses (incomplete sentences),

e.g. *Ohanashi shitai koto ga arundesu ga…* [I have something to discuss with you, and …].

Question sentences: e.g. *Ima sukoshi yoroshii deshoo ka.* [Do you have a minute now?].

Declarative sentences: e.g. *Tenchoo, hanashiga arimsu.* [Mr. –-, I need to talk to you.]

(7) Breakdown of sentence types in Introduction section (A) by each group of native speakers (Sakoda, 2016: 105) Suspended clause: SC, Question sentence: QS, Declarative sentence: DS

Japanese native speakers			French native speakers			Spanish native speakers			English native speakers			Chinese native speakers		
SC	QS	DS	**SC**	QS	DS	**SC**	QS	DS	**SC**	QS	DS	**SC**	QS	DS
90%	0%	10%	**17%**	50%	33%	**33%**	33%	33%	**27%**	55%	18%	**27%**	18%	55%

(8) Breakdown of sentence types in Request section (C) by each group of native speakers (Sakoda, 2016: 106) Suspended clauses: SC, Question sentence: QS, Declarative sentence: DS

Japanese native speakers			French native speakers			Spanish native speakers			English native speakers			Chinese native speakers		
SC	QS	DS	**SC**	QS	DS	**SC**	QS	DS	**SC**	QS	DS	**SC**	QS	DS
73%	20%	7%	**27%**	53%	20%	**13%**	47%	40%	**7%**	73%	20%	**0%**	80%	20%

The results of (7) and (8) indicated the following:

① A high proportion of the sentences used suspended clauses, such as "I am sorry to bother you, but …" by Japanese native speakers in both the introduction section (A) and the request section (C). However, the proportion of such sentences is very low among learners, regardless of their native language.

② On the other hand, a high proportion of the sentences that learners use in both the introduction section (A) and the request section (C) are questions or declarative sentences. Examples are shown in (9) and (10).

(9) Examples of question sentences:

a. *Tenchoo, ima hima desu ka.* [Mr. –-, are you free now?] (Introduction part (A), French native speaker)
b. *Tenchoo, jikan ga arimasu ka.* [Mr. –-, do you have time?] (Introduction part (A), Chinese native speaker)
c. *Futsuka no hi-dake iideshoo ka.* [It is OK to have two days only?] (Request part (C), Spanish native speaker)

(10) Examples of declarative sentences:

a. *Onegai ga arimasu.* [I have a favor to ask you.] (Introduction part (A), French, English, Spanish, and Chinese native speakers)
b. *Hanashi ga arimasu.* [I have something to talk to you about.] (Introduction part (A), Chinese native speaker)
c. *Shuu futsuka shitai to omoimasu.* [I'd like to work twice a week.] (Request part (C) Chinese native speaker)
d. *Mikka-kan wo futsuka-kan ni shite kudasai.* [Please change three days to two days.] (Request part (C) Spanish native speaker)

This result shows that Japanese learners, when compared with Japanese native speakers, have a tendency to use declarative and question sentences but not suspended clauses (incomplete sentences) in the request role play. This tendency is commonly observed among learners despite different native languages: French, Spanish, English, and Chinese. This is, therefore, not a feature specific to Chinese native speakers, so we cannot assume the possibility of influence from their L1, Chinese.

4.2 Expression of Confirmation by L2 Learners of Japanese

It was shown in the previous section that the low rate of suspended clauses use in the request role play was not a phenomenon unique to native speakers of Chinese. Is it true to say that no specific tendencies can be observed among Chinese learners of Japanese? In the request section (C) of Chinese speakers' utterances, expressions like (11) were prominent.

(11) …. since I am busy, working three times a week is not possible for me, and I'd like to work twice. Is it OK?

The learner expressed his/her wishes by saying "I'd like to work twice", then tried to make sure by saying "Is it OK?" For Japanese native speakers, this expression of "confirmation" gives the impression of forceful reminding, and thus has a high risk of leaving a highly unpleasant impression on the listener if they are older or one's superior. We analyzed the data to see if such "confirmation" expressions can be observed particularly among Chinese speakers (Sakoda et al., 2017).

The utterances of 90 learners in the request role play were analyzed in total; 20 French speakers, 20 Spanish speakers, 20 English speakers, 30 Chinese speakers, and 15 Japanese native speakers. The Chinese group was divided into three levels according to their J-CAT scores: top, upper, and lower levels. The other groups were divided into upper and lower levels. The average J-CAT scores in each native language group are shown in (12). The Japanese abilities of learners are not homogeneous; the scores of the upper and lower French and Spanish speaker groups are lower than those of the English and Chinese speaker groups. The scores are almost identical, however, between the Spanish and French groups and the English and Chinese groups, respectively.

(12) Average scores of J-CAT in each native language group (n = 10) (Sakoda et al., 2017: 56)

Level	Lower	Upper	Highest	Level	Lower	Upper	Highest
French	127.4	201.2	-	**English**	185.1	234.4	
Spanish	126.5	198	-	**Chinese**	184.6	234	296.6

The chart in (13) shows the number and proportion of learners who used "confirmation" expressions in the request role play.

(13) The number and proportion of "reminder" expressions in the role play utterances

Speakers	Spanish	French	English	Chinese
Lower level (n = 10)	**1 (10%)**	**2 (20%)**	**1 (10%)**	**5 (50%)**
Upper level (n = 10)	**0**	**1 (10%)**	**0**	**1 (10%)**
Highest level (n = 10)	-	-	-	**1 (10%)**

The results above show that the rate is higher among the lower-level Chinese speakers, half of whom used confirmation expressions. Learners in the upper and highest level were still using the confirmation expressions, which indicates that Chinese speakers specifically have a tendency to use this expression, compared with learners of other native languages. Example utterances are given in (14) ~ (16).

(14) *Ettoo, ee, ni, nichi nichi ga hatarai desu, iidesu ka.* [Well, let's see, two, I want to work for two days. Is it OK?] (Lower level).

(15) *Isshuukan, tabun, futsuka, kite, anoo, ikagadesu ka.* [Per week, maybe two days, I'll come. How about this?] (Upper level).

(16) *Isshuukan futsuka ni natte hoshii to omoimasu, iidesuka.*[Per week, I'd like it to be two days. Would it be all right?] (Highest level).

The language form changed from a regular form, "how about this?" uttered by a learner of the upper-level group, to a polite form, "Would it be all right?", spoken by a learner of the highest group. However, in both cases, it sounds as if the speaker is demanding the listener's response, which gives an unpleasant impression.

Sakoda et al., (2017) conducted an experiment to ascertain whether similar expressions were used in the utterances of native speakers of Chinese. Analysis of conversations by 12 pairs of Chinese native speakers showed that 8 out of 12 speakers (75%) used expressions equivalent to Japanese confirmation expressions. Examples are given in (17) and (18).

(17) 所以最近有点忙, 啊, 没法再延续之前, 三天, 一周三天的打工了, 想改成两天, 您看可以吗?

Therefore, I've been busy recently, I can't continue like before working three days, three days a week. I'd like to change it to two days, do you think that's all right?

(18) 店长最近由于我学业比较忙, 我想把打工时间由三天改到两天,行吗?

Boss, I've been pretty busy recently with my studies, I'd like to change my work from three days to two days, is that okay?

From this result, we were able to infer that Japanese learners of Chinese speakers state their wishes first, using the Japanese format "I want to ~ ", then try to show respect to the listener, the restaurant manager, by "asking for permission" or "asking their opinions".

In Japan, however, speakers tend to avoid directly conveying their wishes to superiors or people older than them. Repeating those wishes using confirmation expressions can frequently leave an even ruder impression. In general, in Japan, frequent use of this kind of "confirmation expressions" can be regarded as pushy. As opposed to "confirmation expressions", Japanese native speakers frequently use "suspended clauses", where they omit the last part of the sentence and have the superior or older listener guess their wishes. An example of this kind of suspended clauses is as follows: "If possible, I would like to change to two days a week, but …". Thus, expressions of "politeness" in Japanese and Chinese clearly differ. Paying attention to grammatical accuracy alone is insufficient. A lack of understanding of differences in communication between different cultures may inadvertently leave a bad impression and lead to trouble in relationships.

5 Discussion on Pragmatic Transfer and Summary

We summarize and discuss findings from the pragmatics data analysis of the two issues shown below:

(1)' What are the specific tendencies among native Chinese learners of Japanese in "request" expressions in Japanese, compared with French, Spanish, and English learners?
(2)' Is there a strong possibility that specific tendencies are affected by learners' native language, Chinese?

Using data from the I-JAS request role play, we compared native speakers of French, Spanish, English, and Chinese. We discovered that "suspended clauses", such as "I have a favor to ask you, but …" and "I'd like to ask you to let me … but…", which are frequently used by Japanese native speakers, are rarely used by learners. Therefore, it became clear that this was not a tendency specific to Chinese native speakers. Some previous studies point out that the use of suspended clauses is rare, but many of them reached that conclusion by targeting learners of only one native language. Our research, on the other hand, reached the same conclusion after comparing and analyzing data from learners of several native languages, divided by Japanese proficiency levels. Although the previous studies indicated that this tendency (of not using suspended clauses) might be affected by the learner's native language, our research is significant because it demonstrated that this may not be correct. We found that native speakers of French, Spanish, English, and Chinese all struggled to use "suspended clauses", and there were no notable differences among the speakers of these languages. However, the effect of native language cannot be discounted unless research clarifies the tendencies of learners whose native language is Korean or Turkish. Predicates in these languages are located at the end of sentences, which is similar to Japanese sentence structure.

As for the confirmation expression such as "Is it OK?", we found that this expression was observed more frequently among Chinese native speakers, compared with the speakers of other languages. We then examined pairs of Chinese native speakers by having them work on the same tasks in Chinese to see if they showed the same tendency. Chinese native speakers were found to frequently use confirmation expressions such as "Is it OK?" or "Would it be all right?", after expressing their requests and wishes. Therefore, it is possible that the use of these confirmation expressions is affected by the Chinese language. While this kind of expression can be considered as a form of politeness in China, it gives the impression of disrespect in Japan. The use of confirmation expressions in Japanese by Chinese speakers can be considered a negative effect on learners' native language, Chinese. This suggests that learning a second or foreign language is not just a matter of learning the language forms, but also requires the teaching of pragmatics, including awareness of differences in politeness between the culture of the target language and that of the native language. We would like to continue our studies in this area.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP24251010, JP16H01934, and JP19KK0055.

References

English References

Ellis, R. (2008). *The Study of Second Language Acquisition* (2nd ed.). Oxford University Press.
Kellerman, E. (1979). Transfer and non-transfer: Where are we now? *Studies in Second Language Acquisition, 2*, 37–57.
Kellerman, E., & Sharwood Smith, M. (Eds.). (1986). *Cross-linguistic influence in second language acquisition.* Pergamon.
Odlin, T. (1989). *Language transfer*. Cambridge University Press.
Odlin, T. (2003). Cross-linguistic influence. In Doughty, C., & Long, M. (Eds.) *The handbook of second language acquisition.* Malden, Mass.: Blackwell.
Schachter, J. (1974). An error in error analysis. *Language Learning, 27*, 205–214.

Japanese References

Inaba, M. (1991). The range of meaning of Japanese conditionals and interlanguage structure. *Journal of JapAnese Language Teaching, 75*, 87–99.
Izaki, Y. (2000). Differences in systems of priority and diverging expectations in conversations involving requests: A study of interactions between Japanese native speakers and French learners of Japanese. *Journal of Japanese Language Teaching, 104*, 79–88.
Ikoma, T., & Shimura, A. (1993). Pragmatic transfer from English to Japanese. *Journal of Japanese Language Teaching, 79*, 41–52.
Kashiwazaki, H. (1992). Discourse analysis of initiation of conversations: A focus on actual expressions of request and demands. *Journal of Japanese Language Teaching, 79*, 53–63.
Lee, Y. (2008). Contrastive research on verbal behavior in Chinese and Japanese: The perspective of politeness. *CAHE Journal of Higher Education, Tohoku University, 3*, 117–129.
Okuno., Y. (2000). Attestation of language transfer in second language acquisition: Issues raised by previous research. The Chugoku-Shikoku Society for the Study of Education. *Education Research Proceedings, 46*(2), 384–389.
Sakoda, K. (1998). *Interlanguage research: The use of the demonstratives ko-, so- and a- by learners of Japanese.* Keisuisha.
Sakoda, K. (2016). Requests observed in learners' role plays. *Japanese Language Education in Europe, 20*, 102–107.
Sakoda, K., Ying, S., & Zhang, P. (2017). A cross-linguistic study of requests in a role-play by Chinese speakers: Pragmatic transfer observed in expressions of emphasis by JFL learners. *Journal of Japanese and Chinese Linguistics and Japanese Language Teaching, 10*, 50–63.
Sakoda, K., & Li, X. (2020). A study on Japanese acquisition of Chinese native speakers from the perspective of intercultural communication — Analysis of the misuse of "Correctness" and "Appropriateness." *Foreign Language Research in Northeast Asia, 4*, 3–7.
Sakoda, K., Ishikawa, S., & Lee, J. (Eds.) (2020). Introduction t*o the Japanese Learner Corpus I-JAS.* Kuroshio.
Sakoda, K. (2020). *The second language acquisition studies for Japanese education* (2nd ed.). ALC.
Samejima, S. (1998). The Acquisition of fixed expressions and sentence-ending expressions by learners of Japanese: The cases of requests, refusals and apologies by Chinese speakers. *Journal of Japanese Language Teaching, 98*, 73–84.
Sugaya, N. (2004). Research on the acquisition of aspect by learners of Japanese through grammatical tests: A consideration of the role of L1. *Journal of Japanese Language Teaching, 123*, 56–65.

Word Order Typology and the Acquisition of Chinese "Verb + Resultative" Compound Verbs: Insights from Brain Science and Learner Corpora

Haining Cui, Hyeonjeong Jeong, Yoshihiro Mochizuki, and Keiko Mochizuki

Abstract This chapter will focus on the acquisition of Chinese "Verb + Complement" resultative compound verbs. Chinese resultative compound verbs are among the most difficult forms for Japanese first language (L1) speakers, although Japanese also has a rich system of compound verbs. This phenomenon is observed in the L1 Chinese Japanese learner corpus. It is difficult for Chinese L1 speakers to learn Japanese compound verbs; non-use is observed rather than misuse. In light of these L2 acquisition difficulties, we present how word order typology affects second language (L2) acquisition in the fields of brain science and learner corpus research and propose Chinese pedagogy based on learners' L1.

1 Introduction

The linear combination of words in a sentence, clause, or phrase is organized following the language's default grammatical sequence. Such a default grammatical sequence is called word order, which plays a vital cue in facilitating language users' encoding and interpreting of who did what to whom during language processing.

H. Cui · H. Jeong
Graduate School of International Cultural Studies, Tohoku University, 41 Kawauchi, Aoba-ku, Sendai 980-8576, Japan
e-mail: cuihn1989@gmail.com

H. Jeong
e-mail: jeong@tohoku.ac.jp

Y. Mochizuki
Graduate School of Medicine, Yokohama City University, 3-9 Fukuura, Kanazawa word, Yokohama City, Kanagawa, Japan
e-mail: dr.jimmyshen@gmail.com

K. Mochizuki (✉)
Institute of Global Studies, Tokyo University of Foreign Studies, 3-11-1 Asahicho, Fuchu City, Tokyo, Japan
e-mail: mkeiko@tufs.ac.jp

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_15

Chinese has a default subject–verb–object (SVO) word order. In contrast, in some languages like Japanese and Korean, the basic word order is subject–object–verb (SOV), and the verb is assigned to the final position of a sentence. Moreover, word order rules manifest at a sentence level and constrain other linguistic elements in a sentence. When the word orders are different between L1 and L2, learners tend to require effort to comprehend L2.

The typology of word order in syntax also reflects in a word structure (Shibatani 1990; Kageyama 1996, 2009, 2010, 2016a, 2018; Kageyama and Kanzaki 2014). Although both Chinese and Japanese have a "compound verb" system, their word syntax is different. This mutual difficulty in learning compound verb systems in Chinese and Japanese is due to different word orders: SVO in Chinese and English, SOV in Japanese. The syntactic word order is also reflected in lexical word order: the <Verb + Resultative> compound verb type in Chinese versus the <Object Clause + Verb> compound verb type in Japanese.

The next section will examine how syntactic differences between Japanese and Chinese lead to differences in aspectual compound verbs in the two languages.

2 Differences in the Structure of Chinese and Japanese Aspectual Compound Verbs

Aoki (2013) suggest that Japanese aspectual compound verbs developed in the form $_{VP}$[$_{VP}$[argument + V1] V2]], with the object complement of V2 compounding with V1. Thishistorical change corroborates the fact that the SOV syntactic structure of Japanese is reflected in the internal structure of compound verbs and the prominence of compound verbs with an OV structure, where V1 functions as the object of V2.

Tang and Hsu (2015a, b), Mochizuki and Shen (2011) state that in Chinese, an SVO word order, verb compounding with an object complement does not occur. In Chinese, SVOC resultatives changed diachronically to S[VC]O constructions, resulting in the prominence of resultative compound verbs. Therefore, in Chinese, inception (start to- "*-kakeru*, *-dasu*, *-hajimeru"* in Japanese), continuation (*-makuru*), and incompletion "fail to (*-sokonau/-sonjiru*)" "forget to (*-wasureru*)" cannot be expressed using compound verbs.

The internal structure of aspectual compound verbs in both Chinese and Japanese reflects the syntactic structure of each language.

Japanese syntactic compound verbs are formed from compounding of the type [[V1(object of V2)] + V2]] (Kageyama, whereas Chinese compound verbs are formed from the VC (V1 + resultative complement) structure.

Therefore, Chinese does not allow "object complement" type compound verbs while Japanese has a rich system of "object complement" type compound verbs.

(1) a. inception (*-kakeru*, *-dasu*, *-hajimeru*)

b. continuation (*-makuru*, *-tsuzukeru*),

Table 1 Differences in phrase structure, compound verb structure and word order in Japanese and Chinese

	Japanese	Chinese
1. **Structure of verb phrase**	**1. OV word order** **2**. verb phrase structure: Verb phrase head: right-sided Resultative predicates are not permitted after the verb	**1. VO word order** **2**.verb phrase structure: Verb phrase head: left-sided Resultative predicates are placed after the verb
	2. Temporal order principle **(1) relation of causation**: *obore-shinu* "drown" 溺れ-死ぬ **(2) antecedent-result relation**: a. *tabe-nokosu* 食べ-残す eat-leave "leave some food left uneaten" b. *ure-nokoru* 売れ-残る sell-remain "remain unsold" cc. *tabe-nokosu* 食べ- 残すeat-leave "leave some food left uneaten" *ure-nokoru* "remain unsold"	**2. Matching phrase structure and word structure principle** → while object complement type compound verbs do not exist, **subject complement type compound verbs** do exist e.g <-完 wán>, <-上 shàng>, <-错 cuò>, <-多 duō>, <-少 shǎo>, <-遍 biàn>

c. incompletion: fail to/miss (*-sokonau*, *-sonjiru*, *-sobireru*, *-shikaneru*) forget to do something (*-wasureru*), repetition (*-naosu*)

Unlike Japanese, Chinese does not allow compounding with an object clause (Table 1).

The following examples in (2) show how Japanese object complement type compound verbs are expressed in Chinese. For example, object complement type compound verbs expressing "inception", "continuation", "incompletion" and "repetition" are expressed in Chinese using a verb phrase with an object clause of the structure [$_{VP}$V2 + [$_{IP}$ …V1…]], or with transitive sentences expressing the impossibility of past events with the structure [IP 没能 méi néng [$_{VP}$…V1…]]. Aspectual compound verbs cannot be used to express these meanings (Mochizuki and Shen 2011).

(2)
a. -hajimeru「～始める」 ([VP 开始 kāishǐ [IP…V1 …]]) “start V1~ing”

b. –tuzukeru 「～続ける」 ([VP 继续 jìxù [IP…V1 …]]) “continue V1~ing”

c. –sokonau 「～損なう/損ねる」 ([IP 没能 méi néng [IP…V1 …]]) “fail to V1”

d. –wasureru「～忘れる」 ([VP 忘 wàng [IP…V1 …]]) “forget to V1”

e. –naosu 「～直す」 ([VP 重新 chóng xīn [IP…V1 …]]) “V1 again”

Why, then, do object complement type compound verbs not exist in Chinese? For example, why is the compound verb <*忘写 wàng xiě, forget-write>, corresponding to “*kaki-wasureru* “書き-忘れる, forget to write” in Japanese, impossible in Chinese? The reason is that in Chinese “resultative” or “antecedent-result” type compound verbs, which follow the “temporal order” of events, are most favored. Tai (1985: 50) states that “SVCs in Chinese, proposes the “principle of temporal sequence”, “the relative word order between two syntactic units is determined by the temporal order of the states which they represent in the conceptual world.” Li (1993: 480, 502) also appeals to a “Temporal Iconicity Condition” which is “a universal condition requiring iconic representation of the temporal relations between two subevents,” and requires that “the constituents involved must be verbal” in order to be constrained by temporal iconicity.

3 L2 Acquisition Difficulties of Word Order Constrained Compound Verbs: Findings from Learner Copora of L1 English and L1 Japanese

We will investigate the difference between Japanese L1 learners and English L1 learners in the acquisition of Chinese resultative complements in “Tokyo University of Foreign Studies and National Taiwan Normal University (henceforth, The TUFS_NTNU) Learners’ Error Corpora of Chinese” https://corpus.icjs.jp/corpus_ch/index.php.

The Chinese learners’ corpus consists of 369 essays by L1 Japanese in Chinese majors at Tokyo University of Foreign Studies. The essays are corrected, error tagged, and include learner information. Data from a wide range of learners is included, from 2nd year students at intermediate proficiency levels to advanced level (4th and 5th year) students at Cefr B2 level with one-year study abroad experience. The data includes homework tasks as well as translations of the Chinese version of the “Memories of Study Abroad in Shanghai” task as the appendix of the Chap. 4 shows. Both types of task allowed the use of a dictionary (Table 2).

In addition, for English L1 learners, while the data cannot be made public because learners’ consent has not been obtained, data obtained from National Taiwan Normal University has also been corrected, error tagged and used for research purposes. The

Table 2 TUFS learner corpus of Chinese collected from May 2013–August 2014

Academic year	Level Chinese major students	Number of essays	Approximate number of words	Number of students
2013	Advanced (4th year)	95	45,500	35
	Intermediate (2nd/3rd year)	132	51,200	58
2014	Advanced (4th year)	21	12,500	23
	Intermediate (2nd/3rd year)	34	25,100	69
Total		282	134,300	185

Table 3 The TOCFL English-native learners' corpus of Chinese

TOCFL (CEFR)	Number of compositions	Number of Chinese characters	Number of students
基礎 (A2)	223	119,971	223
進階(B1)	344	31,852	344

data consists of essays written by native speakers of English as part of the Test of Chinese as a Foreign Language (TOCFL), a Chinese proficiency test developed by National Taiwan Normal University (Table 3).

English has SVO word order system and a rich system of phrasal verb (e.g. get up, work up, pull in, wash out, make out). On the other hand, Japanese has SOV word order and a rich system of compound verbs. However, "Verb + Complement" resultative compound verbs (e.g.V-到 dào/ V-完wán/ V-好 hǎo) in Chinese are the most difficult forms even for B2 level advanced Japanese L1 leaners.

(3) V 到 (underuse)

ID: TUFS_CH_109 4th year/Chinese major/Length of Chinese study: 50 months

现在听妈妈说 → 起 (add, 补语 趋向补语) → 这件事 (add, 宾语), 我 → 才 (add, 副词 语气副词) 知道了 → (delete, 助词 时态助词) 那时候她被我的成长所 → (delete, 助词 其他助词) 感动 → 了 (add, 助词 语气助词) 。我每天吃妈妈做的饭, 碗子 → 碗 (replace, 名词) 也不常洗, 但是那一天妈妈知道了我也成为 → **到 (replace, 表现 动词)** → 了 (add, 助词 时态助词) 会做饭的年龄了, 也体验 → 到 (add, 补语 结果补语) 洗碗的辛苦了。那天做的饭 → 虽然 (add, 连词复句 转折复句) 比在商城卖的包包 → 还 (add, 副词 程度副词) 便宜的 → 得 (replace, 助词 结构助词 补语 程度补语) 多, 但是给妈妈留下了 → 非常 (add, 副词 程度

副词) 深刻的印象, 也帮 → 帮助(replace, 动词) → 了 (add, 助词 时态助词) 我成长。

Hearing my mother talk about this now, I learned for the first time that she was impressed with my growth at that time. I ate the food my mother cooked for me every day and did not wash the dishes often, but that day my mother learned that I was old enough to cook and to understand the difficulty of washing the dishes. The food I made that day was much cheaper than the bags sold in department stores, but it impressed my mother and also made me grow up.

from TUFS_NTNU Learner Error Corpus https://corpus.icjs.jp/corpus_ch/index.php.

(4) V 完 (underuse)

ID: TUFS_CH_104 4th year/Chinese major/Length of Chinese study: 48 months

三年前, 我刚过 20 岁的生日的时候, 对 → 给 (replace, 介词) 父母送了礼物。从 → 我们 (add, 代词 人称代词) 小时候 → 开始 (add, 助词 其他助词 短语 介词短语), 我 → (delete, 代词 人称代词) 父母 → 就 (add, 副词 时间副词) 辛辛苦苦地照顾我们兄弟, 所以我觉得我 → 应该 (add, 动词 能愿动词) 给 → 向 (replace, 介词) 他们表示感谢的意思 → 感激之情 (replace, 表现 短语 “的”字短语), → 于是 (add, 连词 复句 承接复句) 决定送 → 给他们 (add, 短语 介词短语) 礼物。我给爸爸送了一瓶冲绳的传统酒 “泡盛” → , (add, 标点符号) 类似于烧酒。我知道爸爸喜欢烧 酒, 以 前 → 也 (add, 副 词 关 联 副 词) 收 集 的 → 过 (replace, 助 词 时 制助 词) 。我走 → 跑 (replace, 表现 动词) 了很多 → 家 (add, 量词)酒店, 终于 → 到 (add, 补语 结果补语) 买了一瓶口味清淡好喝的泡盛。几天后, 我打 → **完 (add, 补语, 结果补语)** 工后 → (delete, 名词 方位词) 回家时 → 后 (replace, 名词 时间名词), 爸爸对我说 → : (add, 标点符号) “你给我的酒 → 很 (add, 副词 程度副词) 好喝的 → (delete, 助词 语气助词)”。我很高兴 的 → (delete, 助 词 语气助 词) 。○ → 我 (replace, 主 语) 给 妈 妈 → 送 (add, 动词) → 了 (add, 助词 时态助词) 咖啡豆。 → , (replace, 标点符号) 因为妈妈非常喜欢喝咖啡。但是那时候, 妈妈在 → (delete, 副词 时间副词) 开始工作 → 了 (add, 助词 时态助词), → 所以(add, 连词 复句 因果复句) 没有时间买咖啡豆煮咖啡。

Three years ago, on my 20th birthday, I gave my parents some presents. Since our parents had struggled to raise me and my siblings since we were very young, we decided to give them presents to show our appreciation. I gave my father awamori, a traditional Okinawan liquor similar to shochu. My father loved shochu and had been collecting it for a while. I went to many stores and finally managed to buy a bottle of clear, easy-drinking awamori. A few days later, when I came home from my part-time job, my father said to me, "The sake you gave me the other day was delicious." I was very happy. My mother loves coffee, so I gave her some coffee beans. However, my mother started working, so she had no time to make coffee.

from TUFS_NTNU Learner Error Corpus https://corpus.icjs.jp/corpus_ch/index.php.

(5) V 好 (underuse)

ID: TUFS_CH_094 3th year/Chinese major/Length of Chinese study: 25 months

第一, 通过打工, 大学生可以了解社会和工作。在日本, → 有 (add, 存现动词) 很多大学生不升入研究生院 → 读研究所 (replace, 表现)。毕业以后, 就参加工作。大学生最好准备 → 好 **(add, 补语, 结果补语)** 成为一名 (add, 量词短语) 社会成员。还有 → 想 (add, 能愿动词) 打工的学生可以在各种各样的工作单位 → 打工 (add, 动词)。比如说, 当补习班的教师, 做家教, 服务员, 而且 → 或者 (replace, 连词) 做翻译 → 等等 (add, 列举助词)。他们也许 → 能 (add, 能愿动词) 从中作出决定 → 找到 (replace, 表现 动词 结果补语) 将来 → 想 (add, 能愿动词) → 从事 (add, 动词) 的工作, 又可以练习解决种种 → 的 (add, "的" 字短语结构助词) 课题 → 问题 (replace, 表现)。

First, university students can learn about society and work through part-time jobs. Many Japanese university students do not go on to graduate school. They start working right after university. Therefore, university students should be ready to enter society. In addition, students can work part-time in many different types of jobs. For example, they can be cram school teachers, tutors, store clerks, translators, etc. They may be able to find a job they want to do in the future, and can also learn how to solve all kinds of problems.

from TUFS_NTNU Learner Error Corpus https://corpus.icjs.p/corpus_ch/index.php.

We will compare the difference between Japanese L1 learners in "TUFS Learner Corpus of Chinese" and English L1 learners in "The TOCFL English-Native Learners' Corpus of Chinese" in the acquisition of Chinese resultative complements in order to investigate the influence of SVO/SOV word order typology.

Table 4 shows "Raw frequencies of resultative complements" produced by native speakers of Chinese, Japanese, and English. Table 5 shows Adjusted frequencies (per 10,000 word) of resultative complements produced by Japanese and English native speakers and Table 6 shows X^2 significance testing. Results reveal that English native speakers use "V + 到 dào", "V + 完 wán" and "V + 好 hǎo" significantly more frequently than Japanese native speakers.

Results reveal that English native speakers use "V + 到dào", "V + 完 wán" and "V + 好hǎo" significantly more frequently than Japanese native speakers.

We suggest one of the possible factors for this phenomenon in second language acquisition is the word order typology in a word structure. English has SVO word order and "Verb + Resultatives" structure both in syntax and lexical structure as phrasal verb, therefore, both English and Chinese have the same word order in lexicon, it would be easier for English L1 learners of Chinese to acquire resultative complements in Chinese.

On the other hand, Japanese has SOV word order, there is no SVOC construction, therefore, this word order typology might affect the acquisition of "Verb + Resultatives" structure.

Table 4 Raw frequencies of resultative complements produced by Japanese and English native speakers

Total words	109,143	80,736	189,879
	Japanese native speakers	English native speakers	Total
V + 到 dào	430	423	853
V + 成 chéng	64	43	107
V + 完 wán	20	67	87
V + 好 hǎo	17	51	68
V + 掉 diào	6	8	14
V + 错 cuò	12	6	18
V + 开 kāi	2	1	3
V + 住 zhù	15	7	22
V + 坏 huài	4	2	6
V + 满 mǎn	5	6	11
Total	575	614	1,189

Table 5 Adjusted frequencies (per 10,000 word) of resultative complements produced by Japanese and English native speakers

Total words	109,143	80, 736
	Japanese native speakers	English native speakers
V + 到 dào	393.98	523.93
V + 成 chéng	58.64	53.26
V + 完 wán	18.32	82.99
V + 好 hǎo	15.58	63.17
V + 掉 diào	5.50	9.91
V + 错 cuò	10.99	7.43
V + 开 kāi	1.83	1.24
V + 住 zhù	13.74	8.67
V + 坏 huài	3.66	2.48
V + 满 mǎn	4.58	7.43

Table 6 X^2 significance testing

	$\chi 2$	P	df	Significance	Corpus with higher frequency
V + 到 dào	17.23	0.0000	1	Significant at 0.1% ($\chi 2 = 17.24$, $p = 0.000$)	English
V + 成 chéng	0.15	0.6962	1	No significant difference ($\chi 2 = 0.15$, $p = 0.696$)	n/a
V + 完 wán	40.97	0.0000	1	Significant at 0.1% ($\chi 2 = 40.97$, $p = 0.000$)	English
V + 好 hǎo	28.05	0.0000	1	Significant at 0.1% ($\chi 2 = 28.05$, $p = 0.000$)	English
V + 掉 diào	0.70	0.4029	1	No significant difference ($\chi 2 = 0.70$, $p = 0.403$)	n/a
V + 错 cuò	0.30	0.5823	1	No significant difference ($\chi 2 = 0.30$, $p = 0.582$)	n/a
V + 开 kāi	0.00	1.0000	1	No significant difference ($\chi 2 = 0.00$, $p = 1.000$)	n/a
V + 住 zhù	0.64	0.4239	1	No significant difference ($\chi 2 = 0.64$, $p = 0.424$)	n/a
V + 坏 huài	0.00	0.9663	1	No significant difference ($\chi 2 = 0.00$, $p = 0.966$)	n/a
V + 满 mǎn	0.25	0.6158	1	No significant difference ($\chi 2 = 0.25$, $p = 0.616$)	n/a

4 Effect of Word Order Differences on L2 Processing: Findings from Neurological Studies

Behavioral evidence have offered an insight into how the syntactic typological differences between languages influence the usage of word order and compound verbs of L2. These phenomena are in line with psycholinguistics and L2 acquisition studies, which have discussed a phenomenon called language transfer, the impact of cross-linguistic similarities and differences between L1 and L2 (Odlin, 1989; Gass and Selinker, 2001; Koda, 2005). In neurolinguistics, numerous previous neurobiological studies on bilinguals and L2 learners have reported that age of acquisition and level of L2 proficiency are the most influential factors in determining the brain mechanisms of L2. Besides these two factors, the cross-linguistic differences between L1 and L2 are also important factors that facilitate or hinder L2 grammar acquisition in the brain (Jeong et al., 2007a, 2007b; Kotz, 2009; Suh et al., 2007). In particular, these studies that employ neurocognitive measures of event-related potentials (ERPs) and functional magnetic resonance imaging (fMRI) have documented that word order similarities and differences between L1 and L2 can affect the cortical organization of L2 syntactic and sentence processing. In fitting with the aim of this book, which is to provide an overall understanding of L2 acquisition and processing-related to Mandarin Chinese, this section mainly aims at presenting neurolinguistic findings of

L2 word order processing by native Chinese speakers learning various L2s and by nonnative speakers of Chinese.

Over the past two decades, ERPs studies have advanced the understanding of syntactic and word order processing in the human brain. ERPs, which are part of electroencephalography (EEG), manifest averaged electrocortical signals that respond to the onset of the presentation of a particular stimulus (e.g., non-words) or a cognitive operation (e.g., recognition of syntactic violations during the processing of a sentence or phrase (Luck, 2014). Therefore, ERP measures provide relatively accurate temporal information on the neural activation changes for language processing (Hell & Tokowicz, 2010). Several ERP components have been considered to manifest specific syntactic operations. These include an early left anterior negativity (ELAN), left anterior negativity (LAN), and late centroparietal positivity (P600) (Friederici, 2011). According to Friederici (2011), the ELAN component was found between 120 and 200 ms in response to syntactic-structure errors, reflecting word category access and initial syntactic structure building processes. The LAN component, in the time window between 300 and 500 ms, is suggested to be related to the assignment of syntactic argument relations during sentence comprehension. Finally, the P600 is often observed to reflect late- stage syntactic integration and reanalysis (also see Caffarra et al., 2015 for a review).

Converging evidence shows that syntactic typological differences, namely word order and morphosyntactic features, between L2 learners' native language and their L2 play an essential role in affecting the brain response during sentence processing. Such a syntactic difference effect has been observed for native Chinese speakers who study Indo-European languages as their L2, which have long morphosyntactic and syntactic distance with the learners' native Chinese language. For example, in an earlier ERPs study, Chen et al. (2007) examined Chinese English learners' ERP response to syntactic agreement processing in English. Unlike Indo-European languages, which have rich morphosyntactic systems to mark syntactic relations between sentence elements, the Chinese language mainly relies on word order to convey these relations (Li, 1989; Li et al., 1993). As learners may rely on their L1 processing strategies for handling the L2 sentence structures due to these syntactic differences, Chen et al. assumed that the ERP response pattern for processing English subject–verb agreement would differ between native speakers of English and Chinese learners of English. In their experiment, English native speakers and Chinese learners of English were asked to read sentences with agreement violation (e.g., ∗ *The price of the cars were too high*). Their results showed that English native speakers elicited the typical biphasic LAN and P600 effects in reading agreement violation sentences.

In contrast, Chinese L2 learners did not show the biphasic ERP profiles as native speakers did when they detected the subject–verb agreement violations indicated by behavioral measures. Chen et al. also found out that the L2 learners showed ERP effects of N400 following P600-eliciting for processing the grammatical sentences, which included an incongruent local noun in number with the verb (e.g., The price of the cars was too high).

Chen et al. thus suggested that the syntactic distance and morphosyntactic differences between Chinese and English affect L2 syntactic processing and acquisition.

The following studies have extended this line of work to clarify how syntactic distance between the learners' L1 Chinese and L2 affects the parts of the brain responsible for processing L2 syntactic structures (Dowens et al., 2010, 2011; Chang & Wang, 2016).

In two ERPs studies, Dowens and colleagues compared the brain activation in processing Spanish sentences of number and gender agreement by English and Chinese proficient late L2 learners of Spanish and Spanish native speakers. Participants were instructed to read a sentence with violations of number and grammatical gender agreement (adjective–noun agreement and article–noun agreement. The results showed that the P600 effects were found among the English and Chinese L2 learners of Spanish for processing gender and the number violations.

However, a stronger P600 effect was elicited when English L2 learners of Spanish processed sentences with number agreement violations than those with gender agreement violations. In comparison, the P600 effects did not show significant differences between the number and gender agreement violations in the Chinese L2 learners of Spanish. These findings can be explained by syntactic distances between the learners' L1 and the target L2. There is a number agreement feature in English, while the Chinese language does not have any morphosyntactic inflections to indicate syntactic agreement relations.

The functional magnetic resonance image (fMRI) method, which has high spatial resolution by monitoring the blood oxygenation changes to locate the brain activations involved in various language tasks, has the advantage of identifying the anatomical regions and networks relevant to syntactic processing. Numerous previous fMRI studies on bilinguals and L2 learners have reported that the age of acquisition and the level of L2 proficiency are the most influential factors in determining brain areas for L2 processing. Besides these two factors, the linguistic distance between L1 and L2 is also an essential factor that facilitates or hinders L2 grammar acquisition in the brain (Jeong et al., 2007a, 2007b; Kotz, 2009; Suh et al., 2007).

For example, Jeong et al. (2007a) demonstrated that different word order between L1 and L2 influences brain activation during L2 processing. Their participants were native Korean speakers who had learned two typologically different languages (English and Japanese) as their L2s. Using an fMRI method, the participants' brain activation was measured when comprehending auditory sentences consisting of four phrases in Korean and two L2s (English and Japanese). Japanese and Korean have the same word order (S-O-V) and use case particles to indicate thematic roles in a sentence (Shibatani, 1990). In contrast, English has S-V-O word order and relies on word order to signal the subject–object relationship. It was found that comprehending English sentences produced greater activation in the left inferior frontal gyrus (IFG) than comprehending Korean and Japanese sentences (Jeong et al., 2007a, 2007b).

The left inferior frontal gyrus (IFG) is a well-known area for processing syntactic and word order differences (see Friederici, 2011 for a review). Thus, differential

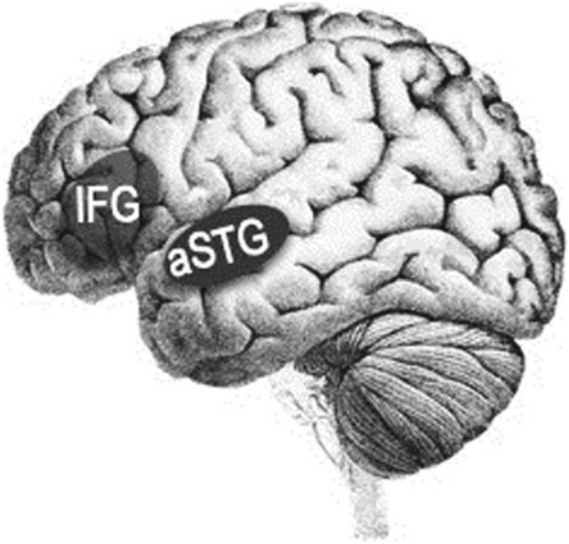


Fig. 1 Neuroimaging findings for effect of word order and morphosyntactic differences on L2 processing. The Chinese group produced greater activation in the anterior part of the superior temporal gyrus (aSTG) involved in processing morphosyntactic features of Japanese (e.g., case particle) than the Korean group (Jeong et al., 2007b). The Korean group produced greater activation in the left inferior frontal gyrus (IFG) for comprehending English sentences than comprehending Korean and Japanese sentences (Jeong et al. 2007a)

activation between English and Japanese reflects cognitive demand for processing distant word order in two languages.

Jeong et al. (2007b) tested the further hypothesis of whether different brain regions are involved in L2 processing depending on the specific linguistic features that differ between L1 and L2. They compared two native-speaker groups (Korean and Chinese) who had learned the same L2s (English and Japanese) controlling the level of L2 proficiency, age of L2 acquisition, and amount of L2 exposure between the two groups. Korean and Chinese groups performed auditory sentence comprehension tasks in their L1 and two L2s.

Brain activation was compared (a) during English processing relative to L1processing and (b) during Japanese processing relative to L1 processing. Syntactically, Chinese and English have SVO basic word order, and this word order is generally used to indicate grammatical relationships between sentence constituents. These characteristics contrast with those of Korean and Japanese, which have the SOV word order.

As a result, different brain activation patterns between the Korean and Chinese groups were observed to match these cross-linguistic characteristics. During Japanese sentence processing, the Chinese group produced greater activation in the anterior part of the superior temporal gyrus involved in processing morphosyntactic features (e.g., case particles) than the Korean group (Fig. 1). During English sentence

processing, the Korean group showed greater engagement of the left inferior frontal gyrus, which is involved in word order processing, than the Chinese group.

This study provides evidence that brain areas are selectively engaged in processing L2 sentences, depending on the different linguistic features (word order and morphosyntactic features) between L1 and L2.

5 Conclusion

This chapter discussed how learner's native language typology affects the second language acquisition of word order and morphology from the viewpoints of learner's 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. Our Chinese learner corpora by L1 English and L1 Japanese display the difference in acquiring "Verb + Complement" resultative compound verbs. English L1 learners seem to be easy to acquire the resultative compound verbs, it might be because English has same word order SVO and similar SOVC resultative construction. On the other hand, for Japanese L1 learners of Chinese, even advanced learners find quite difficult to acquire the resultative compound verbs, since Japanese word order is SOV, there is no SOVC resultative construction.

This word order typology might affect the acquisition of lexical aspect by Japanese and Korean L1 learners, since both Tai's "principle of temporal sequence" (1985: 50) and Li's "Temporal Iconicity Condition"(1993: 480, 502) do not apply to SOV type languages.

Second, morphological typology, Isolating language (Chinese) or Agglutinative language (Japanese and Korean) affects the acquisition of stems and affixes. For Chinese learner's of Japanese, it is difficult to acquire Japanese " transitive/intransitive" pairs (refer to Chap. 11 for in this book).

(6) Break

a. kowa-s-u"弄壞"
transitive break
b. break"壞"
kowa-rer-u
intransitive break

We assume that Japanese/Korean L1 learners and English L1 learners would have different brain activation patterns when learning Chinese because both Japanese/Korean have different typology from Chinese in both word order and morphological typology. Therefore, we need to develop a pedagogy for Japanese/Korean L1 learners based on the typological differences between learner's L1 and Chinese.

References

Aoki, H. (2013). Historical change in Japanese compound verbs. In *New explorations into the mysteries of compound verbs* (pp. 215–241). Tokyo: Hitsuji Shobo.

Caffarra, S., Molinaro, N., Davidson, D., & Carreiras, M. (2015). Second language syntactic processing revealed through event-related potentials: An empirical review. *Neuroscience & Biobehavioral Reviews*, *51*, 31–47. https://doi.org/10.1016/j.neubiorev.2015.01.010.

Chang, X., & Wang, P. (2016). Influence of second language proficiency and syntactic structure similarities on the sensitivity and processing of English passive sentence in late Chinese-English bilinguists: An ERP study. https://doi.org/10.1007/s10936-014-9319-1.

Chen, L., Shu, H., Liu, Y., Zhao, J., & Li, P. (2007). ERP signatures of subject–verb agreement in L2 learning. *Bilingualism: Language and Cognition, 10*(2), 161–174.

Dowens, M. G., Vergara, M., Barber, H. A., & Carreiras, M. (2010). Morphosyntactic processing in late second-language learners. *Journal of Cognitive Neuroscience, 22*(8): 1870–1887.

Dowens, M., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). *Gender and number processing in Chinese learners of Spanish—Evidence from event related potentials.* https://doi.org/10.1016/j.neuropsychologia.2011.02.034.

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews, 91*(4), 1357-1392. https://doi.org/10.1152/physrev.00006.2011.

Gass, S. M., & Selinker, L. (2001). *Second language acquisition, an introductory course* (2nd ed.). Lawrence Erlbaum Associates.

Jeong, H., Sugiura, M., Sassa, Y., Haji, T., Usui, N., Taira, M., … Kawashima, R. (2007a). Effect of syntactic similarity on cortical activation during second language processing: A comparison of English and Japanese among native Korean trilinguals. *Human Brain Mapping, 28*(3), 194–204. https://doi.org/10.1002/hbm.20269.

Jeong, H., Sugiura, M., Sassa, Y., Yokoyama, S., Horie, K., Sato, S., … Kawashima, R. (2007b). Cross-linguistic influence on brain activation during second language processing: An fMRI study. *Bilingualism: Language and Cognition, 10*(2), 175–187. https://doi.org/10.1017/s1366728907002921.

Kageyama, T. (1996). *Doshi Imiron: Gengo to Ninchi no Setten [Verb semantics: The interface of language and cognition]*. Tokyo: Kurosio Publishers.

Kageyama, T. (2009). *Isolate*: Japanese. In R. Lieber, & P. Stekauer (Eds.), *The Oxford handbook of compounding* (pp. 512–526). Oxford: Oxford University Press.

Kageyama, T. (2010). Variation between endocentric and exocentric word structures. *Lingua, 120*, 2405–2423.

Kageyama, T. (2016a). Verb-compounding and verb-incorporation. In T. Kageyama, & H. Kishimoto (Eds.), *Handbook of Japanese Lexicon and word formation* (pp. 273–310). Berlin and Boston, MA: De Gruyter Mouton.

Kageyama, T. (2016b). Agents in anticausative and decausative compound verbs. In T. Kageyama, & W. M. Jacobsen (Eds.), *Transitivity and valency alternations: Studies on Japanese and beyond* (pp. 89–124). Berlin and Boston, MA: De Gruyter Mouton.

Kageyama, T. (2018). Compound and complex predicates in Japanese. In *The Oxford research encyclopedia of linguistics*. Oxford University Press. Retrieved August 4, 2020, from https://oxfordre.com/linguistics.

Kageyama, T., & Kanzaki, K. (2014). *Compound Verb Lexicon (online database)*. Tokyo: National Institute for Japanese Language and Linguistics. Retrieved August 23, 2020, from https://db4.ninjal.ac.jp/vvlexicon/db/.

Kageyama, T., & Shen, L. (2018). Resultative constructions in Japanese from a typological perspective. In P. Pardeshi, & T. Kageyama (Eds.), *Handbook of Japanese contrastive linguistics* (193–226). Berlin and Boston, MA: De Gruyter Mouton.

Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain & Language, 109*(2–3), 68–74. https://doi.org/10.1016/j.bandl.2008.06.002.

Koda, K. (2005). Theoretical underpinnings. In *Insights into second language reading: A cross-linguistic approach* (Cambridge Applied Linguistics, pp. 13–26). Cambridge University Press.

Li, P. (1989). What cues can Chinese speakers use in sentence processing? *Paper presented at the workshop on crosslinguistic study of sentence processing*. Dept. of Psychology, Carnegie Mellon University, Pittsburgh.

Li, Y. (1993). Structural head and aspectuality. *Language, 69*, 480–504.

Li, P., Bates, E., & Macwhinney, B. (1993). Processing a language without inflections: A reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language, 32*(2), 169–192.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Mochizuki, K. (2004). *Causative and inchoative alternation: Comparative studies on verbs in Chinese*. Ph.D. thesis. National Tsing Hua University.

Mochizuki, K. (2007). Patient-orientedness in resultative compound verbs in Chinese. In *Corpus-based perspectives in linguistics* (pp. 287-300). John Benjamins.

Mochizuki, K., & Shen, Y. (2011). Word formation in compound verbs in Japanese and Chinese. In *Comparative linguistics in Chinese and Japanese* (Vol. 2, pp. 46–72). Association for Comparative Linguistics in Chinese and Japanese. Peking University Press (In Chinese and Japanese: 「日本語と中国語の複合動詞の語形成」『汉日语言对比研究论丛第二辑』2 卷, 46–72, 汉日对比语言学研究(协作)会, 北京大学出版社.)

Mochizuki, K., & Shen, Y. (2012). Inheritance of argument structure and compounding constraints of resultative compound verbs in Chinese and Japanese. In L. E. Clemens, & C.-M. Louis Liu (Eds.), *Proceedings of the 22rd North American Conference on Chinese Linguistics (NACCL-22) and the 18th International Conference on Chinese Linguistics (IACL-18)* (Vol. 2, pp. 341–355). Cambridge, MA: Harvard University.

Nishiyama, K. (1998). V-V compounds as serialization. *Journal of East Asian Linguistics, 7*(3), 175–217.

Odlin, T. (1989). *Language transfer: Cross-linguistic Influence in language learning* (p. 224). New York: Cambridge University Press, Inc.

Saur, D., Baumgaertner, A., Moehring, A., Büchel, C., Bonnesen, M., Rose, M., Musso, M., & Meisel, J. M. (2009). Word order processing in the bilingual brain. *Neuropsychologia, 47*(1), 158–168. https://doi.org/10.1016/j.neuropsychologia.2008.08.007.

Suh, S., Yoon, H. W., Lee, S., Chung, J. Y., Cho, Z. H., & Park, H. (2007). Effects of syntactic complexity in L1 and L2; An fMRI study of Korean-English bilinguals. *Brain Research, 1136*(1), 178–189. https://doi.org/10.1016/j.brainres.2006.12.043

Shibatani, M. (1990). *The languages of Japan*. New York: Cambridge University Press Inc.

Tai, J. H.-Y. (1985). Temporal sequence and Chinese word order. In J. Haiman (Ed.), *Iconicity in Syntax: Proceedings of a Symposium on Iconicity in Syntax* (pp. 49–72), Stanford, June 24–26. John Benjamins.

Tai, J. H.-Y. (2004). Cognitive relativism: Resultative construction in Chinese. *Language and Linguistics,* 4(2), 301–316.

Tang, T.-C., & Hsu, S. (2015a). *An introductory study of contrastive linguistics* (Vol. 1). Jhih Liang Publisher: Taipei. (湯廷池, 許淑慎 2015『對比分析研究入門』上冊 台北: 致良出版社.)

Tang, T.-C., & Hsu, S. (2015b). *An introductory study of contrastive linguistics* (Vol. 2). Jhih Liang Publisher: Taipei. (湯廷池, 許淑慎 2015『對比分析研究入門』下冊 台北: 致良出版社.)

van Hell, J. G., & Tokowicz, N. (2010). *Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels*. https://doi.org/10.1177/0267658309337637.