languages

The Acquisition of Chinese as a First and Second Language

Edited by
Xiaohong Wen

Printed Edition of the Special Issue Published in *Languages*

www.mdpi.com/journal/languages

MDPI

The Acquisition of Chinese as a First and Second Language

The Acquisition of Chinese as a First and Second Language

Editor

Xiaohong Wen

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin

Editor
Xiaohong Wen
University of Houston
USA

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Languages* (ISSN 2226-471X) (available at: https://www.mdpi.com/journal/languages/special_issues/Acquisition_Chinese).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03943-270-7 (Hbk)
ISBN 978-3-03943-271-4 (PDF)

Contents

About the Editor

Xiaohong Wen is a Professor of Applied Linguistics and Chinese Language Acquisition at the University of Houston. Dr. Wen has conducted a series of empirical studies from both quantitative and qualitative perspectives. Her publications address second language (L2) learning motivation, acquisition of Chinese as an additional language, interlanguage pragmatics, heritage language, L2 learner factors, and research-based instruction. Her L2 Chinese motivation instruments have been widely adopted for research in various languages. Her empirical studies have been reprinted and translated into Korean and Chinese. She is the Principal Investigator for several major federal and internal grants. Her recent publications include four books: *Studies on Learning and Teaching Chinese as a Second Language* (2019); *Teaching Chinese as a Second Language: Curriculum Design and Instruction* (2015); *Studies of Chinese Language Acquisition by English Speakers* (2012); *Chinese as a Second Language Acquisition and Instruction* (2008). Her research publications also include more than forty articles, mostly in peer-reviewed journals.

languages

MDPI

Editorial

The Acquisition of Chinese as a First and Second Language

Xiaohong Wen

Chinese Studies Program, University of Houston, Houston, TX 77004, USA; xswen@Central.UH.EDU

Received: 9 June 2020; Accepted: 19 August 2020; Published: 3 September 2020

1. Background

The last two decades have witnessed a surge of interest in learning Chinese as a second language (L2 Chinese). The interest is reflected in multiple dimensions including education, Chinese language course enrollment, student body composites, and learning motivation. Chinese language programs and courses have mushroomed, especially in response to the College Board's decision that Advance Placement (AP) Chinese and Culture Examination was to be implemented in 2006 in the USA. Chinese language education was rapidly institutionalized at many American schools; e.g., from 2004 to 2008, there was a 195% increase in enrollment in Chinese language courses in K-12 U.S. public schools (ACTFL 2011).

In addition to an overall increase of interest in Chinese language, the diversity of Chinese language learners' ethnic backgrounds expanded. More than two decades ago, L2 learners were mostly Caucasians. In contrast, a recent large-scale survey found that Caucasians now comprise only 51% of the student body (Li et al. 2014). Chinese heritage students and those from Asian-American backgrounds make up more than 30%, and students from Latin American and African American backgrounds comprise a further 16%. One feature among heritage learners is their variety of linguistic backgrounds and experiences. Along with these significant changes in the student body, learners' goals for studying Chinese have also been evolving. Traditional targets such as going to graduate school and becoming a sinologist have been replaced by functional and instrumental use of the language, as well as competence in Chinese culture and language (Comanaru and Noels 2009; Sung 2013; Xie 2014; Wen 2011; Wen and Piao 2020).

In contrast to the rapid development of Chinese education, research on acquisition of Chinese as a first and additional language has lagged behind. In reviewing the current state of the literature on Chinese language motivation, Wen (2018) located only 16 empirical studies on the topic published up to 2017 after an exhaustive search. Only five Chinese Language Teachers Association (CLTA) monographs have been published since the CLTA was established six decades ago. Recent years have seen more development in L2 Chinese empirical research in books (Everson and Shen 2010; Han 2014; Tao 2016; Wen 2012; Wen and Jiang 2019) and journals. Articles in these books and journals present more rigorous research methodology and a broader scope of inquiries. However, research into Chinese language acquisition, particularly empirical studies, is sparse and lags behind the research development of general language acquisition.

The lack of research into Chinese language acquisition directly affects Chinese language teaching, a deficiency that hampers learning and underserves students. Instructors need to be updated on the current research findings in order to teach effectively. What determines the quality of classroom instruction is research-based knowledge of learners and learning processes. Well informed instructors are able to opt for appropriate pedagogical approaches and instructional conditions for their students.

The acute demands for understanding Chinese language acquisition call for more empirical studies on a wide range of topics to scrutinize the nature of Chinese language learning and learning processes. The authors offer many thanks to *Languages*, a major peer-reviewed journal, which has

recognized the important yet under-represented research in Chinese language acquisition. This special issue, *Acquisition of Chinese as a First and Second Language*, has collected eight empirical studies showcasing research advances in multiple domains. The studies are theoretically motivated and have adopted innovative methodological strategies to achieve a broader understanding of Chinese language acquisition.

The eight empirical studies in this volume encompass a wide range of topics arching from the L1 to L2 acquisition of syntax, semantics, phonetics, and discourse. Two Chapters by Chen, Narasimhan, Chan, W. Yang, & S. Yang, and by Chen & Wang focus on child acquisition of Chinese as a first language whereas chapters by Xu, Liao, Y. Liu, Wang & Chen, Y. Liu, and Yan examine adult acquisition of L2 Chinese. Within L2 acquisition, heritage learners converge, and very frequently, diverge from general L2 learners in the learning process. The last chapter by Yan explores heritage syntax-discourse interface properties. Additional themes examine language skill acquisition, including speaking, listening, and writing. The volume is distinctively featured with comparative studies analyzing learners (e.g., children versus adults; heritage versus general L2 or foreign language learners) across linguistic, cognitive, and research methodological domains (e.g., syntax versus discourse/pragmatics; perceptions versus productions, control versus experimental groups). Last but not least, the volume attempts to connect research to pedagogy. Several chapters built instructional interventions into their designs to scrutinize the effects of consciousness raising, metacognitive strategy training, and student-led versus teacher-led instruction. Implications of study findings bridge the gap between research and instruction in hope of helping teachers to understand their students and their learning processes.

2. Characteristics of the Volume

2.1. Theoretically Motivated with Diverse Research Goals

Studies in this issue are theoretically motivated, reflecting a wide range of research purposes and current issues. Theories adopted encompass cognitive processing strategies, explicit learning, and consciousness raising in L2 acquisition (Gass and Selinker 2008), as Xu's study illustrates. Yanmei Liu's conceptual framework is based on metacognition and metacognitive awareness strategies (Vandergrift et al. 2006) to explore listening comprehension processes. Yu Liu's investigation drew upon the framework of speech production processes (Kormos 2014) to examine lexical accessibility in relation to speaking accuracy and complexity. Adopting the theoretical account of textual organizational devices, Liao analyzed metadiscourse strategies for discourse cohesion in learners' descriptive writing. These studies, motivated by theories on cognition and second language acquisition, investigate the learning process and the development of decoding and encoding skills.

Another group of studies is conducted under the psycholinguistic framework investigating Chinese language acquisition in syntax, semantics, phonetics, and discourse. Chen et al. adopted the theoretical framework of word order and information structure to compare the speech data of Mandarin-speaking children and adults, challenging the established "old-before-new" information structure. Chen and Wang examined the mapping of one form to multiple meanings/functions, testing the continuous derivational and restricted monosemy approaches. Yan drew on the linguistic account of semantic- and syntax-pragmatic interfaces to analyze the Chinese sentence-final particle *ba*, a multifunctionalmark in discourse. Wang and Chen's study bore out the Perceptual Assimilation Model (PAM) after testing both Speech Learning Model (SLM) and PAM via their study of the relationship between perception and production of Mandarin consonants by English speaking college students in the USA.

In addition, two studies (Xu, Yanmei Liu) situated in a complex setting, the classroom, analyzed the effects of instructional intervention in the conceptual framework of the declarative/procedural model and input, interaction, and output model. Other studies (Chen et al., Chen & Wang, Wang & Chen) tested theoretical models and perspectives to support or challenge current theoretical accounts.

In summary, all the studies are theoretically driven, examining current acquisition and learning issues from cognitive, linguistic, and first/second language acquisition perspectives.

2.2. Cognitive Approach to Research on Acquisition of Chinese as an Additional Language

There are a number of studies investigating L2 Chinese acquisition of verb-suffix *-le*, a notoriously difficult but frequently used perfective marker. There are gaps between research and teachers' understanding of the learning task for *-le*. Yi Xu's study connects the research to classroom instruction. The study examines effects of consciousness raising through explicit knowledge construction within the framework of form-focused instruction. It adopted both quantitative and qualitative designs with two learner groups [experimental (E) and control (C) groups] to compare their performance in interactive role-play and written editing in a post-test. The study has revealed several findings. First, underuse of *-le* occurs more than overuse with learners at the elementary proficiency level. Second, form-focused consciousness raising via learning materials but without the instructor's interaction is effective; more importantly, the E group showed knowledge on procedure skills. They demonstrated a more accurate understanding of *-le* rule induction than the C group. Third, given the opportunity for interactive group work and explicit instructional written input, learners were able to notice, categorize, and contrast the linguistic constraints of *-le*. Fourth, learner individual differences were observed particularly in the process of rule induction, as some groups were more competent in metacognitive skills than others. Some learners seemed to be more able to conceptualize the forbidden versus obligatory environments of *-le* leading to explicit rule construction.

This study yields important results that benefit L2 Chinese language teachers. Consistent with previous studies (Duff and Li 2002; Wen 1995), Xu's study revealed that learners underproduced *-le*, particularly in the obligatory resultative verb complement (RVC) environment. Perhaps, learners may consider the resultative complement as an "indicator of completion," thus allowing *-le* to be legally omitted (Wen 1995). Furthermore, since the study was conducted in the classroom as a part of regular instruction, the materials used and the data collected, as well as the instructional intervention, can serve as an exemplary application for task-based instruction to facilitate the acquisition of explicit knowledge. It should be noted that this study is unique in its research design where participants conducted consciousness-raising activities by themselves. The process maximized the student agency role and learning autonomy. It would, however, be interesting to have another experimental group interacting with the instructor. Comparisons then can be made between the effectiveness of self-learning indirectly guided by a teacher versus the teacher's direct interaction with learners.

Diverging from grammar acquisition, Yanmei Liu's study focused on the effects of metacognitive strategies on listening performance, and two instructional methods on metacognition training. Comprehension is a primary step for language acquisition. It is a complex process, involving attention, perception, memory, information processing, problem-solving, and linguistic parsing. In this longitudinal study, instructional intervention was implemented on two experimental groups: teacher-led with teacher's interaction, and student-directed without teacher's interaction but with instructional materials provided to students. In addition to the pre- and post-intervention measures of listening comprehension, the *Metacognitive Awareness Listening Questionnaire* (MALQ) was used to examine five factors: problem-solving, mental translation, person knowledge (learners' self-efficacy beliefs), planning and evaluation, and directed attention. The study produced several findings. The student-directed group showed higher scores on planning and evaluation and on self-efficacy beliefs than the teacher-led and the control groups. Furthermore, the two experimental groups clearly demonstrated more strategy awareness than the control group. Except for the problem-solving factor, the other four factors measured by MALQ showed higher means for the two experimental groups than for the control group. Therefore, the results indicated that the metacognitive intervention had a positive effect on metacognitive awareness development, particularly the significant gains on directed attention for the teacher-led group.

There is a lack of statistically significant effects of metacognitive awareness development on listening performance gains across the three groups. It is not clear if the instructional intervention is sufficiently rigorous; the intervention only included metacognitive strategy training. The metacognitive strategies, however, reflect only a part of listening competence. It is necessary to include other strategies, particularly the cognitive skills, in order to see a fuller picture of comprehension processing. Comprehension, after all, is a process engaging both metacognitive and cognitive strategies including problem solving skills. Future research may benefit from a broader approach to scrutinize the comprehension process.

Jianling Liao's study also explores the cognitive aspect of L2 learning with a focus on descriptive writing in a study abroad setting. Writing is a highly integrative and complex skill. It not only requires linguistic accuracy (words and sentences), cognitive clarity (ideas and logic sequencing, encoding processing), but also writing strategies (textual organization, content cohesion, structure coherence). Liao's research investigated metadiscourse devices used to form text dynamics and textual organization. In addition, the study examined the interrelations among textual organizational features for the two dimensions, interactive metadiscourse (local, global, and text cohesion) and interactional metadiscourse (self-mention and engagement). In addition, the study analyzed correlations between textual organizational features and linguistic competence.

The results demonstrated three ways in which the higher proficiency group was able to write more effectively than the lower-proficiency group. First, the more advanced group was able to use textual organizational devices (cohesion and coherence at three levels: local, global, and textual) to signal their topics/subtopics more effectively and accurately than the lower proficiency group. Second, they were able to apply a higher number of distinctive markers and conjunctions to express ideas, hypothetical meaning, and smooth transitions or contrasts. Third, they were able to engage the reader more frequently and more actively. In addition, participants' linguistic competence in using more diversified lexis was positively associated with their abilities to use more accurate textual organizational devices. Their ability to produce lengthier clauses also highly correlated with their skills in marking organizational features. It should be noted that this study examined finished written products. It is important to further investigate the writing process by developing process-oriented research to track down the complete process and major factors influencing it.

Yu Liu's study explores the process of encoding with a focus on the relationship between lexical access and speaking performance. Yu Liu's study narrowed the scope of the vocabulary within specific tasks to capture a closer look at how lexical access interacts with L2 speaking performance. Lexical access, operationally defined as vocabulary size and lexical retrieval speed, was measured via a timed vocabulary test. Speaking performance, referring to speaking fluency, accuracy, and complexity, was measured within categories including speech rate, syntactic and lexical accuracy, and lexical and syntactic diversities. The speaking tasks required four communicative functions: instructive, descriptive, explanatory, and argumentative. The results demonstrated that participants' vocabulary size significantly correlates with their lexical diversity, word difficulty, and speech rate. Furthermore, lexical retrieval speed also significantly correlates with their speaking accuracy and lexical complexity. The more words that learners know and the faster they can retrieve, the more diverse and advanced words they use in their speaking tasks. The results revealed no significant correlation between lexical retrieval speed and syntactic accuracy and complexity. Further research is needed on the issue, particularly with a larger sample size.

2.3. Linguistic Approach to Acquisition of Chinese as a First and an Additional Language

Linguistic sequential order reflects the prominence of the information deeply rooted in our psycholinguistic concept. A language-general bias stems from conceptual prominence, e.g., mentioning old information before new information. Adults generally produce the old-before-new word order in communication, with a preference for accessing the easily retrievable information first (Chen and Narasimhan 2018). Findings on child language have not reached a consensus. Mandarin

Chinese is distinctively featured with a topic-comment word order where topic represents the old and comment represents the new information in discourse (Li and Thompson 1981). If children's ordering preference is influenced by the language-specific discourse properties of the target language, Mandarin-speaking children may produce "old-before-new", similar to the adult language. If they produce the "new-before-old" word order, it would provide the crosslinguistic evidence for cognitive salience of a new information preference in child language.

Jidong Chen, Bhuvana Narasimhan, Angel Chan, Wenchun Yang, and Shu Yang investigated Mandarin-speaking children's L1 acquisition of information structure and word order preference. They compared two groups, child and adult, to examine word order preference when given two referents in a linear sequence. The order of the "old" and "new" information produced by the two groups revealed differences in word order preference and similarities in using indefinite classifier NPs. First, the adult group differed significantly from the child group in preferring the "old-before-new" word order, suggesting their distinctive conceptual prominence: the easily retrievable information first from adults and the highlighting of novel information from children. Second, children and adults share similarities at the lexical and syntactic level. Both groups predominantly produced bare NPs with no difference in distinguishing the old and the new referents; neither used classifiers to distinguish the old and the new referents. However, if only one classifier phrase was used in the task, both groups tended to use it to refer to the new information, a choice suggesting early sensitivity to language-specific syntactic devices and children's use of these devices to mark the information structure. Mandarin-speaking children's "new-before-old" preference adds crosslinguistic evidence to corroborate language independence in terms of word order preference related to information structure.

One fundamental research topic on language acquisition is form-meaning mapping. Mapping becomes more complex when multiple meanings of a word are represented by an identical form. Jidong Chen and Xincun Wang's study focused on the semantic development of the polysemous word 打*dǎ* (to hit), one of the earliest verbs in Mandarin child speech. They examined the longitudinal corpus data at the age of 1;05–3;10, and another corpus from a child and his caregiver to examine the influence of an adult's input on child production. The study showed several findings. First, children acquired the core meaning of 打 *dǎ*, a physical action involving hand contact, at an early stage. Second, multiple senses of the verb 打 *dǎ*, e.g., *hit open (V.+ resultative complement)* and *play (games)*, emerged in child's production at the same time, with the majority of usages of the verb 打 *dǎ* polysemous. The earliest emergent senses involved a limited set of specific hand contact actions, suggesting that children predominantly produced the verb for multiple senses closely connected to the prototypical meaning. Third, it was at a later stage when the metaphoric meanings with hand action emerged. The study, with the Mandarin-speaking children's data, confirmed that the concrete concepts are produced earlier than less concrete or metaphorical meanings. Fourth, the syntactic and semantic contexts are important because they are inherent to the meaning of the specific senses (e.g., an open event typically involves an animate agent and an inanimate patient). Fifth, the comparison between the individual child and his caregiver's speech showed that the child prevalently produced the prototypical senses of the verb 打*dǎ*, whereas the adult's usage was more evenly distributed across wider syntactic and semantic contexts. Nevertheless, the child data largely approximated that of the adult, with 打*dǎ* in a verb compound being the most frequent, followed by the transitive frame. The acquisition of the Chinese verb 打*dǎ* reveals an early preference for initial one-to-one unambiguous form-meaning mapping, followed later by expansion to other senses associated with the verb and its arguments. These findings, from children's Mandarin speech production, add support for a continuous derivational and restricted monosemy approach. One remaining question is whether the acquisition of the polysemous verb 打*dǎ* represents a typical learning process. Further studies on multiple polysemous verbs/nouns are needed to check whether children initially extract a core feature of a polysemous word, but only use it in a restricted way with a small number of senses in a set of syntactic frames and semantic arguments.

In a separate study on sound perception and production, Xincun Wang and Jidong Chen examined Mandarin consonants that pose difficulties for English-speaking learners. The study

tested the Perceptual Assimilation Model (PAM) and the Speech Learning Model (SLM). Twenty-five English-speaking learners read the eight Mandarin consonants (j/tɕ/, q /tɕʰ/, x /ɕ/, zh /tʂ/, ch /tʂʰ/, sh /ʂ/, z /ts/, and c /tsʰ/) in sentences in addition to identifying the target sounds in a forced-choice task. Findings show that the Mandarin retroflex, palatal, dental fricatives and affricates posed different levels of challenge to learners for listening perception. The misperceived retroflex and palatal sounds were substituted with each other in perception, but mis-produced palatal sounds were substituted with each other, not with retroflex sounds. Specifically, perception data showed that learners had different degrees of difficulties with zh /tʂ/, q /tɕʰ/, and x /ɕ/. The production data demonstrated that learners had difficulties with the c/tsʰ/, zh /tʂ/, and q/tɕʰ/ consonants. The perceived phonetic distances between Mandarin and English consonants predicted the learners' perceptual difficulties, a prediction that lent support to the PAM. On the other hand, reorganizing to establish new Mandarin phonetic categories for the retroflex, palatal, and dental sounds is a learning process whereby learners distinguish the differences between c /tsʰ/, x /ɕ/, z /ts/, and q /tɕʰ/ in order to establish these categories. As such, the SLM also plays a role in the learning process, as the authors proposed.

Correlation analysis showed the weak relationship between perception and production for the majority of the consonants investigated, suggesting that the relationship between Mandarin consonant perception and production is not straightforward. Extended studies, examining the mechanisms that English-speaking learners engage for Mandarin consonant perception and production, would be beneficial, particularly for teachers who need to know whether accurate perception precedes or facilitates accurate production.

Yan's study investigated linguistic interfaces of grammatical, discourse, and pragmatic features. The author chose to examine the acquisition of Chinese sentence-final particles, the interrogative 吧 *ba* and the suggestive 吧 *ba*. What makes learning even more difficult is that interrogative and suggestive meanings share one identical form and sound: 吧 *ba*. Syntactically, both are at the end of a sentence. The function, however, is distinctly different, with the former as a pre-assumptive confirmation question marker and the latter as a polite marker for a suggestive request. Learners must reconfigure the meaning and function in the process. Thirty-five Chinese heritage (CH) speakers at two proficiency levels participated in the study. They can speak or at least understand oral Chinese, since one or both of their parents frequently speak Chinese to them. This influence suggests that they are exposed to authentic communication with rich discourse and pragmatics. As a result, their pragmatic and discourse competence may be more developed than nonheritage learners at the same proficiency level.

Participants completed three tasks: acceptability judgment, discourse completion, and translation. The results demonstrate that the CH learners mixed the presumptive confirmation question marker 吧 *ba* with the grammatical question marker 吗 *ma* in the discourse completion task (DCT). They overproduced the regular interrogative marker 吗 *ma* and underproduced the presumptive confirmation interrogative 吧 *ba*. However, participants, particularly at the advanced level, had little difficulty in accepting grammatical and rejecting ungrammatical sentences with the pre-assumptive confirmation interrogative 吧 *ba*. Furthermore, they correctly translated the suggestive particle 吧 *ba* of a request into an English suggestion. The findings indicate that it is easier for CH learners to acquire the syntactic features than the interface between syntactic and discourse features. Such findings are consistent with previous research (Keating et al. 2011; Polinsky and Scontras 2019; Wen and Jiang 2019) in which the heritage learners did not outperform their counter-group, nonheritage learners, in terms of interface between syntax and discourse/pragmatics. Particularly, heritage learners tended to apply economical strategies including avoidance of ambiguity, resistance to irregularity, and preference for brevity and safety.

2.4. Innovative Research Design and Methodology

Research methods and data collection are central to language acquisition research, largely due to its interdisciplinary nature intersecting with linguistics, psychology, sociology, and education. Chinese language acquisition is exemplified by various linguistic and non-linguistic means drawn

from complex communicative patterns. We need various methods, including quantitative, qualitative, and mixed-methods perspectives, to capture the complex aspects of Chinese acquisition.

Research designs in this volume demonstrate an array of novel concepts. The studies have expanded traditional methods by integrating measures to achieve in-depth analysis of learners' linguistic preference, acquisition stages, cognitive skills, metacognitive strategies, and language use in social interactions. The novelty includes methods such as evaluating learner agency roles by providing instructional intervention without the teachers' interaction. Examples include learner group work for rule induction (Xu's study) and student self-directed activities (Yanmei Liu's study). Instead of using a traditional vocabulary test, Yu Liu developed a task-specific and native-referenced vocabulary assessment for research validity. Despite the focus on child language, Chen et al. and Chen and Wang's studies included Chinese native speakers (NSs) to explore the relationship between age and the linguistic encoding of information structure, and the role of input including syntactic, semantic, and contextual cues in children's polysemous verb development. Table 1 presents a summary of methodology adopted in the studies investigating acquisition of Chinese as an additional language through a cognitive approach.

Table 1. Summary of studies adopting cognitive and L2 acquisition approaches.

Authors	Research Purposes and Focus	Sample	Task Types/Materials	Measurements
Xu, Y.	The role of consciousness-raising Experimental design w/pre- and post-tests. Instruct intervention. Treatment: 1. Role-play sheets with explicit form markings of *-le*, 2. Interactive group work for rule induction	25 elementary CSL Two groups: E and C	1. Grammaticality judgment and error correction, 2. Role-play, 3. Rule induction task among E group 4. Written editing task.	Both quantitative and qualitative analyses 1. Paired sample *t*-tests 2. Coding categories for rule induction of *-le* via cognitive processes
Liu, YM	Effects of Metacognitive Strategy training on listening proficiency Longitudinal study with quantitative quasi-experimental design Intervention treatment: Metacognitive Strategy Training	80 interm. CSL Three groups: Self-direct Teach-direct Control group	1. MALQ Metacogntv Awareness Listening 2. Intervention: MST Metacogntv Strategy Training 3. A listening test 4. A proficiency test 5. Metacognitive awareness worksheet	Quantitative analysis 1. ANOVA for inter vention training effects 2. ANOVA for group differences in listening performance 3. ANCOVA and correlations for group differences in listening proficiency
Liao, J.	Relationship between metadiscourse devices and written discourse, and linguistic performance across three proficiency levels Text analysis on content cohesion and structure coherence features	62 CSL studying abroad In three group levels: elem., interm., adv.	Descriptive essay to introduce one's home university	Quantitative analysis 1. MANOVA:Compare 3 proficiencies 2. Coding categories for metadiscourse mrkrs and linguistic devices 3. Correlation for textual organizational devices 4. Correlation between linguistic and textual organization devices
Liu, Yu	Relationship between vocabulary size, retrieval speed and speaking fluency, accuracy and complexity Correlation study	15 interm-high CSL	1. Native-referenced vocabulary test 2. Four types of speaking tasks: instructive, descriptive explanatory, and argumentative	Correlation analysis 1. Coding categories for fluency, accuracy, & complexity 2. Descriptive analysis 3. Pearson correlation 4. Multiple regression

In addition to traditional tasks and materials for data collection, this volume displays a wide spectrum of innovations, such as combining grammaticality judgment with role-play, rule induction activity via learner groupwork (Xu), with DCT plus translation tasks (Yan), as well as longitudinal corpora from both children and an adult (Chen & Wang). Apart from survey questionnaires,

the qualitative approach was also adopted, including learners' text analysis (Liao; Yu Liu), discourse analyses (Yan), and semantic analyses (Chen & Wang). The statistical analyses spanned a wide range from relationship studies (e.g., Pearson correlation coefficient, various types of regression) to comparison studies (e.g., various types of *t*-tests, ANOVA, MANOVA, ANCOVA), as well as linear mixed-effects models. Table 2 presents a summary of the methodology adopted in the studies investigating L1 and L2 acquisition by Chinese children, CSL learners, and CH learners through a linguistic approach.

Table 2. Summary of studies on L1 and L2 Chinese acquisition adopting a linguistic approach.

	Research Purposes and Focus	**Sample**	**Task Types/Stimuli**	**Measurements**
Chen, J. Narasimhan, B. Chan, A. Yang, W. Yang, S.	Information structure and word order preference cross-linguistically; between Chinese children and adults Age effects on the linguistic encoding of information structure /word order	24 children 4;6 25 NS adults	Elicited production 12 target pairs of objects	Word order comparison between adults' "old-before-new" and children's "new-before-old" 1. Logistic regression: age predicts word order 2. Chi^2 test 3. Independent *t*-test
Chen, J & Wang, X.	Form-meaning mapping for the Polysemous Verb打 The role of input: syntactic, semantic, and contextual cues	9 children (1;05–3;10) 1 caregiver	Corpus longitudinal data Corpus child's and caregiver's data	1. Semantic category coding, syntactic constructs 2. Descriptive analysis 3. Comparison: child and adult
Wang, X. & Chen, J.	Relationship between perception and production Difficult consonants for English-speaking learners	25 beginning CSL	1. Identification task: ten consonants 2. Read eight consonants in sentences	1. Descriptive analysis 2. Separate ANOVA for perception & production 3. Pearson coefficients for perception and production
Yan, S.	Interface of syntactic, pragmatic, and discourse features	CH learners 16 interm. 19 adv. 18 NSs	1. Acceptability judgment 2. Discourse completion 3. Translation	1. Descriptive analysis 2. Linear mixed-effects model 3. Generalized linear mixed-effects models

In summary, this volume presents a variety of research designs and methodological strategies from both quantitative and qualitative perspectives. Cross-sectional and longitudinal studies examine the development of Chinese as a first and additional language; furthermore, they represent the extent of Chinese language acquisition research in conjunction with general language acquisition theories and research.

3. Future Research Direction

The studies in this volume, by incorporating an array of variables and operationalizing them, investigate the effects of cross-linguistic similarities, processing and production strategies, and linguistic constraints. The book reveals insights on learning processes involved in acquiring various Chinese linguistic features, cogitative skills, and metacognitive strategies. These studies provoke additional research questions awaiting future studies. There are several implications for future directions. I will, however, focus on two points: (1) the need for more empirical research to validate the current Chinese

language acquisition research and expand the scope, and (2) the need for innovative research designs with rigorous methodology.

3.1. Validate the Research and Expand the Scope

Further research is warranted to validate the current findings. This is not only because replication studies in Chinese acquisition research are scarce, but also because they are essential to clarify confounding data and to address a mix of findings across studies. Since research is conducted under varied theoretical approaches and methodologies, and in different contexts, empirical findings may yield inconsistency. Research with similar designs but targeting a variety of linguistic or cognitive agendas, as well as with the same linguistic and cognitive topics but conducted from different perspectives and methodology, is needed to build up a body of knowledge that broadens and validates our understanding of the acquisition of Chinese as a first and additional language. With research scope further expanded and different types of data verified, empirical findings should ultimately achieve better consistency.

Studies in this volume discussed further inquiries based on their findings and research limitations, discussions provoking a series of future research agendas. For example, in Liao's study, the relationship between linguistic accuracy and metadiscourse skill development remains intriguing. It is important to examine the learning process to determine if these skills develop simultaneously in parallel, independently, or in an interrelated way. A few studies reported results inconsistent with previous research from other languages. The inconsistencies may be from language specific features, different research settings, and/or variables related to data and data collection. For example, Yanmei Liu's study about the effects of metacognitive strategy on listening performance gains revealed that there was no significant relationship between metacognitive development and listening proficiency. It is necessary to further the inquiry by strengthening the intervention training on cognitive skills in addition to metacognitive strategies.

3.2. Need for Innovative Design with a Rigorous Methodology

Research design and data collection are vital. Robust research methodology and careful design promise more valid studies and reliable results. Drawing on the psycholinguistic framework, this volume may be biased towards the quantitative paradigm, as is true of the majority of research published in journals and books on psycholinguistics. We need to expand traditional methods by integrating a variety of measures. The mixed-methods approach, with both quantitative and qualitative data analyses, is rigorous in terms of data validation. Quantitative methods offer an efficient tool that conceptualizes variables and collects a large amount of data (e.g., via a survey questionnaire) to allow researchers to gain a wider perspective and make generalizations. Qualitative methods provide detailed examinations in context that capture learners' changes in the process. A mixed-methods design, thus, provides a way for a researcher to examine different types of data in order to increase the overall research reliability and validity. In this special issue, qualitative methods are rarely adopted. The majority of the studies, particularly those that adopted only quantitative measures, could have benefited from incorporating various forms of qualitative data to scrutinize the issues and consolidate the study.

One design that deserves more attention is the longitudinal study spanning a substantial length of time. It is a comprehensive tool that keeps track of learners' developments and examines the interactions between learners and their environment. When research is focused on the language development of a group of learners, a longitudinal design, by incorporating multiple factors into the analysis, serves the purpose. It is a "learner focused" design; learners are the "agents" who initiate interactions, which can either positively or negatively influence learning (Ortega and Iberri-Shea 2005; Ushioda and Dörnyei 2012). There is a dearth of longitudinal studies in Chinese language acquisition research. Most studies in this volume employed the cross-sectional design, frequently with a native-speaker group providing the baseline data for comparison.

There are potential improvements in terms of research design and methodology in this volume. Sampling is a first concern. A few studies had small sample sizes upon which, however, rigorous statistical tests were conducted. Even when the sample size is relatively large, if participants are divided into subgroups, the number in each subgroup may be too small for rigorous parametric analysis. Data elicitation methods have been a primary concern. Studies in this volume have illustrated how Chinese-specific features can be identified, elicited, and analyzed to infer learners' linguistic competence. However, qualitative data, such as interviews and portfolio, are sparse. Furthermore, in addition to the learner's production, it is important to conduct process-oriented research to probe into the nature of the learning task and details of the process. Recent years have seen advanced methods of data collection utilizing digital media to capture the accuracy and fluency of learners' performance. Studies should continue to explore innovative data collection and analysis methods that reveal language use "from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction, and the effects their use of language has on other participants in the act of communication." (Crystal 1997, p. 30). All these concerns and issues confronting us promote more robust research designs and innovative research methodology to achieve a broader and more accurate understanding of the acquisition of Chinese as a first and additional language.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

ACTFL. 2011. Foreign Language Enrollments in K-12 Public Schools: Are Students Prepared for a Global Society? Available online: https://www.actfl.org/sites/default/files/pdfs/ReportSummary2011.pdf (accessed on 6 June 2020).

Chen, Jidong, and Bhuvana Narasimhan. 2018. Information structure and ordering preferences in child and adult speech in English. In *The Proceedings of the 42nd Boston University Conference on Language Development*. Edited by Anne B. Bertolini and Maxwell J. Kaplan. Boston: Cascadilla Press, pp. 131–39.

Comanaru, Ruxandra, and Kimberly Noels. 2009. Self-determination, motivation, and the learning of Chinese as a heritage language. *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes* 66. [CrossRef]

Crystal, David. 1997. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Duff, Patricia, and Duanduan Li. 2002. The acquisition and use of perfective aspect in Mandarin. In *Tense-Aspect Morphology in L2 Acquisition*. Edited by Rafael Salaberry and Ysauhiro Shirai. Philadelphia: John Benjamins, pp. 417–54.

Everson, Michael, and Helen Shen, eds. 2010. *Research among Learners of Chinese as a Foreign Language*. Riverdale: National Foreign Language Center, Honolulu: University of Hawaii at Manoa.

Gass, Susan M., and Larry Selinker. 2008. *Second Language Acquisition: An Introductory Course*. Abingdon-on-Thames: Routledge.

Han, Zhaohong, ed. 2014. *Second Language Acquisition of Chinese: A Series of Empirical Studies*. Bristol: Multilingual Matters.

Keating, Gregory D., Bill VanPatten, and Jill Jegerski. 2011. Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Studies in Second Language Acquisition* 33: 193–222. [CrossRef]

Kormos, Judit. 2014. *Speech Production and Second Language Acquisition*. Abingdon-on-Thames: Routledge.

Li, Charles, and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

Li, Yu, Xiaohong Wen, and Tianwei Xie. 2014. CLTA 2012 Survey of college-level Chinese language programs in North America. *Journal of the Chinese Language Teachers Association* 49: 1–49.

Ortega, Lourdes, and Gina Iberri-Shea. 2005. Longitudinal research in second language acquisition: Recent and future directions. *Annual Review of Applied Linguistics* 25: 26–45. [CrossRef]

Polinsky, Maria, and Gregory Scontras. 2019. Understanding heritage languages. *Bilingualism: Language and Cognition* 23: 1–17. [CrossRef]

Sung, Ko-Yin. 2013. L2 motivation in foreign language learning. *Journal of Language and Linguistic Studies* 9: 19–30.
Tao, Hongyin, ed. 2016. *Integrating Chinese Linguistic Research and Language Teaching and Learning*. Amsterdam: John Benjamins Publishing Company.
Ushioda, Ema, and Zoltan Dörnyei. 2012. Motivation. In *The Routledge Handbook of Second Language Acquisition*. Edited by Susan Gass and Alison Mackey. New York: Routledge, pp. 396–409.
Vandergrift, Larry, Christine C. M. Goh, Catherine J. Mareschal, and Marzieh H. Tafaghodtari. 2006. The metacognitive awareness listening questionnaire: Development and validation. *Language Learning* 56: 431–62. [CrossRef]
Wen, Xiaohong, and Meiyu Piao. 2020. Motivational profiles and learning experience across Chinese language proficiency levels. *System* 90: 1–13. [CrossRef]
Wen, Xiaohong, and Xin Jiang, eds. 2019. *Studies on Learning and Teaching Chinese as a Second Language*. London: Routledge.
Wen, Xiaohong. 1995. Second language acquisition of the Chinese particle *le*. *International Journal of Applied Linguistics* 5: 45–62. [CrossRef]
Wen, Xiaohong. 2011. Chinese language learning motivation: A comparative study of heritage and non-heritage learners. *Heritage Language Journal* 8: 41–66.
Wen, Xiaohong. 2012. *Learning and Teaching Chinese as a Second Language*. Beijing: Peking University Press.
Wen, Xiaohong. 2018. Motivation and Chinese second language acquisition. In *Routledge Handbook of Chinese Second Language Acquisition*. Edited by C. Ke. London: Routledge, pp. 352–72.
Xie, Yan. 2014. L2 self of beginning-level heritage and nonheritage postsecondary learners of Chinese. *Foreign Language Annals* 47: 189–203. [CrossRef]

languages

MDPI

Article

Perfective *-le* Use and Consciousness-Raising among Beginner-Level Chinese Learners

Yi Xu

Department of East Asian Languages and Literatures, University of Pittsburgh, Pittsburgh, PA 15260, USA; xuyi@pitt.edu

Received: 3 March 2020; Accepted: 9 April 2020; Published: 17 April 2020

Abstract: Within the framework of explicit learning and consciousness-raising, this study investigates patterns in the use of *-le* in authentic classroom tasks by beginner-level learners of Chinese as a foreign language (CFL). It also explores the role and the processes of student-centered consciousness-raising in explicit knowledge building. Twenty-five participants completed a grammaticality judgment task, an interactive role-play task, and a written editing task. The experiment group received role-play sheets with explicit forms of *-le* provided, and participants engaged in rule induction of *-le* in forbidden context in the role-play session. Results showed that beginner-level learners' difficulty with *-le* use manifested in different ways in these tasks, and *-le* underuse occurred more than overuse in the control group's oral role-play task. Consciousness-raising through unguided small group rule induction supported participants' learning of *-le* usage constraints, shown by differences between the control and experiment groups' performances in the posttest. Through a qualitative analysis of participants' analytical talk transcripts, the processes and outcomes of small group rule induction are examined and discussed.

Keywords: Chinese as a foreign language; consciousness-raising; obligatory and forbidden contexts; perfective aspect marker; rule induction

1. Introduction

Perfective aspect marker *-le* use in Chinese is an important and challenging grammatical area for Chinese as a foreign language (CFL) learners. The topic has drawn interest from linguists and language acquisition researchers since decades ago (e.g., Wen 1995; Wen 1997). More recently, some attempts have been made in investigating effective pedagogical interventions within the framework of form-focused consciousness-raising (Yuan 2012; Yuan 2019). Meanwhile, there are very few empirical studies that have observed learners' performance in regular language classrooms. Pedagogical proposals that integrate form-focused instruction with communicative activities are also lacking. The present research investigates beginner-level CFL learners' use of the perfective aspect in classroom role-play tasks, and discusses the effect of consciousness-raising and rule induction in learners' acquisition of *-le*.

2. Literature Review

2.1. L2 Acquisition Studies of Aspect Marker -le

Particle *-le* as a perfective aspect marker is a crucial component of the Chinese language structure expressing viewpoints. Beginner-level CFL learners are generally introduced to both the perfective marker *-le* and sentence-final *-le*. The two have distinct usages, though they may also incidentally overlap (Xiao and McEnery 2004). Unless otherwise specified, *-le* in this study refers to the former. Studies on *-le* point to a few common usages of the aspect marker, which are directly associated with certain forms. For instance, Li and Thompson (1989) pointed out that *-le* is used with bounded events

and an event can be bounded in one of four ways: (1) by being a quantified event; (2) by being a definite or specific event; (3) when the verb is inherently bounded; or (4) when being the first event in a sequence (pp. 185–86). Such "boundedness" features are respectively reflected in the following forms: (1) V-*le*- Numeral-Measure Word-(O) (referred to as V-*le*-NM below), (2) V-*le*-definite or specific NP, (3) telic and resultative V-*le*, including resultative verb compound (RVC)-*le*, and (4) V1-*le*(O)-V2[1]. In acquisition studies, some earlier researchers summarized common patterns of -*le* usage in similar ways (Ke 2005; Yuan 2012). To some extent, textbooks also introduce the form-function mapping as grammatical rules or language patterns to learners.

Correspondingly, there are a list of contexts where -*le* is disallowed. For instance, habitual and state verbs are generally incompatible with perfective aspect across different languages (Comrie 1976; Smith 1991). -*le* also tends not to go well with "say" types of verbs indicating direct or indirect speech, or verbs taking a clausal object (Duff and Li 2002, pp. 424–25; T'ung and Pollard 1982). Because these negative contexts can be categorized in different ways, and not all of them are suitable pedagogical material for beginner-level CFL learners, the present study deals specifically with the three types of constraints in -*le* usage summarized in Table 1. That is, -*le* is generally not used if the event was a habitual activity, a stative event, or expressions of direct or indirect speech, even if such events took place in the past.[2] These constraints can thus serve as useful pedagogical examples illustrating differences between the English past tense and the Chinese perfective aspect. The constraints are referred to as -*le* forbidden contexts in this paper.

Table 1. Forbidden context of -*le* and examples.

-*le* **Forbidden Environment**	**Sample Sentences**
Habitual activity	*Wo qunian changchang youyong (*le).* "I often swam last year."
State verbs	*Zuotian wo juede (*le) bu-shufu.* "Yesterday I felt uncomfortable."
Expressions of direct/indirect speech	*Wo gaosu (*le) ta, wo jintian bu qu xuexiao.* "I told him that I would not go to school today."

Previous studies also referred to some other -*le* forbidden contexts that may need further scrutiny. Duff and Li (2002, p. 425), citing T'ung and Pollard (1982), claimed that verbs taking a clausal object also do not go with perfective -*le* and gave the following example where -*le* insertion would be ungrammatical: *zuotian women jueding (*le) qu kan na ge dianying* ("Yesterday we decided to go see that movie"). However, the examples given to support this purported "-*le* forbidden rule" tend to involve psych verbs, other state verbs, or verbs related to indirect speech (such as "plan", "decide", "claim"). Thus, it remains dubious to what extent the "clausal object" rule is effective beyond what has already been covered by in stipulations in Table 1. -*le* may also have other constraints. For instance, it typically does not occur in negated form (*wo mei chifan*, "I did not eat"), but as this constraint has more to do with syntactic behaviors rather than semantic aspect, it is of a rather different nature.

In second language acquisition (SLA) studies of Chinese, empirical evidence indicates that -*le* emerges early in learners' language acquisition. In the seminal work of Wen (1995), the author reported that beginner-level learners who had studied Chinese for 14 months used -*le* in several forms, including, V1-*le*(O)-V2, V-NM, V-telic verbs, with 79%, 74%, and 78% accuracy rates, respectively. Similarly high accuracy rates were reported in Wen (1997) among learners with 15–27 months' experience learning

[1] V, O, and NP respectively refer to "verbs", "objects", and "noun phrases". Many but not all RVCs obligatorily require -*le*. According to Jin and Hendriks (2005, p. 71), result-state RVCs with a state verb complement "do not […] depict a process with inherent endpoint" and instead "present the result from a process". In these result-state RVCs (*xi-ganjing* "wash-clean"; *xiuli-hao* "fix well"), -*le* is needed to indicate completion of the event. See *Note 6* for examples of RVCs that can optionally take -*le*.

[2] An anonymous reviewer pointed out that -*le* may be used in some situations of indirect speech, such as in correcting a wrong assumption. There can be marked situations where -*le* is possible in expressing (in)direct speech for emphasis or contrast. For instance, *laoshi yijing shuo-le mingtian bu shangke, ni weishenme hai yao qu xuexiao?* "The teacher already said that there is no class tomorrow, (so) why would you still go to campus?" Care was taken so that materials in this study did not involve these contrasting situations where -*le* may be warranted.

Chinese. Meanwhile, *-le* is infamously difficult even for advanced-level learners (e.g., Yang et al. 1999). One of the major causes of challenge is negative L1 transfer (Tong and Shirai 2016; Wen 1997). Learners with an L1 background of English were found to have a tendency to use *-le* as a past tense marker. As shown by participants' self-reflection reported in Duff and Li (2002), learners may oversupply *-le* in written production tasks, even if they were explicitly aware that the two were not the same. Tong and Shirai (2016) hypothesized that their beginner- to intermediate-level participants were not sensitive enough to the lexical aspect of the verbs. Wen (1997) observed that her participants often adopted a meaning-based approach and relied on contextual cues, such as time adverbials or other expressions (*yiqian* "in the past"; … *de shihou* "at the time of"). While such cues can sometimes be helpful, they are not reliable indicators of *-le* usage, as they are more indicators of past time frame than indicators of viewpoint or boundedness. Researchers also note that the optionality of aspect marking in some contexts in Chinese constitute another source of difficulty for learners. Chen and Shirai (2010) observed that Chinese children erroneously used perfective *-le* with state verbs in early stages of their L1 acquisition, and those authors hypothesized that such deviations from standard usages are due to *-le* use in optional context, making it difficult for learners to induce rules. Corresponding to this finding in L1 acquisition, L2 corpus studies reported similar results. Yang et al. (1999), using corpus of Chinese learners' composition, observed that learners tend to overuse *-le* with stative predicates. In the most recent corpus study examining beginner- to advanced-level L2 learners' speech, Xu et al. (2019) also claimed that across different proficiency levels, learner errors primarily involved ungrammatical use of *-le* with state verbs. These studies argued that overuse was a serious issue in L2 learners' *-le* usage, and attributed it to learners' inadequate knowledge regarding constraints of *-le*.

Researchers have noted that learners' use of *-le* may be subject to task and register variations. For instance, in Duff and Li (2002), the researchers compared learner productions with native speakers' performance. They found that learners underused *-le* in oral narrative tasks, but oversupplied *-le* in a written editing task. Table 2 summarizes the tasks used in previous L2 studies of *-le*.

Table 2. Tasks used in previous L2 studies of perfective *-le* acquisition.

Task Type	Details	Study
Conversation	Questions designed to elicit target aspect markers.	Wen (1995); Wen (1997)
Narrative oral task	Based on a silent film	Duff and Li (2002)
	Personal narrative	Duff and Li (2002)
	Based on sets of picture sequences	Jin and Hendriks (2005)
Picture-based non-narrative oral task	Questions and answers	Wen (1997)
	Description of isolated pictures	
Picture-based writing task	Narrative	Yuan (2019)
	Non-narrative description	Wen (1997)
Judgment and written editing	Edit sentences or paragraph-length texts by supplying aspect markers	Duff and Li (2002); Shi (2013); Tong and Shirai (2016); Yang et al. (2000) Yuan (2012)
Learner corpus studies	Usage and error analysis from corpora data	Yang et al. (1999) (written) Xu et al. (2019) (speech)

These earlier studies, using a variety of tasks, have shed light on features of learner language and its development. However, most of them were conducted using experimental tasks or corpus data, and participants' performance might not directly reflect situations in language classrooms. One pedagogical approach attracting and favored by many practitioners is task-based language teaching (TBLT). Meanwhile, there is evidence that classroom-based oral interactive tasks can yield useful production data for SLA research (e.g., Bygate 2001; Kim 2013). The above review suggests that our

knowledge about students' use of *-le* in language classroom tasks is still very limited, and there exists a disconnection between L2 studies of *-le* in experimental environment and real teaching practice.

2.2. Language Tasks and Consciousness-Raising

Tasks can be used alongside focus-on-forms instruction. That is, in a focused task design, a particular language form can be purposefully embedded in task design to induce learners to use them while performing the task (Ellis 2001). Communicative tasks can also be used alongside other form-focused approaches, such as consciousness-raising (Ellis 2003). Broadly speaking, consciousness-raising is activities or efforts made to make specific language features salient to learners, and can include techniques such as textual enhancement (e.g., bolding, underlining) so that learners' attention can be drawn to certain linguistic features (Leow 2015, p. 167). More often, consciousness-raising tasks are activities in which learners "perform some operations on or with" L2 data, with the goal of achieving an explicit understanding of specific target language properties (Ellis 1997, p. 160). Learners are asked to discover and verbalize target language rules through discussions. Such "metatalks" or "languaging", Swain (2000, 2006), can be operationalized in different ways. Previous L2 studies in other languages have used instructor-guided rule induction (e.g., Adair-Hauck et al. 2010), group interaction with optional intervening from the instructor (e.g., Smart 2014), and individual rule induction in online context (Cerezo et al. 2016). These studies suggested that consciousness-raising rule induction helps promote learners' explicit knowledge of complex linguistic structure.

Given the complexity of rules associated with *-le* usage and learners' continuous difficulties in using it in target-like ways, a form-focused approach to facilitate the L2 acquisition of *-le* is fitting. However, empirical studies that evaluate the effectiveness of pedagogical treatment of *-le* remain scant. Yuan (2012) is the only study that has specifically examined the role of consciousness-raising, form-focused instruction on learners' development of explicit knowledge regarding *-le*. In Yuan (2012) study, ten and eight participants took part in the experimental and control group, respectively, and completed pretest, posttest, and a delayed posttest in the same written paragraph editing task. The experiment group experienced three one-hour consciousness-raising sessions in which participants discussed, reviewed, and did form-function mapping exercises on several categories of *-le* usages. Significant improvement in learners' performances was reported between the pretest and the two posttests for the experiment group, and significant differences were also found between the control and the experiment group in the delayed posttest in overall *-le* use accuracy and several sub-categories of *-le* structures. Yuan's (2012) study confirmed the effectiveness of consciousness-raising activities in CFL learners' *-le* acquisition. However, because her control group did not spend an equal amount of time engaging in comparable learning activities, the comparative effectiveness of consciousness-raising versus other pedagogical approaches is yet to be determined.

In sum, despite the number of attempts made to investigate the L2 acquisition of *-le*, there remains a gap between finely controlled research and the realities of classroom practice. More research is needed on the effect of consciousness-raising as a form-focused pedagogical intervention in today's communicative, student-centered classroom environment. The present study addresses these gaps. The purpose of the study is twofold. First, it investigates what errors tend to occur in beginner-level CFL learners' use of *-le* in their oral and written performances in language classrooms. Second, it explores the role of consciousness-raising by using role-play sheets with explicit markings of language forms and by having learners engage in small group rule induction. As previous studies pointed to learners' lack of knowledge regarding constraints of *-le* usage (Xu et al. 2019; Yang et al. 2000), promoting learners' perception and understandings of *-le* forbidden context is especially important. This study also intends to investigate rule induction within student groups without intervention from instructors. The following are the study's specific research questions:

First, what types of errors can be observed in beginner learners' use of *-le* in classroom activities, including written tasks and oral role-plays?

Second, in communicative language classrooms, does consciousness-raising lead to higher learning outcomes than regular role-play tasks in promoting the learning of *-le* constraints?

Third, how effective are learner-centered rule induction sessions without instructor mediation?

3. Materials and Methods

3.1. Participants

Twenty-five CFL students in a Northeastern American university participated in the study. All participants were native speakers of American English. All participants received a rigorous placement test upon their entrance into the Chinese language program. When the data were collected, participants had had eight months of Chinese learning, and had received approximately 200 hours of classroom instruction. Participants were randomly divided into two groups, with 12 participants in the control group (Group C) and 13 in the consciousness-raising experiment group (Group E). In the role-play and rule induction sessions, participants worked in pairs or small groups of three.

Prior to the study, participants had received instructions on several obligatory contexts of *-le* in their regular course curriculum. Their textbook introduced aspect marker *-le* as a particle indicating "the occurrence or completion of an action or event" (Liu et al. 2008, p. 137). The examples that the textbook gave included patterns of V-*le*-NM, V1-*le*(O)-V2, and V-*le*-proper noun. These illustrations and examples correspond to the "bounded event" environment of obligatory *-le* use specified in Li and Thompson (1989). The textbook also pointed out that in negation of past events, *mei* instead of *mei . . . le* is used. All participants had the same lecture class instructor in their first and second semester study. The instructor confirmed that participants were taught these patterns and usages, and that forbidden contexts other than *mei* negation had not been explicitly taught.

3.2. Materials

The assessment and treatment materials included: a grammaticality judgment pretest, two versions of role-play sheets, one each for Group C and Group E, respectively, and a written passage editing task sheet similar to the one used in Duff and Li (2002).

The pretest of grammaticality judgement included 16 items, 4 of which were fillers. For the remaining 12 items, 6 were grammatical sentences. Three of the six contained obligatory *-le* and the other three were in negative *-le* context (state verb, habitual past activity, and expressions of direct or indirect speech); the other six were ungrammatical sentences, with *-le* missing in *-le* obligatory contexts, or *-le* inserted in the three *-le* forbidden contexts. The 16 items were presented in a randomized order. Participants were instructed to mark the sentence with a "C" ("correct") if they considered the sentence grammatical, and mark it with an "I" ("incorrect") if they considered it ungrammatical. They were further instructed to correct the errors in ungrammatical sentences. Whereas earlier studies tapping into representations in learner language often used grammaticality judgment, the additional step of asking participants to correct errors is important. As Gass and Mackey (2012) pointed out, a caveat of the traditional grammaticality judgment task is that learner language grammar may be "non-native-like in many ways," and "it is necessary to ask learners to correct any sentences that they judge to be unacceptable" (p. 98). Appendix A presents the grammaticality judgment task stimuli.

The role-play sheet described, in English, six scenarios in which *-le* should either be used or is forbidden. In each scenario, speaker A and B received different instructions regarding the situation and how they should act out. The role-play sheets differed minimally between the two versions. The Group E role-play sheet contained explicit written prompts where the perfective marker should or should not be used. For instance, in one scenario, speaker B would tell speaker A about a series of unfortunate incidents, including feeling unwell, taking the wrong bus, and having a stomach problem after eating bad food. In addition to the English prompt, the role-play sheet gave participants keywords with their correct usages of *-le*, i.e., 觉得 (*juede*)∅ "feel", 坐错了 (*zuocuo-le*) "took the wrong (bus)", 吃坏了(*chihuai-le*) "caused a (stomach) problem due to eating bad food", providing the "Verb-*le*" form

when *-le* was obligatory, and the "Verb-∅" form when *-le* was ungrammatical. The scenarios and event verbs included both obligatory *-le* contexts and several examples of *-le* forbidden contexts in each of the following: (1) with state verbs, (2) with habitual activities, and (3) expressions of (in)direct speech in the past time frame. The role-play sheet provided a total of 16 explicit Verb-∅ forms and 9 Verb-*le* forms.[3] For Group E, a rule induction responses sheet was appended to the role-play sheet. Participants were asked to discuss with their partners, in either Chinese or English, to induce when *-le* should/should not be used in past events, and write down their conclusions. Appendix B provides the instructions and sample scenarios in the role-play sheet for Group E.

The role-play sheet for Group C was the same for all descriptions of the scenario, but differed in that it did not explicitly mark when *-le* should be used. Instead, this version of the role-play sheet simply provided participants with the key verbs (e.g., *juede* "feel", *zuocuo* "take.wrong(.bus)", *chihuai* "eat.bad"). No rule induction response sheet was provided to Group C.

For posttest, a written passage editing task sheet was designed. The task sheet contained a passage with two paragraphs, telling a story of a person's trip to and experience in China all in the past time frame. The passage contained 24 blanks and participants were instructed to provide *-le* where it was needed, and write "/" when it should be absent. Among the 24 blanks, 12 were cases where ungrammatical *-le* occurred in forbidden contexts, with four sentences each in the environment of a state verb, habitual activity, or a speech verb. The remaining 12 blanks were cases where *-le* was obligatory, with a telic verb or RVC, or followed by an object that was definite or quantified. See Appendix C for the written editing posttest task.

Only words that participants had previously learned from textbooks were used in the materials. For the grammaticality judgment and written passage editing task sheets, *pinyin* were provided on top of each sentence. All scenarios described in the role-play sheet were appropriate for participants' proficiency level in terms of vocabulary needed and structure usage. The participants' instructor was consulted in finalization of the material.

3.3. Procedure

Both Group E and Group C completed the study on the same day, each in a 50-minute class session in a regular classroom. That is, the two groups completed the study in two different class periods. The researcher, who was incidentally a language instructor herself, monitored classes. Because the study aimed to investigate the effect of consciousness-raising when learner agency is maximized, the researcher instructed participants to complete each step, recorded time, offered assistance with audio-recording, and collected response sheets. She did not offer pedagogical instructions or guidance on completing tasks.

Both groups first completed the grammaticality judgment pretest in a paper-and-pencil format. The procedure took 15 min. Immediately after the pretest sheets were collected, participants were asked to carry out interactive role-play tasks in Chinese with a randomly assigned partner or two partners. Group E and Group C were given their respective role-play sheets.

For Group E, participants used approximately 12 min to act out the scenarios in the role-play sheet in pairs/groups. After they had all completed the role-play tasks, they were instructed to engage in rule induction within their small group by paying attention to usages in the role-play. They were asked to write down their conclusions on a separate sheet and turn in their responses at the end of their discussion. The rule induction discussion took approximately eight minutes. Each pair/group's role-play and rule induction sessions were audio-recorded.

---

[3] This does not mean that participants were necessarily expected to produce all or only these verb forms. As participants often make spontaneous conversations in role-play tasks, these verb forms with and without aspect markers served as examples. In some cases, a specific "Verb-∅" may also be produced by both speakers in a role-play as questions and answers.

For Group C, participants spent approximately 20 min for the role-play task. They were asked to continue studying the role-play situations or switch roles to play again if they completed all scenarios in less than 20 min. Participants in Group C did not engage in rule induction. Their oral production in role-play was audio-recorded.

In the last 15 min of the class, after role-play sheets (and rule induction responses for Group E) were collected, participants were given the written passage editing task as a posttest.

3.4. Scoring and Coding

For the grammaticality judgment task, a score of one was assigned if (1) a participant judged a grammatical sentence to be correct or (2) a participant judged an ungrammatical sentence to be incorrect and also made the right corrections. If a participant judged an ungrammatical sentence to be incorrect but did not identify the location of the ungrammaticality or did not offer target-like corrections, the response received zero score. Zero score was assigned in all other situations.

Audio-recordings of participants' role-play were transcribed. For Group C's oral production, grammatical use and errors of *-le* were coded, including overuse, underuse, and other errors. The researcher and a research assistant coded the usages independently, with 92.5% agreement. Discrepancies were resolved through discussion. For Group E, because correct forms of *-le* use were provided explicitly on the role-play sheet, participants almost always had target-like performance in using *-le*. After confirming that was the case, that part of the data did not enter into further analysis.

For the written passage editing task, each blank was scored either one or zero, based on whether the response was target-like or not.

To answer research question one, namely participants' error patterns in using *-le*, we reported participants' performance in the three tasks. For the grammaticality judgment task, all 25 participants' performance was examined. A paired sample *t*-test revealed no significant differences between the two groups' performance in this task (Group C: Mean = 6.1, S.D. = 1.4; Group E: Mean = 5.3, S.D. = 1.7; $t(23) = 1.25$, $p = 0.22$). For participants' performance in the oral role-play task and the written editing task, only Group C participants' performances were considered relevant.

To answer research question two, namely the outcome of consciousness-raising versus regular role-play tasks, comparisons were made between Group E and Group C participants' scores in *-le* forbidden and *-le* obligatory contexts in the written editing posttest.

To answer research question three, namely the outcome and the processes of rule induction within pairs/small groups, each small group/pair's written responses and transcripts of their rule induction sessions were analyzed qualitatively. Coding categories used in previous studies of consciousness-raising rule induction were used. These codes included *labeling* (observations and naming specific forms and tokens), *categorizing* ("commenting on properties shared by several tokens"), *patterning* ("commenting on links between two categories of forms [. . .] or between form and meaning", and *rule formulation* (generalizing patterns) (Toth et al. 2013; Cerezo et al. 2016, p. 270). In Cerezo et al.'s (2016) study in which they triangulated learning processes with learning outcomes, they focused on depth of L2 awareness and referred to codes including *noticing and reporting* (attention to specific forms and features), *hypothesis formulation* (which overlaps with the definition of *patterning* above), *rule formulation*, *prior knowledge or experience*[4], and *metacognition* (describing feelings about one's progress). As rule induction in this study was unique in several aspects, the researcher considered other types of learning processes to be possible. After going over the overall transcribed data several times, the researcher developed two more codes, *contrasting* and *rationalizing*, explained in the "Results" section. Following a peer review procedure (Merriam 2009), a colleague was asked to read the transcription and assess if these categories of codes effectively captured the data. Due to the exploratory nature

[4] While earlier studies used the term *activation of prior knowledge* (Tomlin and Villa 1994), *prior knowledge* and experience is used here to inclusively refer to activation of existing linguistic knowledge and recalling experiences of language use.

of the research question, we were not interested in drawing statistical correlations between specific processes and outcome in this study. Instead, we aimed to report different types of cognitive processes manifested in student-centered, interactive rule induction, with qualitative analysis of how they may or may not contribute to explicit learning.

4. Results

4.1. Learner Performance across Tasks

To examine learner performances in different tasks, first, all 25 participants' performances in the grammaticality judgment task were analyzed. Participants' scores in judging ungrammatical and grammatical prompts were examined separately, in the *-le* obligatory context and *-le* forbidden context, respectively. Table 3 below shows the descriptive statistics of participants' scores in those situations.

Table 3. Participants' scores in grammaticality judgment.

	***-le* Obligatory Context**		***-le* Forbidden Context**	
	Grammatical prompt	Ungrammatical prompt	Grammatical prompt	Ungrammatical prompt
Mean (S.D.)	2.24 (0.78)	0.44 (0.58)	2.20 (0.76)	0.80 (0.96)

Participants achieved mean scores of 2.24 and 2.20 out of a maximum score of 3 in the *-le* obligatory context and *-le* forbidden context, respectively. In both contexts, their rate of accuracy in identifying the location of error and correcting it was low (0.44 and 0.80, out of a maximum score of 3). In all but three of the zero-scored responses for ungrammatical prompts, participants were actually not able to identify the location of the error.

Stimuli receiving the lowest scores included an item describing habitual activity and an item involving indirect speech, with one and two participants providing target-like responses only. Other items receiving low scores included sentences where *-le* was missing in an obligatory context, with an example in (1).

(1)

zuotian	wo	he	pengyou yiqi	kan	*(le)	Harry Potter
yesterday	I	and	Friend together	read	LE	Harry Potter

"I watched Harry Potter together with friend yesterday."

In general, the pretest results indicate that participants often could not identify situations where *-le* deletion or *-le* insertion were needed. For instance, prompt (1) was directly modified from a sample sentence from the participants' textbook (Liu et al. 2008, p. 139), yet the majority of the participants (18 out of 25) could not identify the location of the error. Participants may not have internalized the knowledge regarding usages and constraints of *-le*.

Second, in the interactive role-play task, both *-le* oversupply and undersupply were witnessed. There were 19 tokens of *-le* undersupply in obligatory context, with nine tokens in RVC context. Except for one role-play small group who correctly produced two RVC-*le* forms, participants in all three other role-play groups constantly failed to supply *-le* in RVC contexts. (2)–(4) provide examples. In addition, there were nine missing *-le* in the V-*le*-NM environment (e.g., *zuotian wo chi *(le) ershi zhi jiaozi* "I had 20 dumplings yesterday"), and one missing *-le* in a specific, completed past event (*Women dou qu *(le)* "We all went").

(2)

Duibuqi	wo zuo-cuo	*(le)	che
sorry	I sit-wrong	LE	bus

"Sorry, I took the wrong bus."

(3)

Wo	chi-huai	*(le)	duzi
I	eat-bad	LE	stomach

"I ate wrong food and had a stomach-ache."

(4)

yinwei	sushe	hen	chao, suoyi	wo	ban-qu	*(le)	gongyu
because	dormitory	very	noisy, so	I	move to	LE	apartment

"Because my dormitory was very noisy, I moved out to an apartment."

In several cases, participants' supply of *-le* seemed arbitrary. They supplied *-le* in one clause containing V-*le*-NM forms, but failed to do so in the conjoined clause with the same structure. (5) provides an example.

(5)

Wo zuotian	chi-le	shi	ge	shuijiao,	hai	wo	he	*(le)	san	bei	kafei
I yesterday	eat-LE	ten	CL	dumplings	also	I	drink	LE	three	cup	coffee

"Yesterday I ate ten dumplings; I also drank three cups of coffee."

Ten tokens of *-le* oversupply in forbidden contexts were found, including four tokens in habitual activities in past time frame, with an example shown in (6), two tokens of ungrammatical *-le* with state verbs, with an example provided in (7), two tokens of ungrammatical *-le* in negation with *mei*, one token of ungrammatical *-le* with indirect speech verbs, and one with a non-telic event.

(6)

Wo	qunian	chang chi	*(le)	henduo	pizza
I	last.year	often eat	LE	many	pizza

"Last year, I used to eat much pizza."

(7)

Wo	zhu zai	Tower C. Wo	ziji	zhu	*(le)
I	live at	Tower C. I	self	live	LE

"I lived in Tower C. I lived by myself."

Although a variety of habitual activity verbs (e.g., *paobu* "run", *youyong* "swim", *wan* "play") and state verbs (e.g., *juede* "feel", *zhidao* "know", *you* "have") were provided as prompts from the role-play tasks, all oversupply cases with habitual and state verbs pertained to *chi* "eat" and *zhu* "live".

There were nine tokens of grammatical *-le* use, two with RVCs, six in the V-*le*-NM environment, and one with a specific, telic event. While there were few tokens of grammatical *-le* used in task, this is within expectation, since the role-play sheets were designed to help learners induce rules of *-le* forbidden context, and *-le* obligatory contexts were provided for the purpose of comparison. Additionally, there were two tokens of non-target-like *le* usage in V-*le*-NM environment, with errors due to word order. An example of word order error is given below as (8). Arguably, despite the word order errors, participants still showed awareness in the V-*le*-NM structure in these cases.

(8)

Wo zai	pida	zhu le	zhu	le	zai sushe	liang	nian	le
I	at University.of.Pittsburgh	live LE	live	LE	dormitory	two	year	LE

"I have lived in the University of Pittsburgh's dormitory for two years."

Finally, for the written editing task, Group C participants' scores and accuracy rate in the *-le* obligatory and *-le* forbidden contexts were examined. In the *-le* obligatory sentences, participants' mean score was 8.83 out of the maximum score of 12, with an accuracy rate of 73.6%. In the *-le* forbidden sentences, the mean was 6.92, with an accuracy rate of 57.6%. One sample, one-tailed *t*-test in comparison with a chance level of 0.50 suggests that the participants' performance in *-le* obligatory context was higher than the chance level ($t(11) = 7.34$, $p < 0.0001$), whereas their performance in *-le*

forbidden context was no more than chance ($t(11) = 1.35$, $p = 0.102$). In other words, there was no evidence that Group C participants had any awareness regarding constraints of *-le* usage.

In sum, beginning level learners' non-target-like performance in *-le* usage manifested in different ways in different tasks. In judgment tasks, their difficulties lay primarily in identifying and correcting errors of perfective *-le* in both obligatory and forbidden contexts. In the role-play task, underuses occurred more often than overuses, and occurred in the context of different verbs and eventualities (quantified, resultative, and specific event). Meanwhile, *-le* overuses appeared to be associated more clearly with certain verbs than others. Other *-le* usage errors related to word order were rare. Finally, results from Group C's written editing task indicate that participants, without experiencing focused treatment, lacked explicit knowledge regarding forbidden contexts of *-le*.

4.2. Effect of Consciousness-raising

To assess if consciousness-raising techniques (including explicit marking of verb-aspect forms and rule induction) led to different results than unfocused interactive activities of role-play, we first examined Group E participants' scores and accuracy rate in the written editing task. Participants' mean scores in *-le* obligatory context was 8.38 out of a maximum score of 12, with an accuracy rate of 69.9%. Participants' mean scores in *-le* forbidden context was 8.92 out of a maximum of 12, with an accuracy rate of 74.4%. One sample, one-tailed *t*-test was performed on accuracy rate in both contexts to make comparisons with the 0.50 chance level. Results showed that participants performed higher than chance level in both contexts: (*-le* obligatory context: $t(12) = 3.94$, $p = 0.001$; *-le* in forbidden context: $t(12) = 4.46$, $p = 0.0004$). In other words, Group E participants showed awareness regarding when *-le* should or should not be supplied in this posttest task.

Next, Group C and Group E participants' performance in the posttest in each of the two situations (*-le* forbidden and *-le* obligatory) was compared. A two-sample *t*-test revealed significant differences between groups in *-le* forbidden context, $t(23) = 2.13$, $p = 0.044$, Hedge's $g = 0.85$. On the other hand, the two-sample *t*-test comparing the two groups' performances in the *-le* obligatory environment revealed no significant differences between groups, $t(23) = 0.63$, $p = 0.54$.

Table 4 presents a summary of participants' mean scores and accuracy rates in the two divergent *-le* contexts. Specifically, Group E participants had better performance than Group C in *-le* forbidden context, demonstrating a development of explicit knowledge regarding *-le* constraints through consciousness-raising.

Table 4. Summary of two groups' performance in posttest.

	***-le* Obligatory Context**		***-le* Forbidden Context**	
	Scores (standard deviations)	Accuracy rate	Scores (standard deviations)	Accuracy rate
Group C	8.83 (1.34)	73.6%	6.92(2.35)	57.6%
Group E	8.38 (2.18)	69.9%	8.92 (2.36)	74.4%

The above shows that in comparison with unfocused role-play tasks, consciousness-raising facilitated participants' explicit learning regarding rules of *-le* forbidden context. Consciousness-raising did not lead to higher outcome compared to unfocused role-play in *-le* obligatory context. This is not surprising, given that participants engaged in rule induction of *-le* forbidden context only in this study. As participants had previously learned about specific rules of obligatory *-le* context, the inclusion of obligatory *-le* in the role-play material did not lead to the development of new knowledge. However, it could have facilitated rule induction by offering participants examples of comparison, as discussed in the next section.

A closer look also revealed that Group E outperformed Group C in all three types of *-le* forbidden environment in the posttest. Their mean scores (out of a maximum of 4) and accuracy rates are reported below in Table 5. Due to the small number of items in each of the three sub-contexts, no inferential

statistical tests were performed. This descriptive result is discussed in triangulation with findings for our third research question, namely the outcome and processes of participants' rule induction.

Table 5. Scores and accuracy rate in three *-le* forbidden situations in posttest.

	Habitual Activities		**State Verbs**		**Speech Verbs**	
	Mean scores	Mean accuracy	Mean scores	Mean accuracy	Mean scores	Mean accuracy
Group C	1.08	27.1%	3.08	77.1%	2.75	68.8%
Group E	2.08	51.9%	3.54	88.5%	3.31	82.7%

4.3. Outcome and Processes of Interactive Rule Induction

Our third question pertained to both the outcome and processes of participants' consciousness-raising rule induction. To examine "outcome", transcripts of participants' paired/small group discussions and their written responses were analyzed to see if they had successfully induced the rules of *-le* forbidden context in the three situations. Table 6 gives a summary of the rules induced by participants, corresponding to the actual linguistic rules. Small groups/pairs within Group E are referred to as "pair 1" to "pair 6" below, though one of the "pairs" involved three participants.

Table 6. Rules induced by participants, corresponding to target language rules [5].

Target Language Rules: *-le* Cannot Be Used	**Rules Induced by Learners**
#1: with habitual activities	• We do not use *-le* when we talk about habit (Pair 1); • Not used when you are just saying "I used to . . . " (Pair 6)
#2: with expressions of (in)direct speech ("say" verbs); Alternatively, with verbs that take a verb construction as their object	• Not used with *gaosu* ("tell"), *wen* ("ask"), *shuo* ("say") (Pair 1); • (Not used with) "speaking words" or "descriptive objects" (Pair 2); • Not used for previous statements (Pair 4)
#3: with state verbs	• Not used with passive verbs; *juede* ("feel") is like passive, *zhidao* ("know") is also passive (Pair 1); • When something stays the same (or) stays constant, you do not use *-le* (Pair 2); • (Not used) when the action cannot be completed, like "like" (*xihuan*) (Pair 3); • If it is like an "opinion" you do not use *-le.* It is like "ongoing" (Pair 4); • (Not used for) something that is felt, not physically done; focused on a state of being (Pair 6)

Note. Quotes were directly taken from transcripts or participants' written response sheets, while words in parentheses were added for clarity.

All three target rules were induced by participants. This corresponds to findings in the posttest, where better a performance was observed in all three *-le* forbidden contexts among Group E in comparison with Group C. Each pair, except for Pair 5, had some success in inducing the rules, though their verbalization of the rules sometimes remained at the level of *patterning* (summarizing specific form-meaning matching without being able to extrapolate it into formal rules), and was not always linguistically accurate.

[5] One pair of participants, Pair 4, also responded that -le should not be used in negation, with mei. While this observation adhered to examples in role-play, this was not considered successful rule induction, since participants had been introduced to that explicit rule in class before the study.

Next, participants' rule induction processes were analyzed. Participants' inputs were coded on two levels: (1) the types of cognitive processes involved (*noticing*, *categorizing*, *labeling*, *prior knowledge or experience*, *patterning*, *rule formulation*, *metacognition*) and (2) whether the process was successful (i.e., leading to rule formulation adhering to target language forms) or not (i.e., incorrect or misleading).

In many cases, the process of rule induction started with *noticing and reporting*. (9) and (10) provide such examples.

(9)
"Ta shuo," she asked him. Don't have *-le*. So how can I say this? (Pair 1)

(10)
They used it here "yesterday", after the verb. (Pair 3)

Participants' *labeling* can naturally lead to *categorization*. Below in (11), participants' induction processes generally followed a path of *noticing-> labeling->categorization-> contrasting*. (1A and 1B refer to the two participants in Pair 1. The same goes for references of other speakers in excerpts that follow.) In this study, *contrasting* is defined as one's effort to make comparisons between rules or features of two or more distinctive linguistic categories. As the focus of participants' rule induction was constraints of *-le* usage, and participants' prior knowledge involved obligatory use of *-le* only, *contrasting* was a useful strategy. In the example below, *categorizing* and *contrasting* can be considered as evidence of participants' achieving deeper awareness (Cerezo et al. 2016), as participants needed to make mental effort to activate several examples, make associations and connections, as well as make distinctions. At the end of the processes, participants arrived at a fairly insightful conclusion that "active verbs" but not "passive verbs" in the past time frame should take *-le*. Although this verbalization of the rule lacked sufficient linguistic accuracy, the participants were able to conceptualize the forbidden versus obligatory environment of *-le*, and developed a heightened sense of awareness through the process.

(11) 1A: It's [not] used when

1B: …eating and drinking,

1A: Hmm???

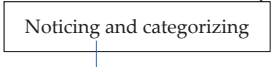
1B: "*chi-le; he-le*"

1A: Hahaha… eating and drinking.

1B: I guess? Maybe for more active verbs.

1A: Hmmm.

1B: Like …*chi-le, he-le, dao-le, pao-le,* which are more like active verbs, than passive verbs, like *dasuan*. Maybe?

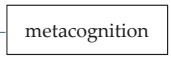


1A: Yeah, hmmm.

1B: [turning page to find other examples] Yeah, *juede* is like passive. It doesn't need *-le* either.

1A: Hmmm.

1B: I am gonna say that. I think it's solid enough.



Since *noticing* reflects low levels of awareness, and higher levels of awareness such as *categorization* require more amount of cognitive effort (Leow 2012), rule induction can progress linearly from *noticing* to other cognitive processes. However, the path of rule induction can also be interactive, especially in this study, in which rule induction was collaborative. For instance, below in (12), a participant hypothesized a pattern first, and then referred back to what they noticed in the role-play. This can

be seen as participant 4B's effort to integrate *noticing* with *patterning*, and it was also 4B's attempt to guide the partner to engage in *noticing*. In this excerpt, participant 4A made reciprocal contribution by offering his *prior experience* to support 4B's generalization. Both participants' input can be seen as acts of scaffolding or knowledge co-construction.

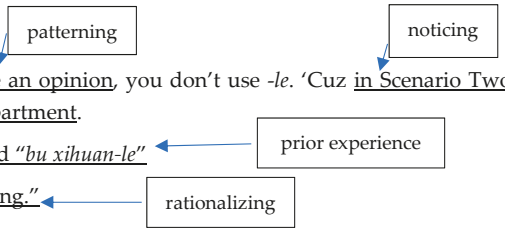
(12) 4A: *-le* is not used when

4B：It's like if anything after is like an opinion, you don't use *-le*. 'Cuz in Scenario Two, you'd ask, if he liked living in his apartment.

4A: Yeah, I don't think I've ever used "*bu xihuan-le*"

4B: Yeah... I guess that is like "ongoing."

4A: *Dui* ("Right")!



While the above examples show successful consciousness-raising through pair interaction, rule induction without instructor guidance can be unsuccessful. First of all, *noticing* did not always lead to successful rule formulation. In the following, unsuccessful rule formulation appeared to be an effect of negative L1 transfer. Both participants felt that it made sense if *-le* was forbidden in the past time frame when a statement remained true, which is a linguistic pattern of the English past tense inflection (*-ed*).

(13) 4A: I don't know why it's not used in this.

4B: "Travel agency said," . . . *shuo* . . . , ah, probably because it's the listing.

4A: Oh, like the price doesn't change?

4B: It's still $800, even though they said it in the past.

4A: Oh, so like, the actual thing, like the thing they are saying still didn't change. 4B: Still true.

4A: Oh yeah, that makes sense.

In some cases, negotiation within pairs could be unhelpful. In excerpt (14), 2A initially started with a *labeling* that was descriptively close to the target rule, mentioning "speaking word". He also had a rough sense that the main verb takes a clausal object in those *-le* forbidden contexts.[6] However, while 2A struggled to conceptualize and verbalize linguistic features and categories, the label "descriptive object" was inaccurate, leading to misunderstandings among group members, and the attempt was unfruitful. (Only relevant part of the transcript is shown.)

(14) 2A: I feel like they are more "speaking" word.

2B: 'Cuz you have *juede*, and *zhidao*.

2C: We can't categorize it.

2B: I don't know.

[. . . . .]

2A: [pause] They are objects that are described.

[6] "Say" verb or "speaking word" itself does not immediately make *-le* forbidden in the environment. While Duff and Li (2002) referred to it generically as a "say" verb environment, it is the expression of (in)direction speech that would make *-le* generally unacceptable. For instance, "say" verbs taking a quantified noun phrase such as *wo shuo-le ta jiju* ('I scolded him a bit.') obligatorily takes *-le* because it fits the V-*le*-NM pattern and it does not express (in)direct speech. Similarly, it is arguable if verbs staking clausal object constitute *-le* forbidden environment. Despite the potential inadequacies of these attempted rule formulations, it is clear that participants were starting to uncover the linguistic features associated with *-le* and that their observations came close to generalizations made by some linguists and grammarians.

2A: [pause] It's like an adverb . . . ; it's like an adverb and a verb.

2B/2C: I don't know.

2A: This one's like describing "when." This one's describing like . . . .

2B/2C: A descriptive adjective. Descriptive adverb?

2C: I am making up new words.

2A: I don't know how to describe it.

2C: Wait. What did you say before?

2A: Oh, it was basically like, all of the . . . all of the like . . . actions are being described. So like like, the Beijing part is, like, for the summer. Like, it's like describing how long you are doing it. And then, for the travel agency, you are describing, like how much, like, the ticket cost. So like, You are describing . . . the action based on how long it takes or how much it costs.

2B: So, like, if you said, *wo chi-le yi-ge pingguo* versus *wo chi* . . . [. . . . .]

2B/2C: *wo qu shangdian mai de* . . .

2A: *mai-wan*, so like . . .

2B: No, that doesn't make sense. You can say *mai-wan-le*

2A: Oh yeah.

2C: So you can say like "Descriptive object." "Descriptive objects" basically?

2A: Yeah, so it's like the object is being described.

Furthermore, due to the complex structures of the *-le* forbidden environment, participants also had difficulty categorizing different *-le* forbidden situations. Excerpt (15) below from Pair 3 can be considered "unsuccessful categorization". Participants attempted to consider different types of verbs (speech verbs, state verbs) as one category, making it difficult for them to identify actual patterns.

(15) 3A: I don't know. 'Cuz this one is like "I planned to go"; "he told me", which is completed, but they didn't use it when it's the past. But "he told me . . . "

3B: Yeah. They used it here yesterday, after the verb.

3A: Ah, you like something. Or you . . . , hmmm. I think it might be something like . . . it's an action, but it's like . . . .

3B: You can't complete it. Right? Something like that.

3A: (Yeah) . . . Like he told me. But . . .

3B: Yeah. like you cannot complete "told."

3B: So, not used when action . . . Just put like [you cannot use *-le* with verbs like] "like" and "asked"

In sum, participants' interactive rule induction led to some level of success in explicit rule formulation. Whereas rule induction processes can start with *noticing* and proceed to deeper levels of processing such as *categorizing* and *patterning*, the induction process within pairs/small groups was interactive in nature, with evidence of knowledge co-construction when the process was successful. Meanwhile, negotiation within pairs/small groups was not always fruitful.

5. Discussion

5.1. Learner Language Features

Learners' challenges in using *-le* manifested in different ways in different tasks in this study. First of all, results indicate that in beginner learners' oral production, *-le* underuse may be a more problematic area than overuse. While there were only 9 prompt verbs in the *-le* obligatory environment versus 16 prompt verbs in the *-le* forbidden environment in the role-play instruction, underuse was the most common error type. In the *-le* obligatory RVC environment, participants correctly supplied *-le* in only two tokens, missing *-le* in nine other tokens. The results corroborate findings from earlier studies that elicited oral production data. For instance, in Duff and Li (2002) film-based oral narrative task, native speakers provided plentiful RVCs, all marked with *-le*, whereas learners produced much fewer RVC tokens and missed *-le* 10 times while correctly providing *-le* only 8 times in the RVC environment. Wen (1995) also noted that beginner-level learners often missed *-le* in RVCs and hypothesized that learners may consider the resultative complement "as an indicator of completion", thus omitting *-le*.[7] Wen's (1995) hypothesis is probable, and in fact, learners would not be entirely wrong if they had such an analysis of resultative complement. According to Xiao and McEnery (2004), RVCs mark the "completive aspect", whereas *-le* marks the "actual aspect". In some situations, a resultative complement can replace *-le* because the two play the same functions in "perfectivis[ing] a situation" (p. 166). For instance, in (16a-b) and (17a-b), either the resultative complement or perfective *-le* can make the following sentences grammatical.

(16)

a.

Wo	chi-wan	fan	jiu	qu kanshu
I	eat-finish	rice	then go	read book

"I will go read books right after I finish my meal."

b.

Wo chi	le	fan	jiu	qu kanshu
I eat	LE	rice	then go	read book

"I will go read books right after I have had my meal."

(17)

a.

Wo xiang-chulai	yi	ge	banfa
I think-out	one	CL	method

"I thought of an idea."

b.

Wo	xiang	le	yi ge	banfa
I	think	LE	one CL	method

"I had an idea."

Given those cases where a resultative complement is interchangeable with *-le* in functions to bound an event or mark completion, it is not surprising that learners tend to drop *-le* in RVCs.[8] Compared

[7] In these tokens of errors in Wen (1995), missing *-le* occurred in positions that were incidentally both verb-final and sentence-final. In many cases, either an aspect marker -le or sentence final -le is needed with RVC to indicate boundedness and completion, or change of state (e.g., *wo xuehui-le zhongwen*; *wo xuehui zhongwen le*, "*I have learned and acquired Chinese*").

[8] In some RVC context, -le can be optional. For instance, *wo xiang-chulai (le) yige banfa* "I got an idea"; *ta dailai (le) yi-ben shu*, "He brought a book"; *wo zoujin wuzi* "I came into the room". The optionality can be explained if the "perfectivization" or boundedness function of *-le* can be achieved with certain complements such as these with [+punctual] features (Xiao and McEnery 2004). The optionality of *-le* in these RVC can make form-function mapping difficult for learners.

to other obligatory context of *-le* such as in V-*le*-NM or V1-*le*(O)-V2 structures, achieving native-like performance in using RVCs with *-le* may be especially difficult for learners.

The findings of more underuse than overuse in oral production but not necessarily in written editing also parallel with Duff and Li (2002). While the corpus study by Xu et al. (2019) referred to overuse as the most prevalent error type in learner speech, the differences may be due to the error tagging methodology used in corpus research.[9] Another possibility is that beginner-level learners in their very initial stage of acquisition tend to undersupply *-le*, whereas oversupply is an error type associated with more "intermediate"-level learners. In Yuan's (2019) study comparing her experimental and control groups' performances in posttest, she found that her experimental groups who had experienced learning tasks overused *-le* significantly more than the control group. Yuan (2019) explained that overuse can be considered learners' effort to experiment with the structure, a sign of language development, whereas underuse reflects an avoidance strategy and thus indicates an earlier developmental stage. As participants in this study had learned Chinese for approximately eight months, it is likely that they had not yet developed into the stage of experimenting with more usages. Future cross-sectional studies are needed to test learners' underuse and overuse in different developmental stages of acquisition.

Participants' oral production also indicates that *-le* oversupply may be associated with specific verbs. Specifically, participants oversupplied *-le* with *zhu* "live", but not with other state verbs such as *xihuan* "like". Participants also oversupplied the marker in *chi-le* in habitual activities but not with other verbs (*youyong* "swim"; *yundong* "physical exercise"). The specific semantic features of verbs may play a role here. *zhu* "live", for instance, is an "interval state" (physical or spatial configurations similar to *sit*, *stand*, etc.), in the words of Dowty (1979). It is different from mental state verbs such as "like" or "feel". In English, "live" can be in progressive aspect, and in both Chinese and English, *zhu* may be more prone to expressions such as "for a certain period of time" (thus requiring *-le* in Chinese) than mental state verbs. In Yang et al. (1999), all six tokens of state V-*le* tokens in their corpus were associated with *zhu* "live" and they occurred in V-*le*-NM structures such as *wo zai shanghai zhu-le liangnian* "I lived in Shanghai for two years". In other words, learners may have encountered forms of *zhu-le* in V-*le*-NM structure, while such usages of *xihuan-le*, *juede-le* are rare or ungrammatical. As to why *(changchang) chi-le* occurred in learner production but not *youyong-le* or *yundong-le*, one possibility could be their different levels of eventiveness: *chi* is a prototypical verb in Chinese, with high "eventiveness" (Monahan and Brunson 2014), whereas "swim", "exercise" can be nouns indicating events in Chinese. As *-le* is a "dynamic" particle (Liu et al. 2008, p. 137), learners may have more tendency in using it with verbs that are more prototypically associated with dynamic features. Another possible explanation could be participants' L2 exposure: in the textbooks that participants used, *yundong* was initially introduced as a noun. When it was used as a verb, it occurred in a habitual activity context (*yige xingqi yundong liang san ci*, "exercise a couple of times a week"). Similarly, *youyong* was initially introduced in a nominal-like way in the textbook (e.g., *ni qu youyong ba*, "How about you do swimming as an exercise?"). In addition, it is reasonable to expect that *yundong* and *youyong* occurred in much lower frequencies with *-le* than *chi-le* in L2 learners' exposure. L2 participants in this study may perceive *youyong* and *yundong* without *-le* as the more salient usages. If so, then the absence of *-le* oversupply with these verbs did not indicate acquisition; rather, participants likely had not developed an awareness regarding usages or prohibitions of *-le* yet, and were simply not using *youyong*/*yundong* with *-le* because it was not a prominent usage in their input.[10] In sum, oversupply and absence of *-le* may be related to

---

[9] Xu et al. (2019) relied on the corpus' existing tags to code errors and found their data by extracting all samples containing *-le*. They found zero tokens of *-le* underuse in their corpus of 443,712 tokens. This is surprising, as the researchers themselves acknowledged, and they suggested further investigations in future research.

[10] I thank an anonymous reviewer for pointing out this plausible explanation from the L2 input perspective. It is also interesting to note that among the six grammatical prompts in the grammaticality judgment pretest, participants had the lowest performance in item #4 (*wo zuotian da-le lanqiu, hai you-le-yong*, "yesterday I played basketball and swam"), with 11 out of the 25 participants (44%) correctly judging it to be grammatical. In comparison, average accuracy rate for the six grammatical

finely categorized semantic features of specific verbs or prominent usages in participants' L2 exposure. These potential explanations warrant further investigations in future research.

Overuse and underuse are not competing categories of errors, and pedagogical attention is needed for both. The written editing task results show that beginner-level learners had little to no knowledge regarding when *-le* should not be used, resulting in only chance level performance in *-le* forbidden sentences in the task. Explicit learning of *-le* constraints can help learners differentiate *-le* from past tense marker and encourage more target-like uses of perfective *-le*.

5.2. Consciousness-Raising

To facilitate the learning of explicit knowledge, consciousness-raising in this study was implemented through both explicit markings on the role-play sheet, and rule induction sessions within pairs/small groups. Compared to earlier studies (Wagner and Toth 2013; Yuan 2012), the present research is unique in its attempt to maximize student agency and autonomy, and there could be several advantages. First of all, without teacher mediation, group members had equal status as non-experts and they may be more at ease in contributing to the rule induction process. Secondly, compared to guided rule induction, where learners engage in form-function mapping tasks with examples and categories readily available (e.g., Wagner and Toth 2013; Yuan 2012), the tasks involved in unguided rule induction were more challenging. In finding relevant examples from role-play, coming up with appropriate categorization, participants needed to stretch the limits of their prior knowledge. After interpreting language samples, they also needed to convey meaning to others. In verbalizing their "analyses" and offering explanations to peers, they needed to construct clear and coherent representations of their own understanding (Van Lier 1996). This in itself required participants to build on higher levels of awareness. Third, in paired/small group rule induction, there was evidence of scaffolding within pairs and small groups through peer interaction, and they helped fill in each other's knowledge gaps, as we witnessed in some examples. Participants' cognitive process appeared multi-directional in this study, and *noticing* (or referencing to what one had noticed) was intertwined with one's rule or hypothesis formulation (i.e., *patterning*) and *rationalizing*. While the effectiveness of rule induction is generally believed to be relevant to greater depth of processing (i.e., one's increased mental effort in problem-solving) (Craik and Lockhart 1972), with the interactive and challenging natures of small group analytical talks, unguided rule induction can lead to particularly meaningful and memorable experiences that help learning.

The use of small group/paired rule induction without instructor mediation had obvious limitations. As Table 6 illustrates, participants' rule formulation may not have always been descriptively accurate, even if they had developed some insight into the underlying mechanisms governing linguistic phenomena. As learners' competency in "languaging" varied, they may also have used informal or inaccurate terminologies due to incomplete metalinguistic knowledge, and this may lead to misunderstandings or ineffective negotiation, as shown by excerpt (14). Learners are not always attuned to peers' needs and it is possible that some group/pair interactions may be dominated by one particular member. For instance, in the successful induction example of (11), it remains unclear if speaker B's rule formulation was explained in accessible ways to speaker A. Without expert guidance, there was a lack of opportunities to check if group members had understood the rules inducted. Further, the effectiveness of small group/pair rule induction may be dependent on specific group dynamics. In this study, one particular pair did not succeed in inducing any of the target rules. A review of their interaction transcript indicated that the pair was not fully committed to the task and relied primarily on prior knowledge instead of observations and analytical talks. Nevertheless, as participants in the

prompts was 74%. In other words, participants might be generally unfamiliar with the *youyong* in the perfective form, and were thus unlikely to oversupply *-le* with this verb in prohibitory environment in the role-play task.

consciousness-raising condition had better performance in the posttest than the control group, the study confirmed the effectiveness of student-centered interactive rule-induction within small groups.

6. Limitations and Implications

The present research studied beginner-level learners' usage and errors of *-le* in interactive oral tasks and in written editing tasks. We found that consciousness-raising through explicit form making and student-centered rule induction supported explicit learning of *-le* constraints. The study was carried out in a real classroom setting, and the pedagogical intervention was short and effective. As Yuan (2012, p. 85) pointed out, consciousness-raising sessions that stretch over several class periods may not represent actual practices that teachers can implement. Thus, short pedagogical intervention sessions have high ecological validity. Though the study did not involve a large number of stimuli in each task, they included rules of aspect marker *-le* in a variety of obligatory and forbidden contexts, so that results can meaningfully indicate learner language patterns. Future studies on learners' *-le* acquisition can further investigate learners' use of the aspect marker with RVCs and with specific verbs, and include considerations of the *-le* optional context. Due to the timing of the experiment, a delayed posttest was not implemented, but future researchers should attend to this issue, so as to assess the extent to which learners internalize explicit knowledge gained through consciousness-raising. This study is one of the first few to use an unguided interactive rule induction approach, and the comparative effect of guided instruction versus student-centered or self-guided rule induction should be further studied.

Development of linguistic competence in complex grammatical rules such as aspect *-le* is a slow process, requiring elaborate and intensive treatment. The techniques used in this study can provide a successful example of embedding focus-on-forms in language tasks in the classroom. It is also hoped that this classroom-based research takes one step further in the field's effort to bridge the gap between controlled lab-based experiments and natural interactions in language classrooms.

Funding: This research received no external funding.

Acknowledgments: The author wishes to thank the participants of this study for their time. She is also immensely grateful for the three anonymous reviewers and the guest editor for their constructive feedback. All errors are her own.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Grammaticality Judgement Pretest

The original pretest was given in Chinese characters, with *pinyin* on top. Ungrammaticality is indicated by * here and not in the task. English translations for stimuli are provided here and not in the actual task. Filler sentences are not included below.

Instructions: Judge the grammaticality of the following sentences. If it is grammatical, put "C" for "correct" in the parentheses. If the sentence has a grammatical error, put "I" for "incorrect". Correct the errors for those incorrect sentences by editing directly on the sentences.

(C) 1 *Wo xiaoshihou hen xihuan chongwu, danshi xianzai wo mei yang chongwu*. "I used to like pets when I was little, but I do not have pets now."

(I) 2 *Zai jiazhou shixi de shihou, wo youde shihou qu (*le) zhongcanguan chifan*. "When I was in California for my internship, I sometimes went to Chinese restaurants to eat."

(C) 3 *Xiao Ming gaosu wo jintian you kaoshi, suoyi wo zhunbei de henhao*. "Xiao Ming told me that there is a test today, so I prepared very well."

(C) 4 *Wo zuotian da le lanqiu, hai you le yong*. "Yesterday I played basketball and swam."

(I) 5 *Zhongxue de shihou wo xiwang (*le) xue xibanya yu, keshi xianzai wo xue zhongwen*. "When I was in middle school, I hoped to learn Spanish, but now I study Chinese."

(C) 6 *Wo zhongwu he le san-bei kele, xianzai bu ke*. "I drank three cups of coke at noon, and I am not thirsty now."

(I) 7 *Xiao Gao shuo (*le) ta bu qu jintian de wanhui, yinwei ta mei shijian*. "Little Gao said that he would not go to tonight's party, because he does not have time."

(I) 8 *Wo yijing zuo-hao *(le) jintian de zhongwen gongke*. "I already finished today's Chinese homework."

(C) 9 *Wo qunian xuexi hen mang, suoyi changchang 12 dian cai shuijiao*. "I studied very busily last year, so I often went to sleep as late as 12AM."

(I) 10 *Zuotian wanshang wo he pengyou yiqi kan *(le) "Harry Potter."* "My friend and I watched Harry Potter together last night."

(C) 11 *Zhe pian kewen youdian'er nan, ni kandong le ma?* "This text is a bit difficult; do you understand it (after reading)?"

(I) 12 *Zuotian de wanhui, Xiao Zhang chi *(le) shi ge jiaozi*. "Xiao Zhang ate ten dumplings at yesterday's party."

Appendix B. Instructions and Sample Scenarios in the Role-play Sheet for Group E

In the role-play sheets participants received, Chinese characters instead of *pinyin* were shown for each prompt verb. Instructions for two scenarios were provided as examples, while the original role-play sheet contained four scenarios.

Instructions: The following are scenarios that you would act out with your partner, with one person taking up the role of either A or B. Use Chinese only to act out the scenario.

If the key verbs in parentheses appears as (VØ) such as (*zuo* Ø), it means *le* is not used. If it appears as (V-*le*) such as (*zuo-le*), *le* should be used. Pay attention to these when role-playing.

After completing all the four scenarios, turn to page 3 and discuss with your partner regarding the use of *–le*.

Appendix B.1. Scenario One

A: Last week, you and your friends made a number of plans from Monday to Thursday, including going shopping, playing tennis, going to a movie, having dinner, etc. But one of your friends did not show up in any of the events. Find out why he/she missed for each of these activities. (*weishenme mei* **VØ**)

B: You previously made plans with your friends for several activities last week, but didn't go to any, because of the following on different occasions:

You felt unwell (*juede* **Ø**);

You knew that the weather was going to be bad (*zhidao* **Ø**);

You took the wrong bus (*zuocuo* ***le*** *che*);

You ruined your stomach by having bad food (*chi-huai* ***le*** *duzi*).

Apologize and propose new plans with your friend.

Appendix B.2. Scenario Two

A: You are new to campus living. Talk to your friend, who lived in a dorm in the past but just moved out. Ask him/her about his/her former dorm life, including:

where he lived (*zhu* **Ø**)'

how many roommates he had (*you* **Ø**) and if he/she frequently spent time with the roommate(s) (*gen . . . yiqi wan* **Ø**)'

If he frequently ate out (*zai waimian chifan* **Ø**);

If he liked (*xihuan* **Ø**) living in an apartment;

Further, ask him/her why he moved.

B: You lived in a dorm last semester but moved to an apartment this semester. Your friend asked you about your former dorm life. Use *zhu*; *you*; *wan*; *chifan*; *xihuan*, etc. to answer his/her questions.

Then, tell him that you decided to change to an apartment now because you had been at Pitt for two years (*xuexi* ***le***) and had lived in a dorm for two years (*zhu* ***le***), and you want to live in someplace new.

Appendix C. Written Editing Posttest

The original posttest was given with Chinese characters and pinyin on top. [+le] indicates the 12 obligatory contexts for *-le* usage. [h], [s], and [i] respectively indicate cases where *-le* is ungrammatical due to being in "habitual activity," "state verb," and "indirect/direct speech verb" contexts.

Instructions: Please provide *le* in the blank where it is needed. If *le* should be absent in that sentence, write /.

Qunian wo de nanpengyou Xiao Lin zai Zhongguo gongzuo [h], suoyi wo qu Zhonguo lvxing [+le] yi-nian. Wo xingqi-liu dao [+le] Beijing jichang yihou, juede [s] youdian ke. Kandao jichang li you [s] yi-ge Xingbake (Starbucks), jiu zou [+le] jinqu. Wo yiqian zai Meiguo hen ai [s] he kafei. Danshi wo bu xihuan [s] na bei kafei, suoyi wo huan [+le] yi-bei cha. Yibian he yibian deng Xiao Lin. Wo zai Xingbake zuo [+le] ban ge xiaoshi, Xiao Lin jiu lai [+le]. Xiao Lin wen [i] wo: "Ni zai feiji shang zuo [+le] shenme?" Wo shuo [i], "Wo zai feiji shang shui [+le] wu-ge xiaoshi de jiao, kanwan [+le] yi-ben shu." Xiao Lin gaosu [i] wo, ta you e you ke. Wo mashang shuo [i]: "Na women yiqi qu chi wanfan ba!" Na tian women dian [+le] henduo cai, zuihou ba qian dou yongwan [+le].

Zai Zhongguo de shihou, wo mei ge xingqi dou gei babamama da [h] dianhua. Wo ye kaishi xihuan-shang henduo xin dongxi. Biru, wo yiqian he [h] kafei, danshi xianzai ai he cha. Wo yiqian ye changchang chi [h] Faguo cai, keshi xianzai geng xihuan chi Zhongguo jiaozi. Wo hai xuehui [+le] xie hanzi.

English Translations of the Passage (Not Provided in Posttest to Participants)

Last year, my boyfriend, Xiao Lin, went to China to work, so I travelled and went to China for a year. I arrived at the Beijing Airport on a Saturday and felt thirsty when I arrived. I saw a Starbucks coffeeshop in the airport, and went in. I used to love drinking coffee when I was in America, but I did not like that cup of coffee, so I exchanged it for a cup of tea. I was waiting for Xiao Lin and drinking tea at the same time. I sat in Starbucks for half an hour before Xiao Lin came. Xiao Lin asked me, "What did you do on the plane?" I said, "I had a five-hour sleep on the airplane and finished reading a book." Xiao Lin told me that he was both thirsty and hungry. I immediately said that we (should) go get dinner. On that day, we ordered several dishes, and in the end spent all the money.

While in China, I called my parents every week. I started to like many new things. For instance, I used to drink coffee, but I now love drinking tea. I used to eat French food frequently, but now I prefer Chinese dumplings. I also learned how to write Chinese characters.

References

Adair-Hauck, Bonnie, Richard Donato, and Philomenna Cumo-Johanssen. 2010. Using a story-based approach to teach grammar. In *Teacher's Handbook: Contextualized Foreign Language Instruction*, 4th ed. Edited by Judith L. Shrum and Eileen W. Glisan. Boston: Heinle, Cengage Learning, pp. 216–44.

Bygate, Martin. 2001. Effects of task repetition on the structure and control of oral language. In *Researching Pedagogical Tasks: Second Language Learning, Teaching and Testing*. Edited by Martin Bygate Peter Skehan and Merrill Swain. Harlow: Pearson Education, pp. 23–48.

Cerezo, Luis, Allison Caras, and Ronald P. Leow. 2016. The effectiveness of guided induction versus deductive instruction on the development of complex Spanish *gustar* structures. *Studies in Second Language Acquisition* 38: 265–91. [CrossRef]

Chen, Jidong, and Yasuhiro Shirai. 2010. The development of aspectual marking in child Mandarin Chinese. *Applied Psycholinguistics* 31: 1–28. [CrossRef]

Comrie, Bernard. 1976. *Aspect*. Cambridge: CUP.

Craik, Fergus. I. M., and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11: 671–84. [CrossRef]

Dowty, David R. 1979. *Word Meaning and Montague Grammar*. Dordrecht: D. Reidel.

Duff, Patricia A., and Duanduan Li. 2002. The acquisition and use of perfective aspect in Mandarin. In *Tense-Aspect Morphology in L2 Acquisition*. Edited by Rafael M. Salaberry and Yasuhiro Shirai. Philadelphia: John Benjamins, pp. 417–54.

Ellis, Rod. 1997. *SLA Research and Language Teaching*. Oxford: Oxford University Press.

Ellis, Rod. 2001. Investigating form-focused instruction. In *Form-focused Instruction and Second Language Learning*. Edited by Rod Ellis. Malden: Blackwell, pp. 1–46.

Ellis, Rod. 2003. *Task-based Language Learning and Teaching*. Oxford: Oxford University Press.

Gass, Susan, and Alison Mackey. 2012. *Data Elicitation for Second and Foreign Language Research*. Beijing: Foreign Language Teaching & Research Press.

Jin, Limin, and Henriëtte Hendriks. 2005. The development of aspect marking in L1 and L2 Chinese. *Working Papers in English and Applied Linguistics* 9: 69–99.

Ke, Chuanren. 2005. Acquisition patterns of Chinese linguistics features for CFL learners. *Journal of the Chinese Language Teachers Association* 40: 1–24.

Kim, YouJin. 2013. Promoting attention to form through task repetition in a Korean EFL context. In *Second Language Interaction in Diverse Educational Contexts*. Edited by Kim McDonough and Alison Mackey. Philadelphia: Benjamins, pp. 3–24.

Leow, Ronald P. 2012. Explicit and implicit learning in the L2 classroom. What does research suggest? *The European Journal of Applied Linguistics and TEFL* 2: 117–29.

Leow, Ronald P. 2015. *Explicit Learning in the L2 Classroom*. New York: Routledge.

Li, Charles N., and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Liu, Yuehua, Tao-Chung Yao, Nyan-Ping Bi, Liangyan Ge, and Yaohua Shi. 2008. *Integrated Chinese*, 3rd ed. Boston: Cheng & Tsui.

Merriam, Sharan B. 2009. *Qualitative Research: A Guide to Design and Implementation*. San Francisco: Jossey-Bass.

Monahan, Sean, and Mary Brunson. 2014. Qualities of eventiveness. Paper presented at the 2nd Workshop on EVENTS: Definition, Detection, Conference, and Representation, Baltimore, MA, USA, June 22–27; pp. 59–67.

Shi, Yasuhiro. 2013. Analysis of L2 learners' knowledge of perfective *le* in Mandarin. *Arizona Working Papers in SLA and Teaching* 19: 40–54.

Smart, Jonathan. 2014. The role of guided induction in paper-based data-driven learning. *ReCALL* 26: 184–201. [CrossRef]

Smith, Carlota S. 1991. *The Parameter of Aspect*. Dordrecht: Kluwer.

Swain, Merrill. 2000. The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In *Sociocultural Theory and Second Language Learning*. Edited by James P. Lantolf. Oxford: Oxford University Press, pp. 97–114.

Swain, Merrill. 2006. Languaging, agency and collaboration in advanced language proficiency. In *Advancing Language Learning: The Contribution of Halliday and Vygotsky*. Edited by Heidi Byrnes. London: Continuum, pp. 95–108.

Tomlin, Russell S., and Victor Villa. 1994. Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition* 16: 183–203. [CrossRef]

Tong, Xiner, and Yasuhiro Shirai. 2016. L2 Acquisition of Mandarin *zai* and *-le*. *Chinese as a Second Language Research* 5: 1–25. [CrossRef]

Toth, Paul D., Elvis Wagner, and Kara Moranski. 2013. "Co-constructing" explicit L2 knowledge with high school Spanish learners through guided induction. *Applied Linguistics* 34: 279–303. [CrossRef]

T'ung, Ping-cheng, and David E. Pollard. 1982. *Colloquial Chinese*. New York: Routledge.

Van Lier, Leo. 1996. *Interaction in the Language Curriculum: Awareness, Autonomy and Authenticity*. London: Longman.

Wagner, Elvis, and Paul Toth. 2013. Building explicit L2 Spanish knowledge through guided induction in small-group and whole-class interaction. In *Second Language Interaction in Diverse Educational Contexts*. Edited by Kim McDonough and Alison Mackey. Philadelphia: Benjamins, pp. 90–125.

Wen, Xiaohong. 1995. Second language acquisition of the Chinese particle *le*. *International Journal of Applied Linguistics* 5: 45–62. [CrossRef]

Wen, Xiaohong. 1997. Acquisition of Chinese aspect: An analysis of the interlanguage of learners of Chinese as a foreign language. *ITL: Review of Applied Linguistics* 117–118: 1–26. [CrossRef]

Xiao, Richard, and Tony McEnery. 2004. *Aspect in Mandarin Chinese*. Philadelphia: John Benjamins.

Xu, Hai, Xiaofei Lu, and Vaclav Brezina. 2019. Acquisition of the Chinese Particle le by L2 Learners: A Corpus-Based Approach. In *Computational and Corpus Approaches to Chinese Language Learning*. *Chinese Language Learning Sciences*. Edited by Xiaofei Lu and Berlin Chen. Singapore: Springer, pp. 196–216.

Yang, Suying, Yueyuan Huang, and Xiuling Cao. 2000. Underuse of temporal markers in Chinese as a second language. *Journal of Chinese Language Teachers Association* 35: 87–116.

Yang, Suying, Yueyuan Huang, and Dejin Sun. 1999. Acquisition of aspect in Chinese as a second language. *Journal of the Chinese Language Teachers Association* 34: 31–54.

Yuan, Fangyuan. 2012. Effects of consciousness-raising on le in L2 Chinese—A pilot study in a classroom setting. *Journal of the Chinese Language Teachers Association* 47: 65–90.

Yuan, Fangyuan. 2019. Effects of task repetition on the interlanguage development of Chinese aspect marker *le*. In *Classroom Research on Chinese as a Second Language*. Edited by Fangyuan Yuan and Shuai Li. New York: Routledge, pp. 77–98.

languages

MDPI

Article

Information Structure and Word Order Preference in Child and Adult Speech of Mandarin Chinese

Jidong Chen [1,*], Bhuvana Narasimhan [2], Angel Chan [3], Wenchun Yang [3] and Shu Yang [1]

[1] Department of Linguistics, California State University, Fresno, CA 93740, USA; sukkie68@mail.fresnostate.edu
[2] Department of Linguistics, University of Colorado, Boulder, CO 80309, USA; bhuvana.narasimhan@colorado.edu
[3] Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China; angel.ws.chan@polyu.edu.hk (A.C.); wenchunchun.yang@connect.polyu.hk (W.Y.)
* Correspondence: jchen@csufresno.edu

Received: 16 February 2020; Accepted: 2 April 2020; Published: 14 April 2020

Abstract: The acquisition of appropriate linguistic markers of information structure (IS), e.g., word order and specific lexical and syntactic constructions, is a rather late development. This study revisits the debate on language-general preferred word order in IS and examines the use of language-specific means to encode IS in Mandarin Chinese. An elicited production study of conjunct noun phrases (NPs) of new and old referents was conducted with native Mandarin-speaking children (N = 24, mean age 4;6) and adults (N = 25, mean age 26). (The age of children is conventionally notated as years;months). The result shows that adults differ significantly from children in preferring the "old-before-new" word order. This corroborates prior findings in other languages (e.g., German, English, Arabic) that adults prefer a language-general "old-before-new" IS, whereas children disprefer or show no preference for that order. Despite different word order preferences, Mandarin-speaking children and adults resemble each other in the lexical and syntactic forms to encode old and new referents: bare NPs dominate the conjunct NPs, and indefinite classifier NPs are used for both the old and the new referents, but when only one classifier phrase is produced, it is predominantly used to refer to the new referents, which suggests children's early sensitivity to language-specific syntactic devices to mark IS.

Keywords: information structure; child language acquisition; Mandarin Chinese; word order

1. Introduction

Children acquire the core components of a language (e.g., phonology, morphology, and syntax) based on experience with the ambient language by the age of four or five years (e.g., Hoff 2009). However, their knowledge of information structure (IS)—adapting the production of language to the appropriate informational needs of the interlocutors and specific speech contexts—tends to lag behind (e.g., Höhle et al. 2016). An important dimension of IS or information packaging (Chafe 1976) involves a distinction between "old" or "given" information (recently activated information, e.g., a referent mentioned in previous discourse) versus "new" information (e.g., a referent introduced for the first time) (Birner and Ward 2006). Bock et al. (2004) suggest that speakers' choice of ordering information that is old versus new in discourse is influenced by conceptual prominence, i.e., which information is activated and accessible at the time of speaking in discourse. However, they also reason that, paradoxically, conceptual prominence could also be associated with new information that involves novelty and change, leading it to be mentioned first.

Previous studies of adult language production show that adult speakers typically order old referents before mentioning new referents when communicating with their interlocutors (Arnold et

al. 2000; Bock and Irwin 1980; Ferreira and Yoshita 2003). Research findings in child language are inconclusive, suggesting that children prefer "old before new" (Stephens 2010), "new before old" (Bates 1976; MacWhinney and Bates 1978), or exhibit no significant ordering preference (MacWhinney and Bates 1978). Research on the acquisition of IS is still comparatively scarce (Höhle et al. 2016), and little is known about when children acquiring different languages develop adult-like use of linguistic devices to encode old versus new information. Recent studies using an experimental paradigm of elicited conjunct NPs suggest an early cognitive or communicative tendency influencing children's production crosslinguistically. Children exhibit a preference for the "new-before-old" order in languages such as German (Narasimhan and Dimroth 2008), Spanish (Ceja Tel Toro et al. 2016), and Arabic (Semsem and Chen 2019), in contrast to adult speakers of these languages who exhibit the opposite preference for "old-before-new" word order. However, an elicited production study of conjunct NPs in English-speaking children (mean age 4;4, age range 3;10–5;1) show that English-speaking children do not show a significant preference for the "new-before-old" word order; however, they are less likely to employ the "old-before-new" word order compared to adults (Chen and Narasimhan 2018). From a psycholinguistic perspective, the age-related differences may be explained in terms of the influence of different facets of conceptual prominence on word order in conjunct NPs: adults prefer to mention accessible, easily retrievable, information first, whereas children lack this preference and may even prefer to highlight novel information first.

This study revisits the preference for the "old-before-new" or "new-before-old" word order in IS and examines how it is manifested in the speech of child and adult speakers of Mandarin Chinese (henceforth Mandarin). If adult Mandarin speakers are guided by a language-general bias stemming from conceptual prominence—i.e., for mentioning old information before new in adult language production—they are predicted to prefer the "old-before-new" word order (e.g., Arnold et al. 2000; Bock and Irwin 1980; Ferreira and Yoshita 2003). Turning to acquisition, if the previously observed preference for the "new-before-old" word order in children is a language-independent bias influencing children's production crosslinguistically, we would expect Mandarin children to exhibit a "new-before-old" preference (as was found in children acquiring German, Spanish, and Arabic). However, if children's ordering preference is also influenced by the language-specific discourse properties of the target language, children acquiring Mandarin may be more similar to their adult counterparts in preferring to use the "old-before-new" word order. Mandarin has a canonical SVO (Subject-Verb-Object) word order, and word order variation is allowed to a certain extent (Li and Thompson 1981). Typologically, it is known as a discourse-prominent language with prevalence of topic-comment structure and a morphologically impoverished language that does not have overt morphological markers for old versus new information (Li and Thompson 1981). Information that is "old" is frequently omitted if retrievable from the speech context, e.g., arguments and adjuncts whose referents are "given" in the discourse-pragmatic context. Syntactic positioning has also been argued to reflect information structure. For example, information focus is typically located in the sentence final position (Xu 2004). Because topic is often correlated with old information and focus is correlated with new information (e.g., Von Stutterheim and Klein 2002), adult Mandarin speakers may be more likely to reserve the sentence-final position for new information and either omit old information or mention the information in sentence-initial position.

The developmental study of Mandarin, therefore, provides a new testing ground for the interplay between language-specific encoding of IS and cognitive or communicative biases for IS in adults and children. The findings will shed light on whether the "old-before-new" preference in adults and "new-before-old" preference in children is a universal pattern or whether information status influences word order differently in speakers of different ages and languages.

2. Materials and Methods

Against the theoretical and empirical background described above, our study specifically explores the nature of age effects on the linguistic encoding of IS, namely, word ordering preferences, by asking:

how do monolingual Mandarin-speaking children and adults order "old" and "new" referents in conjunct NPs? Conjunct NPs (e.g., a book and a flower) were chosen, as they are simple to produce and allow for information status to be manipulated in noun phrases that do not otherwise differ in topicality or semantic or grammatical role.

The specific research questions that we are examining are the following:

1. How are "old" and "new" referents ordered in conjunct NPs in the speech of Mandarin-speaking children and adults?
2. Is the "old-before-new" order a natural preference in adult language crosslinguistically?
3. Is the "new-before-old" preference a cognitive bias in child language, or is it modulated by the possibility for pragmatically driven word order variation in the target language?

2.1. Participants

An elicited production study of conjunct NPs was conducted, following the paradigm adapted from Narasimhan and Dimroth (2008). Two groups of native Mandarin speakers, 25 adults (mean age 26, age range 19–32, 11 females) and 24 children (mean age 4;6, age range 4;0–5;5, 13 females) were recruited and participated in the elicitation task in China.

2.2. Stimuli

The stimuli were composed of a total of 30 trials, including 4 warm-ups, 12 target trials, and 14 filler trials. The trials consisted of colored pictures of commonly encountered inanimate objects presented singly or in pairs on slides on a laptop. The pictures of the object pairs in the 12 target trials were matched in color and size. To avoid potential spatial bias that might affect the ordering of the nouns, the object pairs in all the trials appeared simultaneously and moved randomly across the laptop screen, and the spatial locations of the initial occurrence of the two objects (old versus new) were also counterbalanced.

The 12 target pairs of objects and their Mandarin labels are shown in Table 1. The names of the objects in the target trials (i.e., 24 target nouns) were matched on the number of syllables and frequency of use based on two longitudinal child-caregiver corpora (children's age range: 1;4–3;4), including the Tong corpus (Deng and Yip 2018) and Beijing corpus (Tardif 1996) in the CHILDES database (MacWhinney 2000). The target nouns were also checked against the word list in the Mandarin Early Vocabulary Inventory (Hao et al. 2008) to ensure that they occur as part of the early productive vocabulary of monolingual Mandarin-learning children. The names of the target objects were also controlled for phonological (e.g., syllable weight) and semantic similarities. To control for any effects of the salience of individual objects, the object introduced first in each target pair (i.e., the "old" referent) was counterbalanced across subjects (i.e., object 1 presented first versus object 2 presented first). The target and filler stimuli, the test trials, and the presentation order of the test trials were randomized and counterbalanced into four different orders, and participants were randomly assigned to one of the orders.

Table 1. Labels for target object pairs in stimuli.

	Object Label 1			Object Label 2		
1	书	*shu*	"book"	花	*hua*	"flower"
2	钟	*zhong*	"clock"	碗	*wan*	"bowl"
3	气球	*qiqiu*	"balloon"	蜡笔	*labi*	"crayon"
4	杯子	*beizi*	"cup"	鞋子	*xiezi*	"shoe"
5	钥匙	*yaoshi*	"key"	扣子	*kouzi*	"button"
6	帽子	*maozi*	"hat"	鸡蛋	*jidan*	"egg"
7	饼干	*binggan*	"cookie"	瓶子	*pingzi*	"bottle"
8	树	*shu*	"tree"	床	*chuang*	"bed"
9	桌子	*zhuozi*	"table"	勺子	*shaozi*	"spoon"
10	汽车	*qiche*	"car"	椅子	*yizi*	"chair"
11	苹果	*pingguo*	"apple"	铅笔	*qianbi*	"pencil"
12	盘子	*panzi*	"plate"	衬衫	*chenshan*	"shirt"

Labels for the target objects are shown in Chinese characters and Pinyin, the official Romanized transcription of Chinese characters, followed by English translations in quotation marks.

2.3. Procedure

Each participant watched the stimuli one by one in a slide show on a laptop individually in a quiet room with an experimenter. The experimenter played the slide show and thus was able to see the laptop screen. Within each of the 12 target pairs, one of the objects was presented first; the participant had to name the object that he/she had seen; and the experimenter repeated once the name of the object that the participant provided (the "old" referent). Then, the second object (the "new" referent) appeared simultaneously with the first object in the following slide. The participant was asked what he/she had seen on the screen. With the child participants, this procedure was slightly adapted in a child-friendly manner to keep them engaged. The experimenter introduced a stuffed animal at the beginning of the task, a toy teddy bear, who could not see the slide and wanted to know what the child had seen on the screen. Each child was first invited to make friends with the teddy bear by patting it. Then, she or he (henceforth "she") was asked if she would like to help the teddy bear learn what she had seen. All the children agreed. All the elicitation sessions were audio recorded.[1]

2.4. Data Treatment

The participants' responses to the target trials were transcribed in simplified Chinese characters following the Codes for the Human Analysis of Transcripts (CHAT) convention (MacWhinney 2000) and coded for the ordering of the referents: (1) n/o: new referent before old; (2) o/n: old referent before new; and (3) missing responses. The total number of target responses was 551, including 300 (12 target trials × 25 adults) from the adults and 251 from the children, excluding 37 missing responses from the 288 expected responses (12 target trials × 24 children) due to PowerPoint failure during the experiment.

[1] Although the children were asked to provide information about what they saw on the computer screen to the teddy bear toy who could not see the screen, they did share visual access to the screen with the experimenter. Informal observations of the children during the experiment indicate that most children tended to look at the experimenter when describing what they saw on the screen and did not often pay attention to, or interact with, the teddy bear. The same procedure with a toy teddy bear was employed with the English-learning children (Chen and Narasimhan 2018). But in the study of German-learning children (Narasimhan and Dimroth 2008), upon which the current study is based, the interactional situation differed from the present study and the study with children acquiring English. In that study, the child and adult participants also shared access to the screen with one experimenter as in the current study. But they interacted mainly with a second experimenter who could not see the screen and who engaged the participant by matching the participant's description (e.g., apple and spoon) with the corresponding picture (from a set of pictures they had available). Despite these differences, the information status of the referents was similar in all three studies in terms of newness in the discourse: the first object of the paired objects in each target trial is discourse-old (labeled and repeated prior to presentation of the target trial containing the pair of objects) and the second object is discourse-new for the participants as well as the experimenters with whom the participants interacted.

3. Results

3.1. Word Order Preference

Figure 1 presents the mean proportions of the "old-before-new" and the "new-before-old" word orders in the adult and the child speech, respectively. The adults showed an overall preference for the "old-before-new" word order: 82.33% "old-before-new" responses in contrast to 17.67% "new-before-old". The children, on the other hand, showed a reduced preference for the "old-before-new" order (mean proportion 55.07%) and a much higher use of the "new-before-old" order (mean proportion 44.93%) than the adults.

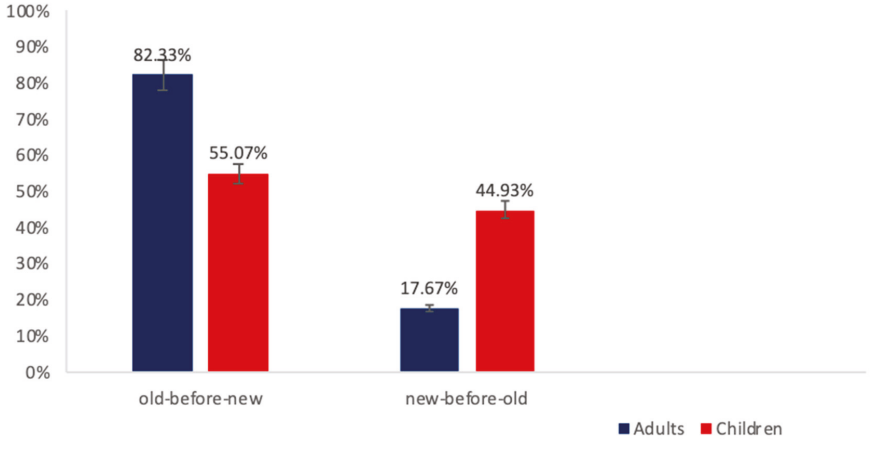


Figure 1. Mean proportions of the "old-before-new" and the "new-before-old" responses in Mandarin child and adult speech.

The adults also showed less variation in word order preference than the children. As shown in Figure 2, the majority, 88% (22 out of 25) of the adults, preferred the "old-before-new" order (with percentages of "old-before-new" response rates ranging from 67% to 100%), and 68% (17 out of 25) of them predominantly used the "old-before-new" (with percentages of "old-before-new" rates ranging between 83% and 100%). In contrast, the children exhibited much greater variation in their responses. As shown in Figure 3, children's percentages of "old-before-new" responses varied between 25% and 100%. Further, 58% (14 out of 24) of the children exhibited a low preference for the "old-before-new" order (with percentages of "old-before-new" responses ranging between 25% and 44%), 29% (7 out of 24) of the children preferred the "old-before-new" order (with percentages of "old-before-new" responses ranging between 78% and 100%), and 13% (3 out of 24) of the children were at chance level.

Even given the individual variation among children, there may be some developmental trends. For instance, older children may be more likely to employ the adult-like "old-before-new" word order compared with younger children. A further examination of the results shows only partial support for this possibility. The seven children who preferred the "old-before-new" order (i.e., children 18–24 in Figure 3) were relatively older, mean age 4;7 (age range 4;4–5;4), and the youngest five children (four 4;0-year-olds and one 4;1-year-old) showed a low preference (mean proportion 38.9%) for the "old-before-new" word order. However, some of the older children (e.g., children 3, 5, 9, 13, and 14, age 5;0 and above) did not prefer the "old-before-new" word order, whereas some younger children (e.g., children 19, age 4;4) preferred that word order (see Figure 3).

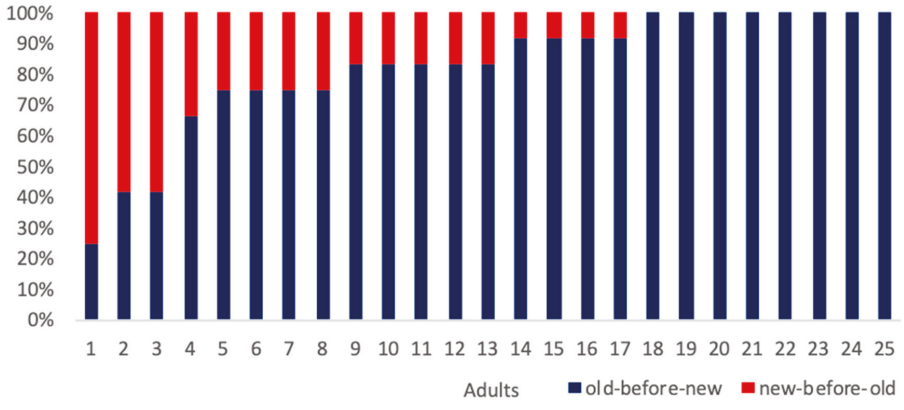


Figure 2. Mean proportions of "old-before-new" and "new-before-old" responses by adult.

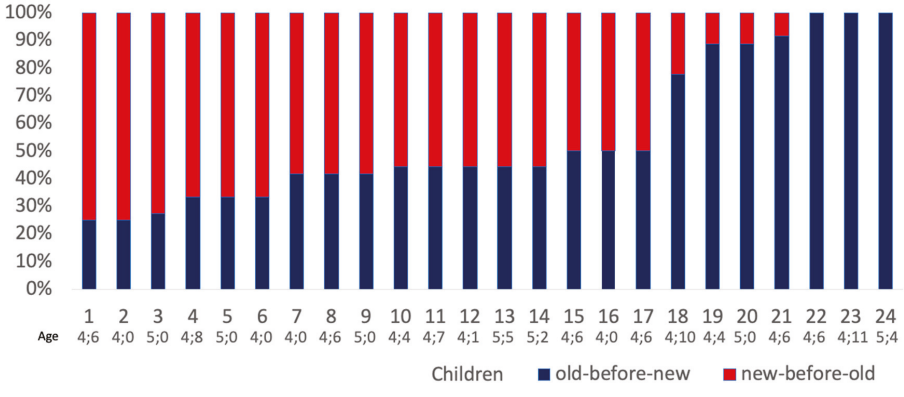


Figure 3. Mean proportions of "old-before-new" and "new-before-old" responses by child.

The descriptive results above reveal a strong and consistent preference for the "old-before-new" order in Mandarin adult speech, but a much weaker and varied preference in the child speech. In order to examine whether the children and the adults differed significantly in their responses, a logistic regression analysis was conducted with age as the predictor variable and word order as the outcome variable. The results show a significant effect of age: children were significantly less likely to use the "old-before-new" word order than were adults (82.33% versus 55.07%, $\beta = 1.371$, $p < 0.000$; see Table 2). A chi-square test further reveals that the children did not show a preference for the "new-before-old" either because the preference for either word order does not differ from chance (X^2 (20, N = 251) = 26.455, $p = 0.151$).

Table 2. Effects of age on the choice of "old-new" versus "new-old" order in children and adults.

	Estimate	Std. Error	Z Value	*p* Value
(intercept)	0.354	0.168	4.469	0.035 *
Age: child	1.371	0.198	47.806	0.000 ***

(Asterisk is used conventionally to indicate degree of significance: * means $p \leq 0.05$; ** means $p \leq 0.01$; and *** means $p \leq 0.001$.).

As our study is essentially a free naming/description task, both the adults and the children exhibited variation in their responses from the expected target forms (e.g., 一朵花和一本书, *yi duo hua*

he yi ben shu, "one CLF flower and one CLF book" (CLF = classifier) or 书和紫色的花, *shu he zise de hua*, "book and purple flower" instead of the target noun forms 书和花, *shu he hua*, "book and flower"). Prior studies show that word order is influenced by factors other than information status, including the "weight" or length of noun phrases. In many languages, noun phrases that are longer (e.g., they contain more words or syllables) tend to be placed last, e.g., the "heavy NP shift" in English (Arnold et al. 2000). Here, we performed a post-hoc analysis to investigate whether the second nominal in the conjunct NP tends to be heavier than the first nominal, and whether information status interacts with the weight of noun phrases in influencing ordering preferences. To address this question, the number of the syllables for each of the NPs in the conjunct NPs were extracted by the Computerized Language ANalysis (CLAN) program (MacWhinney 2000) and categorized by the combinations of syllable numbers in the first and the second NPs (e.g., syllable1 + 1, syllable2 + 1, etc.).

The results show that the number of syllables of each nominal in the conjunct NPs range widely from 1 to as high as 11. However, 84% of the adults' and 85% of the children's conjunct NPs were composed of words with only one or two syllables, and complex NPs or multisyllabic nouns with more than 3 syllables were infrequent in both the child and the adult speech. The comparison of the weight of the first and the second nominals shows that the majority of the nominals in the conjunct NPs had the same weight (i.e., contain the same number of syllables) for both the children (58.17%) and the adults (68.33%). Further, two-syllable words were the most frequent for the children (69.23%) and the adults (61.64%) in the set of conjunct NPs with equally weighed noun phrases. This pattern suggests that the length of the referents does not affect the ordering of the nouns or NPs in the majority of the conjunct NPs.

We further looked into the conjunct NPs that have nominals with unequal weights. Figure 4 summarizes the mean proportions of conjunct NPs with a longer first nominal (HL = heavy-light) and a longer second nominal (LH = light-heavy) in the "old-before-new" versus "new-before-old" word orders in the child and the adult speech, respectively. It shows that the adults used HL and LH similarly frequently for the "old-before-new" (81.82% versus 80.39%) and the "new-before-old" (18.18% versus 19.61%) orders. A similar pattern was found in the child speech: 48.15% HL and 58.82% LH in the "old-before-new" order, and 51.85% HL and 41.18% LH in the "new-before-old" order. To summarize, the weight of the nominals in the conjunct NPs does not appear to affect the ordering of the old and the new referents in the child and the adult speech, as indicated by (1) the dominance of equal-weight nouns or NPs and (2) the similar frequency distribution of heavier or lighter first nouns or NPs in the "old-before-new" and the "new-before-old" orders (see Figure 4).

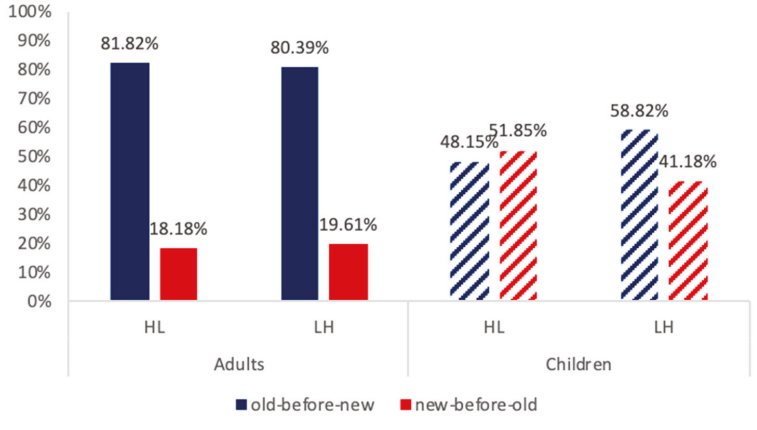


Figure 4. Mean proportions of heavy-light (HL) and LH (light-heavy) conjunct NPs by information structure (IS) in child and adult speech. (HL = first nominal with a heavier weight than the second noun or NP; LH = second nominal with a heavier weight than the first noun or NP).

3.2. Lexical and Syntactic Features of the Conjunct NPs

We further analyzed the lexical and syntactic features of the conjunct NPs to see if adults and children use lexical or syntactic means other than word order to distinguish between old and new referents in the conjunct NPs. All the nominals for the 12 target trials (300 from the adults and 251 from the children) were further coded for the response types based on the actual forms produced in the adult and the child speech: (1) bare nouns (both nouns are bare, e.g., 书和花, *shu he hua*, "book and flower"), (2) classifier NP (at least one of the two NPs involve a classifier, e.g., 一本书和花 *yi ben shu he hua*, "one CLF book and flower"), (3) NP with a modifier (at least one of the NPs involves a modifier such as an adjective, 红书和花, *hongshu he hua*, "red book and flower"), and (4) nominalized verb phrase (VP) using the nominalizer (NOM) 的 *de* "de" (e.g., 喝水的 *he shui de*, "drink-water-NOM" (glass for drinking water)). Classifier phrases were placed in a separate category, as they represent a Mandarin-specific syntactic construction for nominal referents. All the classifier NPs contain a numerical (i.e., one) followed by a classifier and a noun (e.g., 一本书, *yi ben shu*, "one CLF book", 一朵花, *yi duo hua*, "one CLF flower") in the child and the adult speech. The weight/length of all the NPs were also coded and measured by number of syllables. The CLAN program (MacWhinney 2000) was used to extract the mean length of utterance (MLU) of the conjunct NPs and the number of different response types.

The children's conjunct NPs (MLU = 4.68, SD = 2.43) were on average about one morpheme/character longer than those of the adults (MLU = 3.78, SD = 1.1), but an independent-samples t-test shows that the difference was not significant ($t(47) = 1.656$, $p = 0.1$). Figure 5 shows the mean proportions of different types of the conjunct NPs. Both the children and the adults were similar in the overall frequency in the use of different types of NPs: bare nouns were dominant in the child and the adult speech (64.14% versus 73.67%), followed by classifier NPs (27.89% versus 17.33%) and modifier NPs (7.18% versus 8.66%), whereas the nominalized NPs were minimal (0.8% in the child speech). The dominance of bare NPs suggests that neither the adults nor the children tended to mark the old and the new referents differentially in nominal forms.

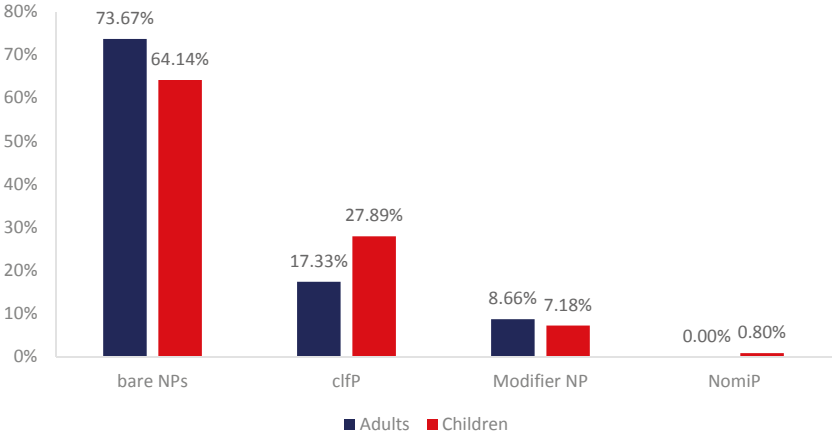


Figure 5. Mean proportions of the types of conjunct NPs in child and adult speech. (clfp = classifier phrase; NomiP = nominalized verb phrase).

A further examination of the classifier NPs also reveals remarkable similarities between the children and the adults. The general classifier 个 *ge* was used mostly frequently among all the different classifiers for both the children (75%) and the adults (55%). A variety of sortal classifiers were used by both the children (8 types) and the adults (13 types). The majority of the conjunct NPs with classifiers contained two classifier NPs for both the children (61.34%) and the adults (63.46%), which suggests

that a classifier was not used distinctively for IS in the conjunct NPs with two classifier phrases (see Figure 6). However, when the conjunct NP contains only one classifier phrase, both the adults and the children tended to use it to refer to the new referent (31.47% and 30.71%) in contrast to the use on the old referents (7.14% and 7.69%), indicated by the second and the third pairs of bars in Figure 6. Such a pattern suggests a very subtle differentiated use of indefinite classifier phrases to indicate new referents.

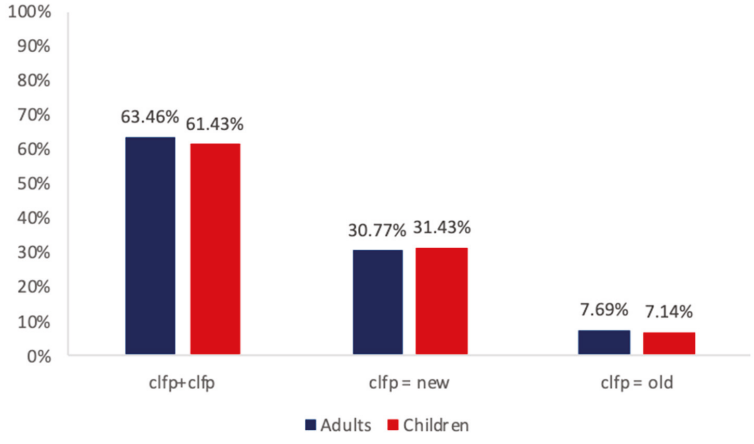


Figure 6. Mean proportions classifier NPs in conjunct NPs by IS in child and adult speech. (clfp = classifier phrase).

4. Discussion

The current study investigated how labels for "old" and "new" referents are ordered in conjunct NPs in the speech of Mandarin-speaking adults and children: "old-before-new" or "new-before-old". We conjectured that the two types of patterns reflect different aspects of the influence of conceptual prominence on word order, accessibility, and novelty. We asked if prior research demonstrating an "old-before-new" preference in adult language and a "new-before-old" preference in child language are language-independent preferences crosslinguistically, or whether these patterns could be modulated by the role of discourse-pragmatically motivated word order variations in the target language.

Our results offer further evidence from Mandarin for the robust "old-before-new" word order preference in adult language: Mandarin-speaking adults produced the "old-before-new" order dominantly and consistently in the conjunct NPs, congruent with the "old-before-new" preference documented in adult speakers of German, English, and Arabic using a similar task of elicited production of conjunct NPs (Chen and Narasimhan 2018; De Ruiter et al. 2018; Narasimhan and Dimroth 2008; Semsem and Chen 2019).[2]

However, the "old-before-new" preference is not a global preference; it is modulated by age. Mandarin-speaking four-year-olds differed from their adult counterparts in exhibiting no such preference. Nor did they employ the "new-before-old" order at rates significantly above chance, similar to the findings in children acquiring English (Chen and Narasimhan 2018), but unlike the patterns

[2] The study of Spanish speakers, using a similar task of elicited production of conjunct NPs (Ceja Tel Toro et al. 2016), had a wide age range (from 31 to 72 years) of adult bilingual Spanish speakers in a small sample (12 speakers in total), who were living in the USA with varied length of residence (3 to 46 years) and varied proficiency in English. The younger Spanish adults (31 to 40 years) showed a higher mean proportion of the "old-before-new" word order, but the older adult Spanish speakers (42–70 years) did not. Individual variation was also found among the older and the younger speakers. It is possible that variation within adult age and bilingualism plays some role in the choice of word order for IS. It is thus unclear if monolingual Spanish adult speakers may also show a preference for the "old-before-new" word order in conjunct NPs.

found in children learning German, Spanish, or Arabic who exhibit a significant "new-before-old" preference (Ceja Tel Toro et al. 2016; Narasimhan and Dimroth 2008; Semsem and Chen 2019). The crosslinguistic differences may arise from multiple sources.

We conjecture that children acquiring any language are likely to find novel referents more salient than old referents. However, it is possible that children acquiring a relatively rigid word order language, such as English (Callies 2009), are less likely to reorder noun phrases based on their information status, even though adult speakers of English are willing to do so when producing conjunct NPs. On the other hand, children acquiring Mandarin are exposed to grammatical patterns that are frequently motivated by discourse-pragmatic considerations in constructions other than conjunct NPs alone. In particular, they may be frequently exposed to the use of the "old-before-new" order in the input. Even though no studies have analyzed the distribution of the "old-before-new" word order at sentence and discourse levels in naturalistic longitudinal children-directed speech in Mandarin, the topic-prominent property of Mandarin predicts that children are likely to hear the "old-before-new" order frequently in the input. Subject NPs in Mandarin are usually definite, referring to old information, and object position tends to be reserved for an indefinite NP that is new information (Hole 2012). Topics (typically old information) tend to occur in sentence-initial positions (Li and Thompson 1981), and focused elements (typically new information) are placed in sentence final position (Xu 2004). If a cognitive bias to produce the "new-before-old" order is in competition with an input-driven "old-before-new" preference, children may produce both patterns frequently, giving rise to the overall non-significant patterns in children's production in the present study. Although German, Spanish, and Arabic are also languages with relatively less rigid word order, pragmatically driven word order variation ("old-before-new") may be a less frequent phenomenon in these languages relative to Mandarin. Hence, although adults produce the "old-before-new" pattern, children acquiring these languages may be more strongly influenced by the cognitive salience of novel information than pragmatically based word order patterns in the input compared to children acquiring Mandarin.

The absence of a preference for the "new-before-old" order in English and Mandarin child speech may also result from methodological differences across studies that relate to the communicative situation in which the experimental task was performed. In the present study, children interacted mainly with an experimenter, even though they were instructed to address their responses to a toy teddy bear who could not see the screen. However, in the study by Narasimhan and Dimroth (2008), children acquiring German addressed a second experimenter who could not see the screen during the experiment and had to select a picture that matched the description of the experimental stimuli produced by the child. The study of children acquiring Arabic (Semsem and Chen 2019) was similar to the study of the children acquiring German in that it involved an adult confederate who had to repeat what the child described. However, no picture-matching was employed as was the case in the German study. Nevertheless Arabic-speaking children preferred the "new-before-old" order just like the German-speaking children. In both studies, the children were engaged in a more communicative interaction as compared with the procedure used in the English study (Chen and Narasimhan 2018) and the current study, where children simply described what they saw on the computer screen to the experimenter (or a teddy bear). This methodological difference (i.e., less communicative contexts) may have contributed to children's sensitivity to the informational needs of the addressee and thus the less frequent production of the "new-before-old" order.

Individual variation may be another confounding factor. As our results show (cf. Figure 3), the mean proportion of the "new-before-old" order is 44.93%, ranging from 0% to 75%; 25% of the children exhibited a preference for the "new-before-old" order (67%–75% of their responses), 30% of the children exhibited a preference for the "old-before-new" order (78%–100% of their responses), and 45% of the children were at chance level. Age variation among the sampled children may have also contributed to the results. Our results show an emerging developmental trend in Mandarin children from age 4;0 to 5;5. The younger children (4;0–4;1) tended to use the "new-before-old" order more frequently, and the older children (4;10–5;5) employed the "old-before-new" order more frequently but with

considerable individual variation. A clearer developmental trajectory has been found in the study of Arabic speakers (Semsem and Chen 2019), where two groups of children (four- and six-year-olds) were compared: there was a significant increase in the use of the "old-before-new" order in the speech of the six-year-olds (mean age 6;4) than the four-year-olds (mean age 4;7), even though the six-year-olds still differed from the adults in using significantly less "old-before-new" word order. Dimroth and Narasimhan (2012) found that German-learning children exhibited adult-like word order preference by around nine years of age whereas five-year-olds still patterned like three-year-olds in preferring the "new-before-old" order (Narasimhan and Dimroth 2008). Hence the shift towards the "old-before-new" pattern occurs sometime between five and nine years of age in children acquiring German. These developmental trajectories suggest that it may take time for children to develop adult-like word order strategy to adapt to the IS needs.

Our study also reveals remarkable similarities between Mandarin children and adults in using language-specific lexical and syntactic means to express old and new referents in conjunct NPs. Bare noun forms dominate the production of both the old and the new referents. However, when an indefinite classifier phrase is used in the conjunct NP, it is typically used to refer to the new referent. Thus, young Mandarin-speaking children, similar to adults, use indefinite classifier phrases to mark IS in a subtle manner. Mandarin-speaking children also resemble adults in producing nouns or NPs with similar weight in the majority of their conjunct NPs. Even when the nouns or NPs in the conjunct NPs varied in weight, both the children and the adults used heavy or light nouns or NPs similarly as the first or the second referent in the "old-before-new" and the "new-before-old" orders.

5. Conclusions

This study revisits the debate on language-independent preferred word order in IS and the use of language-specific means to encode IS in Mandarin. Our results from the elicited production of conjunct NPs of new and old referents show that Mandarin-speaking adults differ significantly from children in preferring the "old-before-new" word order. This finding corroborates prior research of monolingual adult speakers of English, German, and Arabic, supporting that adults prefer a language-general "old-before-new" IS, whereas children (e.g., learning German, Spanish, or Arabic) disprefer or show no preference for that order (e.g., in English or Mandarin). The difference between children and adults in all the languages studied thus far nicely captures the paradoxical role of conceptual prominence in influencing speakers' choice of word order for IS as discussed in Bock et al. (2004). Our results reveal that adults are more likely to place first the old/given referent that is activated and accessible at the time of speaking, whereas children tend not to be similarly motivated, preferring (in some languages) to place first the new referent that involves novelty and change. Children and adults thus exhibit different biases in arranging the order of new versus old information for IS, at least in conjunct NPs. The preference for the given-before-new word order has been argued to hold true crosslinguistically to account for word variation for IS (Neeleman and Koot 2016), and it is ultimately "an effect of a general cognitive principle according to which integration of new information is easier if framed within old information" (Neeleman and Koot 2016, p. 401, see also Clark and Haviland 1977). Young children (around the age of 4;6) are therefore still in the process of developing the discourse-pragmatic sensitivity and competence to facilitate the integration of new information in an adult-like manner. This development may be gradual and subject to extensive individual variation (e.g., age of acquisition, gender, influence of a second language, and other potential random variables). Further, it may be also sensitive to the communicative contexts in which utterances are produced (e.g., in terms of shared information between the speaker and addressee) as well as language-specific patterns in the input: the lack of a significant preference to order "new" information first in Mandarin-learning children may arise from exposure to relatively frequent "old-before-new" patterns in the ambient language. Despite different word order preferences, Mandarin-speaking children and adults resemble each other in their lexical and syntactic forms to encode old and new referents: bare NPs dominate the conjunct NPs, and indefinite classifier NPs are used for both the old and new referents, but when only one

classifier phrase is produced, the classifier NP is predominantly used to refer to the new referents, which suggests children's early sensitivity to language-specific syntactic devices to mark IS. Future research should examine large samples of Mandarin-learning children at different ages to explore how individual differences (e.g., age, gender, lexical and syntactic proficiency, etc.) and communicative contexts may affect the use of word order to mark IS, and when Mandarin-learning children become adult-like in adapting word order for the need of IS.

Author Contributions: Conceptualization, J.C. and B.N.; Formal analysis, J.C. and B.N.; Methodology, J.C., B.N., A.C., W.Y., and S.Y.; Writing—original draft, J.C.; Writing—review & editing, J.C. and B.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Research, Scholarship, and Creativity Award to the first author at California State University, Fresno.

Acknowledgments: We thank all the adult and child participants for their participation in the experiments.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Ethics Statement: All participants gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Committee on the Protection of Human Subjects of California State University, Fresno (Project ID #786).

References

Arnold, Jennifer E., Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76: 28–55. [CrossRef]

Bates, Elizabeth. 1976. *Language and Context: The Acquisition of Pragmatics*. Cambridge: Academic Press.

Birner, Betty J., and Gregory Ward. 2006. Information structure. In *The Handbook of English Linguistics*. Edited by B. Aarts and A. McMahon. Malden: Blackwell Publishing, pp. 291–317.

Bock, Kathryn J., and David E. Irwin. 1980. Syntactic effects of information availability in sentence production. *Journal of Verbal Memory and Verbal Behavior* 19: 467–84. [CrossRef]

Bock, Kathryn, David E. Irwin, and Douglas J. Davidson. 2004. Putting first things first. In *The Interface of Language, Vision, and Action: Eye Movements and the Visual World*. New York: Psychology Press, pp. 249–78.

Callies, Marcus. 2009. *Information Highlighting in Advanced Learner English*. Amsterdam: John Benjamins Publishing.

Ceja Tel Toro, Pablo, Jidong Chen, and Bhuvana Narasimhan. 2016. Information structure in bilingual Spanish-English child speech. Paper presented at the 2016 International Workshop on Language Processing and Production, San Diego, CA, USA, June 15–16.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In *Subject and Topic*. Edited by Charles N. Li. New York: Academic Press, pp. 27–55.

Chen, Jidong, and Bhuvana Narasimhan. 2018. Information structure and ordering preferences in child and adult speech in English. In *The Proceedings of the 42nd Boston University Conference on Language Development*. Edited by A. B. Bertolini and M. J. Kaplan. Boston: Cascadilla Press, pp. 131–39.

Clark, Herbert H., and Susan E. Haviland. 1977. Comprehension and the given-new contract. *Discourse Production and Comprehension. Discourse Processes: Advances in Research and Theory* 1: 1–40.

De Ruiter, L., Bhuvana Narasimhan, Jidong Chen, and Jonah Lack. 2018. Children's use of prosody and word order to indicate information status in English phrasal conjuncts. *Proceedings of the Linguistic Society of America* 3: 40. [CrossRef]

Deng, Xiangjun, and Virginia Yip. 2018. A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics* 46: 69–92.

Dimroth, Christine, and Bhuvana Narasimhan. 2012. The development of linear ordering preferences in child language: The influence of accessibility and topicality. *Language Acquisition* 19: 312–23. [CrossRef]

Ferreira, Victor S., and Hiromi Yoshita. 2003. Given-new ordering effects on the production of scrambled sentences in Japanese. *Journal of Psycholinguistic Research* 32: 669–92. [CrossRef] [PubMed]

Hao, Meiling, Hua Shu, Ailing Xing, and Ping Li. 2008. Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods* 40: 728–33. [CrossRef] [PubMed]

Hoff, Erika. 2009. *Language Development*. Belmont: Wadsworth.

Höhle, Barbara, Frauke Berger, and Antje Sauermann. 2016. Information structure in first language acquisition. In *The Oxford Handbook of Information Structure*. Edited by C. Féry and S. Ishihara. Oxford: Oxford University Press, pp. 562–80.

Hole, Daniel. 2012. The information structure of Chinese. In *The Expression of Information Structure*. Edited by M. Krifka and R. Musan. Berlin: De Gruyter Mouton, pp. 45–70.

Li, Charles, and Sandra Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Los Angeles: University of California Press.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum.

MacWhinney, Brian, and Elizabeth Bates. 1978. Sentential devices for conveying givenness and newness: A crosscultural developmental study. *Journal of Verbal Learning and Verbal Behavior* 17: 539–58. [CrossRef]

Narasimhan, Bhuvana, and Christine Dimroth. 2008. Word order and information status in child language. *Cognition* 107: 317–29. [CrossRef] [PubMed]

Neeleman, Ad, and Hans Van De Koot. 2016. Word order and information structure. In *The Oxford Handbook of Information Structure*. Edited by C. Féry and S. Ishihara. Oxford: Oxford University Press, pp. 383–401.

Semsem, Mashael, and Jidong Chen. 2019. The use of word order to mark information status in adult and child Arabic. In *The Proceedings of the 30th Western Conference on Linguistics*. Edited by T. Driscoll. Fresno: Department of Linguistics, California State University, pp. 173–78.

Stephens, Nola Marie. 2010. Given-Before-New: The Effects of Discourse on Argument Structure in Early Child Language. Ph.D. dissertation, Stanford University, Stanford, CA, USA.

Tardif, Twila. 1996. Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology* 32: 492–504. [CrossRef]

Von Stutterheim, Christiane, and Wolfgang Klein. 2002. Quaestio and L-perspectivation. In *Perspective and Perspectivation in Discourse*. Edited by C. F. Graumann and W. Kallmeyer. Amsterdam: John Benjamins, pp. 59–88.

Xu, Liejiong. 2004. Manifestation of informational focus. *Lingua* 114: 277–99. [CrossRef]

languages

MDPI

Article

Metadiscourse, Cohesion, and Engagement in L2 Written Discourse

Jianling Liao

School of International Letters and Cultures, Arizona State University, Tempe, AZ 85287-0202, USA; jianling.liao@asu.edu

Received: 27 February 2020; Accepted: 21 May 2020; Published: 5 June 2020

Abstract: The current study examines how L2 Chinese writers at different proficiencies employed various metadiscourse devices to shape their written descriptive discourse and also whether various metadiscourse features may distinguish levels of writing proficiency. The study also looks at how L2 learners' use of metadiscourse devices is related to their linguistic performances in descriptive writing. The findings revealed differential metadiscourse use by learners at different proficiencies on local, global, and textual organizational dimensions. For instance, compared to low-proficiency writers, more proficient writers used significantly more conditional/hypothetical markers, frame markers, and engagement markers. Multiple metadiscourse features also demonstrated significant positive and negative correlations with each other, suggesting patterns of decreases and increases in the use of particular organizational features. Several metadiscourse features characteristic of more advanced writers also displayed positive relationships with linguistic features.

Keywords: L2 Chinese; descriptive writing; metadiscourse; cohesion; local; global; text; interactive; interactional

1. Introduction

Writing in an L2 involves not only an effort to monitor linguistic quality, such as linguistic accuracy or complexity, but also an effort to make metadiscourse choices that will result in cohesive written discourse. An examination of L2 writers' metadiscourse performances will allow for a fuller understanding of L2 writing skills, in terms of how learners allocate their cognitive resources to different areas of writing and how successful they may be in each specific area.

Currently, studies have examined organizational quality in L2 texts mainly by analyzing cohesion and coherence (e.g., Chiang 2003; Crossley et al. 2016a; Ferris 1994; Guo et al. 2013; Harman 2013; Jafarpur 1991; Kormos 2011; Liu and Braine 2005; Yang and Sun 2012). It has been argued that using more logical operators and cohesive devices, such as metadiscourse markers, semantic repetitions, and co-referentiality, will contribute to a more cohesive text (Bardovi-Harlig 1990; Chen and Baker 2016; Connor 1990; Crossley et al. 2016a; Ferris 1994; Guo et al. 2013; Halliday and Hasan 1976; Reid 1992; Yang and Sun 2012). Studies have also investigated the interpersonal dimensions of L2 writing by examining authorial identity and engagement with the reader. It is claimed that the presence of devices that express authorial voice and involve readers will enhance the effectiveness of a text (e.g., Hyland 2005; Lee and Deakin 2016; Zhao 2013).

Despite increased empirical understanding of L2 textual organizational performances, overall knowledge of cohesion and other aspects of textual organization in L2 texts is still limited (Crossley et al. 2016a). For instance, it remains unclear what types of metadiscourse devices L2 writers at different proficiencies may apply to shape their writing on local, global, and text levels. How different types of metadiscourse devices may work together to affect the organizational quality of an L2 written text is also understudied. Furthermore, L2 written organizational features have often been researched in isolation from other aspects of writing. Their relationship to such areas such as

linguistic accuracy or complexity still needs to be explored to allow a more complete picture of L2 writing performance and development. Additionally, current studies have mainly examined writing in English as a second or foreign language (ESL/EFL). We still need to understand textual organizations in other L2s. The current study attempts to address these research gaps by investigating how L2 Chinese writers at different proficiencies deploy metadiscourse devices to form text dynamics and how textual organizational features interconnect with linguistic features in descriptive writing.

2. Literature Review

The literature is surveyed in three areas to provide relevant background on (a) how L2 written organizational quality is currently theorized; (b) how L2 written organizational performances are operationalized; and (c) how L2 learners' textual organizational skills develop.

2.1. Organizational Quality in L2 Texts

L2 writing researchers have investigated textual organizational features in two different yet related dimensions: text structure and interpersonal engagement (with the reader). Text structure is often characterized by two frequently researched textual organizational constructs: cohesion and coherence (e.g., Chiang 2003; Crossley et al. 2016a; Ferris 1994; Guo et al. 2013; Harman 2013; Jafarpur 1991; Kormos 2011; Liu and Braine 2005; Yang and Sun 2012). Although definitions may vary, cohesion in general refers to making connections between ideas for the creation of a coherent and comprehensible discourse (Halliday and Hasan 1976). The cohesiveness or coherence of a text concerns not only whether the preceding and incoming discourses are appropriately linked to advance meaning, but also whether the presented meaning representation may be effectively understood by the reader. Researchers have argued that higher quality writing displays stronger textual cohesion and coherence (Connor 1990; Crossley et al. 2016a; Ferris 1994; Yang and Sun 2012).

To a certain degree, cohesion and coherence appear to be a pair of related traits of textual organizational quality from the writers' and the readers' perspectives, respectively (Crossley et al. 2016a, 2016b; McNamara et al. 1996). From the writer's perspective, cohesion involves the writer's intention to create a text that flows logically; from the reader's side, coherence concerns whether a text is perceived as flowing effectively. Thus, the interpretation of either cohesion or coherence may involve a certain level of subjectivity, whether from the writer or from the reader. An evaluation of cohesion or coherence in an L2 text may also involve an additional level of addressing the possible transfer effects from the writer's L1. A piece of text deemed cohesive in the learner's L1 may be incoherent in the L2, due to possibly distinct rhetorical norms observed in the two languages.

The cohesive ties in a text can be explicit or implicit. Explicit cohesion markers often refer to logical connectives, such as conjunctions, adverbs, or lexical bundles (Chen and Baker 2016; Crossley and McNamara 2012; Guo et al. 2013; Yang 2013). Logical connectives can serve a useful role in terms of creating explicit links and relations between the ideas in a text, which have also been classified into different logical categories by researchers, further explained in the next section. Less explicit cohesive devices include global cohesion features such as lexical, argument, or semantic overlap and co-referentiality (Halliday and Hasan 1976). Nowadays, global cohesive features are often analyzed using computerized programs, especially in ESL/EFL studies (e.g., Crossley et al. 2016a; Crossley and McNamara 2012; Crossley et al. 2011; Guo et al. 2013; Kormos 2011). For instance, through computational tools, latent semantic analysis computes sentence-to-sentence conceptual similarity in a text, by examining meaning overlap between explicit words or words that are implicitly related in meaning (Graesser et al. 2004; Guo et al. 2013; Mazgutova and Kormos 2015).

Hyland (2005), however, argued that there are limitations for observing metadiscourse usages without considering the interaction between the text and the reader. He proposed an interpersonal framework of metadiscourse and posited that one essential purpose for the writer to employ metadiscourse devices is to guide the reader's understanding of the text towards his or her preferred interpretations. Hyland further categorized metadiscourse devices into two taxonomies: interactive

and interactional. Interactive metadiscourse realizes functions similar to cohesion conceptualization, but with a stronger focus on the consequential interpretations that may be made available to the reader. Second, Hyland argued for the need to examine interactional metadiscourse devices in a text, through which the writer brings in authorial voice and engages the reader. How metadiscourse features are specifically operationalized is elaborated next.

2.2. Textual Organizational Devices in L2 Texts

Researchers have proposed various frameworks to operationalize written organizational performances. Crossley et al. (2016a, 2016b) categorized cohesive indices into local, global, and text levels, to allow for a more fine-grained understanding of text cohesion. Local cohesive devices refer to the connectives within/between clauses/sentences. The quantity of preposition usages has also been evaluated to understand intra-clausal cohesion (Crossley et al. 2016a; Reid 1992; Smith and Frawley 1983). As mentioned earlier, global and text cohesion devices tend to be more implicit. Global cohesive devices include connectives between paragraphs or larger chunks of texts, as well as lexical and semantic overlap between the paragraphs in a text (Guo et al. 2013; Halliday and Hasan 1976; Li 2014). Text cohesion concerns cohesiveness across the text and is often assessed through features such as proportion of given/new information (e.g., pronoun/noun ratio, pronoun density), lexical repetitions, or lexical diversity (Crossley et al. 2016a; Kyle and Crossley 2017; Reid 1992).

As discussed earlier, Hyland's (2005) interpersonal analysis framework divides metadiscourse into interactive and interactional devices. Interactive devices build links between ideas in line with the writer's intended interpretation for the reader. According to Hyland, such connectors may include the use of transitional markers, which indicate relations between clauses (e.g., addition, adversative); frame markers, which signal discourse acts, such as sequencers, stage labels, announcements of goals, and topic shifters; endophoric markers, which provide references to information in other parts of the text; evidentials, which provide citations within a community-based literature; and code glosses, which provide reformulations and exemplifications. In contrast, interactional metadiscourse markers serve interpersonal functions, which include hedges used to withhold authorial commitment and open dialogue, boosters used to emphasize writer's certainty, attitude markers that express the writer's attitude to proposition, self-mentions explicitly referencing to authors, and engagement markers that involve the reader in the discourse.

Other researchers have proposed situational models of cohesion, which identify various situational dimensions of cohesion, such as causation, time, space, intentionality, or protagonists, expressed through particles, nouns, prepositions, verbs, or word inflection features (Kintsch 1998; Kormos 2011; Van Dijk and Kintsch 1983; Zwaan and Radvansky 1998). For example, causal cohesion evaluates the extent to which causal links between sentences are expressed; temporal cohesion reflects the extent to which tense and aspect assist in the formation of cohesion; and spatial cohesion looks at how different contents are linked by spatial particles or relations, such as the incidences of location nouns, prepositions, and motion verbs. Additionally, researchers have classified coherence relations into positive relations, i.e., extending the information provided in the text; and negative relations, i.e., restricting or ceasing to elaborate information (Louwerse 2002; Sanders et al. 1992).

Studies on L1 and L2 English writing have reported various kinds of relationships between the use of specific textual organizational features and essay quality. Studies on L1 English writers have found that global cohesion (e.g., semantic links between paragraphs) positively relates to human judgments of writing quality (McNamara et al. 2013; Neuner 1987). Local and text cohesions, however, are not strong indicators of human judgments of writing quality (Crossley et al. 2016b; Evola et al. 1980; McNamara et al. 2010). Guo et al. (2013) examined how features, including lexical sophistication, syntactic complexity, cohesion, and text length, may predict human judgments of quality of TOEFL iBT integrated essays (i.e., reading–listening to summary writing) and independent essays (i.e., argumentative writing). They found that lexical sophistication, text length, and use of past participle verbs significantly predicted essay scores for both types of essays. Nevertheless, cohesion

features including semantic similarity, noun overlap, and tense repetition predicted only writing quality for integrated essays. The number of conditional connectives, content-word overlap, and aspect repetition negatively predicted or correlated with the writing quality of independent essays. Guo et al. argued that the two writing tasks may be assessed with similar and distinct criteria. Zhao (2013) investigated authorial voice in EFL argumentative writing. He found a positive correlation between authorial voice and ratings of writing quality.

Thus, textual organizational features in L2 writing have been operationalized at multiple discourse levels (e.g., clause, sentence, paragraph, text), in intra-text or writer–reader interpersonal dimensions, as well as in various logical categories. Taken together, these perspectives improve our understanding of how L2 writers shape written discourse at various textual levels and how they communicate their intended meaning to potential readers. To obtain a fuller picture of metadiscourse measures that learners take to form their writing, the current study incorporates relevant perspectives from the theoretical frameworks discussed above, to examine how L2 Chinese learners shape written descriptive discourse. For instance, self-mentions and engagement markers in Hyland's (2005) interpersonal metadiscourse framework were included in the current analysis because they are applicable to the current writing prompt (i.e., introducing one's institution to friends), further explained in Section 3.2.

2.3. Development of L2 Textual Organzational Skills

Studies have investigated whether and, if so, how, L2 learners at different proficiencies may demonstrate differential patterning in using textual organizational features in their writing. Researchers have argued that low-proficiency writers may need to allocate significant attentional resources to low-level processing, such as spelling or linguistic encoding due to limited language skills (Kormos 2011). Consequently, they may not be able to devote sufficient cognitive resources to more global aspects of writing, such as textual organization (McCutchen 1996). In contrast, more proficient writers may attend to multiple writing areas more successfully, use more effective metadiscourse devices, and produce more cohesive texts (Bardovi-Harlig 1990; Chen and Baker 2016; Connor 1990; Crossley et al. 2016a; Lee and Deakin 2016; Yang and Sun 2012). More advanced writers may also have a greater assortment of lexical and referential devices at their disposal to promote textual cohesion (Halliday and Hasan 1976).

A number of studies have provided empirical evidence to show that compared to low-level writers, more proficient writers deploy more sophisticated and a greater range of metadiscourse devices, use cohesive devices more accurately, and present authorial voice and engage the reader more effectively. Yang and Sun (2012) discovered that L1-Chinese fourth-year college English learners used a greater number of cohesive devices in their argumentative writing and used them more accurately than second-year learners. Similarly, Ferris (1994) reported that higher-proficiency English learners used more cohesive devices that showed pragmatic appropriateness. Over a semester-long upper-level English for Academic Purposes course, Crossley et al. (2016a) found that students increased their use of local, global, and text cohesive devices in their writing. The usages of cohesive features at the local, global, and text levels predicted with a 71% accuracy rate whether an essay was written at the beginning or at the end of the semester. The cohesion features also explained 42% of the variance in the judgments of writing quality. Chen and Baker (2016) examined lexical bundles in argumentative and expository English writing. They discovered that the lexical bundles in lower-proficiency writing shared more similarity with conversational language, whereas more proficient essays were characterized by more formal lexical bundles that were closer to the register of academic prose. Reid (1992) reported that English learners, regardless of their L1s, used a lower percentage of prepositions than native writers in their essays of two topic types (comparison/contrast; chart/graph). Adopting Hyland's (2005) interpersonal metadiscourse framework, Lee and Deakin (2016) compared the usages of stance and engagement resources among three corpora of college English learners' argumentative essays: successful and less-successful essays produced by L1-Chinese learners; and successful native English essays. Their analyses revealed that successful essays by both native and L2 writers contained

significantly greater instances of hedges than less-successful essays. Compared to native writers, both groups of L2 writers were overwhelmingly resistant to establishing an authorial identity in their essays. A comparative study of English and Chinese academic writing, Hu and Cao (2011) examined the use of hedges and boosters in academic article abstracts published in applied linguistics English-medium (by both native and non-native writers) and Chinese-medium journals. They found that the abstracts published in Chinese-medium journals featured hedges markedly less frequently than those in English-medium journals which, according to Hu and Cao, can be attributed to distinct culturally preferred rhetorical norms in the Chinese and Anglo-American academic communities, respectively. Thus, the use of interactional devices may be culture-specific.

Similar to the L2 English findings, two studies on L2 Chinese writing have reported that, in comparison with native Chinese writers, L2 Chinese writers used a lower number or a narrower range of cohesive devices. Using a corpus-based approach, Li (2014) compared lexical cohesion in 50 argumentative compositions produced by advanced L1-English Chinese learners in a proficiency test with that in 50 native Chinese argumentative essays produced for the Chinese National College Entrance Examination. Li investigated various lexical cohesion features, including simple and complex repetitions, simple and complex paraphrases, superordinate and hyponymy, co-reference, and bond density (i.e., lexical repetitions across sentences). He found that, compared to native writers, L2 Chinese writers applied a lower frequency of simple and complex paraphrases and superordinate and hyponym relations, as well as a lower ratio of bond-forming sentences. Yang (2013) investigated the use of textual conjunctives and topicalizers in 30 written summaries produced by three fourth-year college Chinese learners. He compared the usages with those in the original texts produced by native Chinese authors and found that L2 Chinese learners applied a narrower range of cohesive devices.

On the other side, the findings, however, suggest that more is not necessarily better. Crossley and McNamara (2012) discovered that higher-proficiency L2 English writers produced texts with fewer cohesive devices than lower-proficiency writers. They explained their findings as a reverse cohesion effect: More proficient writers may assume that their audience includes high-knowledge readers, who benefit more from lower-cohesion texts (p. 130). Kennedy and Thorp (2007) similarly reported that compared to learners who received lower band scores, more proficient English learners applied many fewer lexico-grammatical and enumerative markers and subordinators in their argumentative essays, which appeared to be more similar to native-speaker use. In the study on L2 Chinese discussed earlier, Yang (2013) found that compared to the original texts produced by native authors, L2 Chinese learners overused certain types of cohesive devices in their written summaries, such as adversative (e.g., but, *danshi*), causative (e.g., therefore, *suoyi*), and additive (e.g., but also, *erqie*) connectives. Two previously discussed L2 English studies found that a higher ratio of pronouns was associated with low writing proficiency. Reid (1992) discovered that, in comparison with native writers, ESL writers used significantly higher percentages of pronouns and coordinate conjunctions, which was similar to the register of interactive or oral English communications. Crossley et al. (2016a) found that a higher pronoun/noun ratio negatively predicted human judgments of writing quality of academic English essays. Together, these findings suggest that more proficient L2 writers likely use certain cohesive devices more concisely, such as enumerative markers, subordinators, or pronouns.

The findings thus far have enabled us to better understand how L2 writers with different proficiencies may use metadiscourse features in distinct patterns. For instance, we know that higher-level writers deploy a greater range of cohesive devices and use them more accurately (Chen and Baker 2016; Ferris 1994; Li 2014; Yang 2013). We also know that as L2 learners grow their writing skills, they may rely less on coordinate connectives, subordinators, or pronouns, and resort more frequently to lexical cohesive devices or prepositions (Crossley et al. 2016a; Kennedy and Thorp 2007; Li 2014; Reid 1992; Yang 2013). Despite increased knowledge of L2 learners' textual organizational skills, the understandings we have obtained are derived from different studies that have used different writing genres, tasks, or learner proficiencies. It is, thus, difficult to compare the results across studies

and develop more integrated knowledge. We do not yet know in a systematic way how various textual organizational features interrelate to influence writing.

Another gap in previous research is that the majority of the studies have examined either linguistic features or discourse features in L2 texts, but not both, which does not allow us to observe how L2 writers pull together linguistic resources to form global meaning. Only a handful of studies include both linguistic and discourse features in their analyses. A previously discussed study by Crossley and McNamara (2012) examined the predictive effects of text cohesion and linguistic sophistication on L2 writing proficiency among high school English learners. Their results showed that highly proficient writers produced essays that were linguistically more sophisticated, but not more cohesive. Several linguistic and cohesive features, including lexical diversity, word frequency, word meaningfulness, aspect repetition, and word familiarity, significantly predicted writing proficiency. Kormos (2011) investigated the effects of task complexity on linguistic and discourse characteristics of narrative texts produced by upper-intermediate secondary school English learners. She found that a task variable—whether learners had to narrate a story with predetermined content or plan their own story—did not result in substantial linguistic or cohesive differences. The task conditions exerted a major impact on only one measure of lexical sophistication and had a minor effect on the explicit signaling of temporal cohesion. Guo et al. (2013), discussed earlier, examined both linguistic and cohesion features in integrated summary writing and independent argumentative writing, regarding their predictive power for human judgments of writing quality. They found that cohesive features predicted the writing quality of integrated essays only (see Section 2.2 for more details). Ferris (1994) compared lower-level and higher-level ESL texts using 28 linguistic and textual organizational measures. He found that the 28 variables divided the subjects into groups with 82% accuracy, and that higher-level students used a greater variety of lexis, syntactic constructions, and cohesive devices. Thus, these studies have examined written linguistic and cohesion features mainly in terms of their predictive capacity for human judgments of writing quality. They provide little knowledge regarding the interrelations between linguistic and discourse performances in L2 writing. Without such knowledge, we will not understand appropriately the dynamic development of L2 writing ability as a whole.

Furthermore, the previous studies have mainly analyzed argumentative writing (e.g., Chen and Baker 2016; Kennedy and Thorp 2007; Lee and Deakin 2016; Li 2014; Yang and Sun 2012). We still need to explore how learners apply organizational features in other types of writing, such as descriptive writing. The current study examines textual organizational features in L2 Chinese descriptive writing, as well as how organizational features differ between proficiencies. How various organizational features interrelate with each other, as well as how organizational features correlate with linguistic features were also investigated. Three questions guided this study:

1. What kinds of textual organizational features exist in low-score, middle-score, and high-score L2 Chinese descriptive essays, respectively, and how are the organizational features different across the groups?
2. What are the interrelations among various textual organizational features in L2 Chinese descriptive essays?
3. How do textual organizational features relate to linguistic features in L2 Chinese descriptive essays?

3. Methods

3.1. Participants and Dataset

The participants in the current study were 62 L1-English college Chinse learners from the United States, who were in China on a study-abroad program when the data were collected. The dataset comprised 62 descriptive Chinese essays produced by the participants during the placement test administered by the program. There were 27 females and 35 males, with ages from 19 to 22 years.

Students hand-wrote their essays within 30 min, based on the topic of introducing one's home university to one's Chinese friends. A descriptive writing task was used because it was suitable for both lower-level and higher-level learners. The essays were scored on a 6-point holistic scale (see Appendix C). Scores 1–2, 3–4, and 5–6 correspond roughly to the Novice, Intermediate, and Advanced levels of the proficiency scale of American Council on the Teaching of Foreign Languages (ACTFL), respectively (ACTFL 2012). The scale focused on overall writing quality and included general descriptors of the overall quality of language, content, and organization. Specific criteria on linguistic accuracy or complexity were not included; instead, they were incorporated into the descriptors of overall language and content quality. The author and a second rater evaluated the essays. Both raters were experienced college Chinese language educators. For 56 of the 62 essays (90.32%), the two raters' ratings were identical or they differed by one point, which were considered acceptable scores. The two raters' scores were averaged to derive the final score for each essay. Therefore, the final score may be an integer or a 0.5 value. For the six essays whose ratings differed by two points, the final ratings were determined through discussion.

The 62 essays had three score levels: 19 low-score (1.0–2.5), 20 middle-score (3.0–4.0), and 23 high-score (4.5–6.0). According to the program's placement results and course syllabi, the low-score students were mostly placed into Chinese first-year and second-year part I classes, roughly equivalent to the ACTFL Novice Low to Novice High levels; the middle-score students were mostly placed into Chinese second-year part II and third-year classes, roughly equivalent to the ACTFL Intermediate Low to Intermediate High levels; and the high-score students were often placed into Chinese fourth-year and fifth-year classes, roughly equivalent to the ACTFL Advanced Low to Advanced Mid levels. The low-score, middle-score, and high-score essays had a mean length of 152, 230, and 298 characters, respectively, and they contained a total of 2895, 4603, and 6864 Chinese characters, respectively. Table 1 provides a summary of the dataset in this study.

Table 1. Dataset in the current study.

	Low-Score	Middle-Score	High-Score
Score range	1.0–2.5	3.0–4.0	4.5–6.0
Number of essays	19	20	23
Mean essay length (characters)	152	230	298
Total number of characters	2895	4603	6864

3.2. Measures of Textual Organizational Features

To obtain a comprehensive picture of the metadiscourse choices that the learners made in their descriptive writing, a range of theoretically driven indices related to text cohesion and interpersonal features were designated as variables for the data analysis. First, Hyland's (2005) interactive and interactional metadiscourse framework was adopted to capture both text structure and interpersonal characteristics. Second, to analyze organizational features on a finer level, following the methods used in Crossley et al. (2016a), the interactive metadiscourse features were further classified into local (between/within clauses/sentences), global (across idea units), and text indices (across a text), based on the specific metadiscourse functions that individual indices served. Measures drawn from situation models (Kintsch 1998; Kormos 2011; Van Dijk and Kintsch 1983; Zwaan and Radvansky 1998) were also used to classify the metadiscourse features into logical categories. Categories not found in the current dataset, including interactive metadiscourse features, such as evidential, code-gloss, and endophoric markers and interactional metadiscourse features such as hedge, booster, and attitude markers, which are more relevant to genres such as argumentative or academic writing, were not included in the current analysis. Measures not relevant to the Chinese language were also excluded, for example, cohesions that concern aspect and tense. Since there are no effective computerized tools for analyzing textual organizational features in Chinese texts, features that are difficult to analyze manually, such as lexical and semantic overlap, were not included. Kormos (2011) also argued that cohesive features,

such as semantic overlap, co-reference, or latent semantic analysis, may not be fit well with short texts (p. 155), such as the essays in the current dataset.

In particular, local markers denote logical relations between or within clauses and adjacent sentences. Local transitional conjunctions, adverbs, and phrasal bundles were coded into logical categories, including continuative/additive (e.g., moreover, also, next), comparison/contrast (e.g., but), causative (e.g., therefore), and conditional/hypothetical (e.g., only if, if) markers. Adopting the misuse category in Li and Wharton (2012), incorrectly used logical devices that expressed an inaccurate semantic relation between clauses or adjacent sentences were categorized as misuse (p. 348). Moreover, the frequency of preposition usage in an essay was computed to further understand intra-clausal cohesion (Crossley et al. 2016a; Reid 1992; Smith and Frawley 1983). For example, the two prepositions (bold and underlined) in sentences (1) and (2), 跟 *gen* 'with' and 对 *dui* 'to', connect我 *wo* 'I' and 朋友们 *pengyoumen* 'friends', and 我 *wo* 'I' and 我的大学 *wodedaxue* 'my university', respectively.

(1)

我	星期五	晚上	**跟**	朋友们	吃	很	好吃	的	饭。
Wo	*xingqiwu*	*wanshang*	***gen***	*pengyoumen*	*chi*	*hen*	*haochi*	*de*	*fan.*
I	Friday	evening	**with**	friends	eat	very	delicious	Auxiliary	food

I eat very delicious food with friends on Friday evenings.

(2)

我	**对**	我的	大学	有	很	深	的	了解。
Wo	***dui***	*wode*	*daxue*	*you*	*hen*	*shen*	*de*	*liaojie.*
I	**to**	my	university	have	very	deep		Auxiliary knowledge

I have very deep knowledge of my university.

Global cohesive devices signal interconnectedness between idea units. In particular, frame markers that signal discourse acts, sequences or stages, or introduce new topics/subtopics were identified (e.g., first of all, finally, in conclusion).

Cohesion across a text was examined by evaluating the amount of given information. The proportion of third-person pronouns, as well as the third-person pronoun/noun ratio in an essay, were calculated to observe givenness and referentiality in a text (Crossley et al. 2016a, 2016b; Kyle and Crossley 2017; Reid 1992; Yang and Sun 2012). A greater proportion of third-person pronouns will indicate a higher amount of given information in a text.

Since the current writing task involved a topic of introducing one's institution to friends, an examination of interactional metadiscourse usages is relevant for understanding how the writer established authorial presence and engaged the reader. Following Hyland's (2005) framework, interactional devices were categorized into self-mention (referencing to the author; e.g., I) and engagement markers (address the reader; e.g., you). Possessive first-person pronouns including 我的 *wo de* 'my' and 我们的 *women de* 'our' were not counted, since these pronouns would be naturally needed to address the current topic and, thus, may not necessarily represent authorial voice.

The analysis of organizational performances was also supplemented with an investigation of their relationships with linguistic performances. Linguistic performances were evaluated for both accuracy and complexity. Complexity was analyzed for lexical and syntactic complexity. Lexical complexity was operationalized as lexical diversity; syntactic complexity was evaluated by clause length. See Table 2 for details on how the measures were operationalized. Accuracy was analyzed by the ratio of correct clauses in a text. Clauses containing lexical or syntactic errors were counted as incorrect clauses. Since the current writing task was timed (30 min) handwriting, students had to write fast and may produce imperfect characters with inaccurate or missing strokes. These types of errors, however, often did not prevent effective character recognition. Since character accuracy concerns a rather unique language ability and it is not the focus of the current study, characters with incorrect or missing strokes that did

not affect recognition were corrected during the transcribing process. More significant character errors that made characters unrecognizable were marked with the symbol *.

Table 2. Measures used in the current study.

Indices	Analysis Methods
Interactive Metadiscourse Markers	
Local cohesion between/within clauses/sentences Transitional markers: - continuative/additive (e.g., moreover) - comparison/contrast (e.g., but, however) - causative (e.g., therefore) - conditional/hypothetical (e.g., if) - misuse: inaccurate logical relations	Proportion of the number of transitional markers in each category against the total number of interactive metadiscourse markers (i.e., total number of transitional and frame markers) in an essay
Percentage of prepositions: relations among clausal constituents	Proportion of the number of prepositions to the total number of words in an essay
Global cohesion across idea units	
Frame markers: signal discourse acts, sequences, and stages (e.g., first, finally)	Proportion of the number of frame markers to the total number of interactive metadiscourse markers in an essay
Text cohesion	
Givenness: proportion of given to new information	Third-person pronoun/noun ratio: number of third-person pronouns divided by the number of nouns in an essay Third-person pronoun density: proportion of the number of third-person pronouns to the total number of words in an essay
Interactional Metadiscourse Markers	
Self-mention: referencing to the author (e.g., I); Engagement: addressing and involving the reader (e.g., you)	Proportion of the number of interactional markers in each category to the total number of interactional metadiscourse markers (i.e., total number of self-mention and engagement markers) in an essay
Linguistic Indices	
Linguistic accuracy: ratio of correct clauses	Number of error-free clauses divided by the total number of clauses in an essay
Lexical complexity: lexical diversity	Number of word types divided by the square root of the total number of word tokens in an essay
Syntactic complexity: clause length	Total number of words divided by the total number of clauses in an essay

3.3. Analysis

Since the essays varied in length, ratios and frequencies were calculated to control for the effect of length. Specifically, the proportions of metadiscourse markers in each category against the total number of interactive or interactional metadiscourse markers were computed to observe which types of metadiscourse features were most or least used to organize ideas or to engage the reader. The percentages of prepositions and third-person pronouns and the ratios of third-person pronoun/noun and correct clauses in an essay were also calculated.

To answer RQ 1, one-way MANOVA test was employed to identify significant differences in textual organizational features among the groups. To answer RQs 2 and 3, Pearson correlations were calculated among the metadiscourse indices, as well as between the metadiscourse and linguistic indices. Descriptive statistical analysis was also conducted. The measures and methods of analysis are summarized in Table 2.

The data were coded by the author and a second rater. The two raters independently coded the same 20% data sample, reaching interrater agreement of 88.24–97.59% for the coding of interactive and interactional metadiscourse measures. Clause accuracy had a lower interrater agreement of 86.84%. Given the challenges in achieving high interrater reliability on accuracy (Polio 1997; Polio and Shea 2014), these relatively low agreement values were considered acceptable.

4. Results

4.1. RQ1: Texual Organizational Features in the Essays

Table 3 presents the mean values of the textual organizational measures. In comparison with the low-score group, the middle-score and high-score groups produced notably higher percentages of conditional/hypothetical, frame, and engagement markers, as well as lower percentages of misuse and self-mention markers. The high-score group also produced the highest percentage (28.76%) of causative markers. The percentage of third-person pronouns and the third-person pronoun/noun ratio consistently increased from the low-score to the high-score group. Across the groups, the continuative/additive, comparison/contrast, and causative markers displayed high percentages. Figure 1 provides a visual illustration of the use of various organizational markers in the three essay groups.

Table 3. Textual organizational measures descriptive statistics.

	Low (n = 19) ***M* (*SD*)**	**Middle (n = 20)** ***M* (*SD*)**	**High (n = 23)** ***M* (*SD*)**
Interactive Metadiscourse Markers			
Local cohesion between/within clauses/sentences:			
Continuative/additive marker	17.97% (0.2318)	17.99% (0.2086)	15.47% (0.1371)
Comparison/contrast marker	16.18% (0.2648)	22.20% (0.2398)	17.96% (0.1795)
Causative marker	21.67% (0.2383)	20.42% (0.1849)	28.76% (0.1988)
Conditional/hypothetical marker	1.50% (0.0451)	5.90% (0.0872)	7.64% (0.0951)
Misuse marker	26.36% (0.3262)	9.53% (0.1213)	8.23% (0.0868)
Preposition	3.29% (0.0209)	3.20% (0.0160)	3.27% (0.0115)
Global cohesion across idea units:			
Frame marker	5.80% (0.1076)	23.95% (0.2217)	21.93% (0.1473)
Text cohesion:			
Third-person pronoun	1.49% (0.0118)	1.56% (0.0140)	2.21% (0.0120)
Third-person pronoun/noun ratio	5.54% (0.0466)	5.82% (0.0525)	7.84% (0.0454)
Interactional Metadiscourse Markers			
Self-mention marker	83.49% (0.3194)	77.00% (0.2181)	76.26% (0.2367)
Engagement marker	5.98% (0.1260)	23.00% (0.2181)	19.40% (0.1737)

The MANOVA analysis revealed a significant multivariate effect (see Table 4), Wilks' $\Lambda = 0.494$, F (22, 98) = 1.882, $p = 0.019$, partial $\eta^2 = 0.297$. The tests of between-subjects effects (see Table 5) showed that the percentages of conditional/hypothetical, frame, misuse, and engagement markers had significant differences among the groups. The frame markers, F (2, 59) = 7.079, $p = 0.002$, $\eta^2 = 0.194$, and misuse markers, F (2, 59) = 5.079, $p = 0.009$, $\eta^2 = 0.147$ displayed the highest significance level.

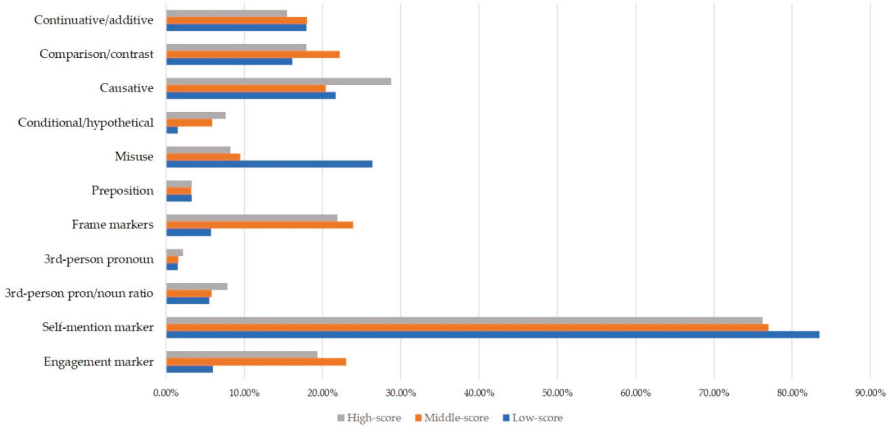


Figure 1. Use of textual organizational markers by groups.

Table 4. Textual organizational measures multivariate test.

	Value	*F*	*Df*	Error *df*	*p*	Partial eta Squared
Pillai's trace	0.567	1.799 *	22	100	0.027	0.284
Wilks' Λ	0.494	1.882 *	22	98	0.019	0.297
Hotelling's trace	0.900	1.963 *	22	96	0.013	0.310
Roy's largest root	0.730	3.316 *	11	50	0.002	0.422

* $p < 0.05$.

Table 5. Textual organizational measures tests of between-subjects effects.

Source	Dependent Variable	*Df*	*F*	*p*	Partial eta Squared
	Interactive Metadiscourse Markers				
	Local cohesion between/withinclauses/sentences:				
Groups	Continuative/additive marker	2	0.122	0.886	0.004
	Comparison/contrast marker	2	0.364	0.696	0.012
	Causative marker	2	1.021	0.367	0.033
	Conditional/hypothetical marker	2	3.150 *	0.050	0.096
	Misuse marker	2	5.079 *	0.009	0.147
	Preposition	2	0.015	0.985	0.001
	Global cohesion across idea units:				
	Frame marker	2	7.079 *	0.002	0.194
	Text cohesion:				
	Third-person pronoun	2	2.151	0.125	0.068
	Third-person pronoun/noun ratio	2	1.468	0.239	0.047
	Interactional Metadiscourse Markers				
	Self-mention marker	2	0.468	0.628	0.016
	Engagement marker	2	4.995 *	0.010	0.145
Error	Continuative/additive marker	59			
	Comparison/contrast marker	59			
	Causative marker	59			
	Conditional/hypothetical marker	59			
	Misuse marker	59			
	Preposition	59			
	Frame marker	59			
	Third-person pronoun	59			
	Third-person pronoun/noun ratio	59			
	Self-mention marker	59			
	Engagement marker	59			

* $p < 0.05$.

The post-hoc analysis results with the Bonferroni correction showed that the middle-score and high-score groups produced significantly greater percentages of frame markers and lower percentages of misuse markers than the low-score group (see Table A1 in Appendix A). The high-score group also produced a significantly higher percentage of conditional/hypothetical markers than the low-score group. The middle-score and high-score groups produced a significantly higher (p = 0.012) or near-significantly (p = 0.053) higher percentage of engagement markers, respectively, than the low-score group. The differences in the other textual organizational measures were non-significant across the groups, including continuative/additive, comparison/contrast, causative, and self-mention markers, prepositions, third-person pronouns, and the third-person pronoun/noun ratio. Thus, compared to the low-proficiency writers, more proficient writers used organizational devices more accurately, applied a higher number of frame markers to signal topics, expressed conditional/hypothetical meaning more frequently, and engaged the reader more often.

4.2. Interrelations among Textual Organizational Features

Table 6 displays the interrelations among the textual organizational measures. Since causative marker was not significantly correlated with other organizational measures, it was not included for space limitations. The correlation analysis results demonstrated several interesting patterns.

Table 6. Correlations among textual organizational measures (n = 62).

	1	2	3	4	5	6	7	8	9
1. Continuative/additive marker	—								
2. Comparison/contrast marker	−0.402 *	—							
3. Conditional/hypothetical marker	−0.094	−0.038	—						
4. Frame marker	−0.034	−0.143	−0.005	—					
5. Misuse marker	−0.093	−0.207	−0.189	−0.336 *	—				
6. Preposition	−0.285 *	0.151	−0.067	−0.032	0.211	—			
7. Third-person pron.	−0.039	−0.186	0.028	0.272 *	0.046	0.054	—		
8. Third-person pron./ noun ratio	0.012	−0.184	0.129	0.257 *	0.071	0.050	0.962 **	—	
9. Self-mention marker	0.091	0.100	−0.318 *	0.021	0.008	0.077	0.123	0.118	—
10. Engagement marker	−0.062	−0.112	0.598 **	0.165	−0.092	0.001	0.055	0.087	−0.566 **

** $p < 0.001$, * $p < 0.05$.

First, several textual organizational measures revealed significant negative correlations with each other, suggesting that a decrease in particular organizational measures was accompanied by an increase in some other organizational measures. In particular, the percentage of continuative/additive markers correlated negatively with the percentages of both comparison/contrast markers and prepositions, indicating that the writers who used more comparison/contrast markers and prepositions tended to use fewer continuative/additive markers. The percentage of conditional/hypothetical markers correlated negatively with the percentage of self-mention markers. Thus, the writers who used more conditional/hypothetical markers reduced their use of first-person accounts in their writing. The percentage of frame markers correlated negatively with the percentage of misuse markers, implying that the writers who were better at signaling their topics tended to use organizational features more accurately. Interestingly, the two types of interactional metadiscourse indices—self-mention (e.g., I) and engagement (e.g., you) markers—correlated negatively with each other (r = −0.566, $p < 0.001$), indicating that as the writers became more skillful at engaging the reader, they reduced their use of first-person voice.

Second, positive correlations were displayed among several textual organizational measures. The percentage of third-person pronouns and the third-person pronoun/noun ratio correlated positively with frame marker (r = 0.272, 0.257, $p < 0.05$), indicating that the writers who employed more third-person pronouns (e.g., he/they/it) to describe their schools were also better at signaling their topics. In addition, the percentage of conditional/hypothetical markers correlated positively with the

percentage of engagement markers ($r = 0.598$, $p < 0.001$), suggesting that the learners who used more conditional/hypothetical markers also better engaged their readers.

Last, not surprisingly, the percentage of third-person pronouns and the third-person pronoun/noun ratio correlated strongly with each other ($r = 0.962$, $p < 0.001$), suggesting that they may signal rather similar constructs. Using one of the two measures may satisfy the relevant analysis purposes.

4.3. Interrelations between Textual Organizational and Linguistic Features

Before discussing the correlation results between the organizational and linguistic measures, the results for the linguistic measures are summarized to facilitate an understanding of the relationships between organizational and linguistic performances. The results show that the mean values of all three linguistic measures—ratio of correct clauses, lexical diversity, and clause length—consistently increased from the low-score to the high-score group (see Table A2 in Appendix B). The test of between-subjects effects revealed significant group differences for all three measures (see Table A3 in Appendix B). The results of the post-hoc analysis revealed that the high-score group produced a significantly higher ratio of correct clauses than the middle-score and low-score groups. The high-score group also produced significantly greater lexical diversity and clause length than the middle-score group, which also had greater values in both measures than the low-score group (see Table A4 in Appendix B).

Table 7 presents the correlation results between the textual organizational measures and the linguistic measures. The results demonstrate that the percentage of conditional/hypothetical markers correlated positively with lexical diversity, suggesting that the learners who used more conditional/hypothetical markers also applied more diversified lexis. The percentage of frame markers correlated positively with clause length, indicating that an ability to signal topics/subtopics was aligned with an ability to produce lengthier clauses in writing. The percentage of misused markers correlated negatively with the ratio of correct clauses and lexical diversity, suggesting that when the learners improved their accurate use of organizational features, their ability to use more accurate clauses and more diversified lexis also improved. The percentage of engagement markers correlated positively with lexical diversity and clause length. Thus, the writers who used more devices to engage the reader also produced more diversified lexis and lengthier clauses. In sum, the learners' ability to use more diversified lexis in writing was positively associated with their ability to apply more accurate textual organizational devices, to use more conditional/hypothetical markers, and to better engage the reader. The learners' ability to produce lengthier clauses also aligned well with their skills to signal topics/subtopics and apply devices to engage the reader.

Table 7. Correlations between organizational measures and linguistic measures ($n = 62$).

	Ratio of Correct Clauses	Lexical Diversity	Clause Length
Continuative/additive marker	0.026	0.018	−0.045
Comparison/contrast marker	0.098	0.084	−0.047
Causative marker	0.244	0.162	0.219
Conditional/hypothetical marker	0.007	0.263 *	0.194
Misuse marker	−0.269 *	−0.296 *	−0.188
Preposition	−0.088	0.009	0.204
Frame marker	0.150	0.182	0.266 *
Third-person pronoun	0.052	0.171	0.214
Third-person pronoun/noun ratio	0.025	0.116	0.143
Self-mention marker	0.201	−0.157	−0.053
Engagement marker	0.016	0.296 *	0.337 *

* $p < 0.05$.

5. Discussion

The current findings revealed differential textual organizational features for L2 writers at different proficiencies on local, global, and text levels. Multiple organizational features also display significant negative correlations with each other. Textual organizational features characteristic of advanced writers demonstrate some positive associations with linguistic performances.

5.1. RQ1: Textual Organizational Features in the Essays

For local cohesion, the findings demonstrate that the learners across levels frequently employed continuative/additive, comparison/contrast, and causative markers to signal transitions and establish cohesion between clauses/sentences. Thus, the learners seem to already possess the ability to deploy these transitional markers to shape their writing at an early stage of development. The higher-level writers showed a significantly stronger ability to use conditional/hypothetical markers in their texts. This finding contradicts those of previous studies of L2 English, in which conditional connectives negatively predicted the quality of EFL argumentative essays (Guo et al. 2013). The discrepancy in the findings may relate to the different genres examined in the two studies, i.e., argumentative essays in Guo et al. (2013) and descriptive essays in the current study. The use of conditional connectives may be more relevant and, therefore, more needed in the current descriptive writing task. The discrepancy may also be associated with the current analysis method of aggregating conditional and hypothetical markers into one analysis category. Additionally, the higher-proficiency writers used organizational devices more accurately than the lower-level writers, which corroborates L2 English findings that fourth-year Chinese-L1 college English learners used more accurate cohesive devices than second-year learners in argumentative writing (Yang and Sun 2012). One possible explanation is that low-proficiency writers need to focus more on low-level linguistic encoding, which may have taken away attentional resources that would otherwise be available for appropriate signaling of cohesion (Halliday and Hasan 1976; Kormos 2011; McCutchen 1996).

With respect to global cohesion, the middle-score and high-score writers have demonstrated a greater ability to use frame markers to signal relations between idea units, suggesting that advanced writers are more capable of connecting ideas on a higher textual level. This result corroborates the previous findings that English learners increased their use of global cohesive devices in academic writing over a semester-long course (Crossley et al. 2016a). Regarding text cohesion, although non-significant, the percentage of third-person pronouns and the third-person pronoun/noun ratio consistently increased from the low-score group to the high-score group, suggesting that more proficient writers are able to describe their schools beyond merely discussing their first-person experience. This finding contradicts previous L2 studies, which reported that less proficient writers used a significantly higher percentage of pronouns in comparison with native writers (Reid 1992) and that pronoun-to-noun ratio negatively predicted human judgments of essay quality (Crossley et al. 2016a). The reason for the divergent findings may lie in the fact that Reid (1992) and Crossley et al. (2016a) counted all pronouns, whereas the current study only counted third-person pronouns to meet the needs of the analysis.

Concerning the use of interactional metadiscourse markers, the middle-score and high-score writers have demonstrated a stronger ability to use engagement markers to involve the reader. The differences in the number of self-mention markers, however, were non-significant among the groups, suggesting that low-level writers refer to themselves as frequently as more advanced writers while describing their institutions. This result differs from the findings in Lee and Deakin (2016) that Chinese-L1 English learners were overwhelmingly resistant to establishing an authorial identity in their argumentative essays. There are two possible reasons. First, Lee and Deakin (2016) examined Chinese-L1 English learners whose writing may have been influenced by the Chinese rhetorical tradition of preferring indirect authorial presence, whereas the current study analyzed English-L1 Chinese learners whose essays may have been impacted by the English rhetorical norms of advocating more direct authorial identity. Second, Lee and Deakin (2016) examined argumentative writing, whereas the current study investigated descriptive writing on a personalized topic. The latter may have naturally elicited

a higher level of authorial presence to address the current topic. Thus, frequent use of self-mention markers in the current descriptive writing may not necessarily reflect better writing quality.

No statistical differences were found across the groups in several interactive and interactional organizational features, including the percentages of continuative/additive, comparison/contrast, causative, and self-mention markers, the percentages of prepositions and third-person pronouns, as well as the third-person pronoun/noun ratio. There may be two reasons for these insignificant differences. First, the learners may have learnt how to use continuative/additive, comparison/contrast, causative, and self-mention markers from an early stage of learning, which may have resulted in a similar number of usages of across the groups. Second, the topic used in the current task, i.e., introducing one's institution, may have allowed a limited context for applying prepositions and third-person pronouns and the resulted low frequency of usages (e.g., 1.49% third-person pronouns, 3.29% prepositions in the low-score group) may have weakened the statistical power to detect differences between the groups.

5.2. Interrelations among the Textual Organizational Features

Regarding the interrelations among the textual organizational features, both positive and negative correlations were identified. Both of the text-level cohesive markers—percentage of third-person pronouns and the third-person pronoun/noun ratio—correlate positively with the percentage of frame markers. Thus, learners' ability to provide discussions beyond first-person accounts is positively associated with their ability to signal change of topics in writing. Given the current finding that the high-score group uses significantly more frame markers than the low-score group, more frequent use of third-person pronouns to refer to given information is likely to result in stronger cohesion.

More interesting interrelation findings lie in that multiple pairs of organizational features demonstrate significantly negative correlations, implying connected decreases and increases in specific textual organizational features. In particular, significant negative correlations were found in the following pairs of measures: (a) continuative/additive and comparison markers; (b) continuative/additive markers and preposition; (c) conditional/hypothetical and self-mention markers; (d) frame and misuse markers; and (e) self-mention and engagement markers. Thus, a higher use of continuative/additive markers is accompanied by a lower use of comparison markers and prepositions. A more frequent use of conditional/hypothetical markers is associated with a reduced use of first-person discussions. When learners become more adept at signaling their topics/sub-topics, their misuse of organizational features also declines. The negative relation between self-mention and engagement markers seems to be somewhat intuitive. When learners downplay first-person experiences, they become more aware of involving their readers. Given that the high-score group uses significantly more conditional/hypothetical, frame, and engagement markers than the low-score group, we may infer that their corresponding negative correlators—use of self-mention and misuse markers—may be characteristics of low-proficiency writers.

Combining the findings of RQs 1 and 2, we can see that in the current descriptive writing task, compared to low-proficiency writers, more advanced writers use organizational features more accurately, apply more conditional/hypothetical transitional markers, provide more third-person discussions, signal their topics/subtopics more effectively, and engage the reader more actively.

5.3. Interrelations between Textual Organizational Features and Linguistic Features

The analysis shows that the organizational features characteristic of more advanced writers, including the use of conditional/hypothetical, frame, and engagement markers, third-person accounts, and accuracy of organizational features, display positive relationships with the linguistic measures. Specifically, the conditional/hypothetical marker correlates positively with lexical diversity; the frame marker correlates positively with clause length; and the engagement marker correlates positively with both lexical diversity and clause length. Misuse marker correlates negatively with clause accuracy and lexical diversity, suggesting a connected growth between the ability to control the accuracy of organizational features and the ability to produce accurate clauses and use diversified lexis in writing.

We can see that learners' ability to apply diversified lexis in writing, an indicator of lexical complexity, is positively associated with multiple textual organizational features: accurate use of metadiscourse devices, application of devices to engage the reader, and use of conditional/hypothetical markers. These findings suggest that learners' effective use of metadiscourse devices and conditional/hypothetical markers in particular may relate to their lexical skills in the complexity dimension. Learners' ability to produce lengthier clauses, an indicator of syntactic complexity, aligns well with their metadiscourse skills in framing new topics and engaging the reader. These results indicate that as learners become more capable of developing complex clauses, they are also more skillful at signaling topic shifts and involving the reader, thus better guiding the reader's interpretations of the text towards their preferred ones (Hyland 2005). In contrast, although the high-score learners produced significantly higher ratio of correct clauses than the other two groups, clause accuracy has non-significant correlations with all metadiscourse features, except for misuse marker. This finding suggests that linguistic accuracy develops somewhat independently from the development of written metadiscourse skills.

Although only a few linguistic measures have been analyzed in the current study, the findings have provided useful knowledge regarding the connections, or lack thereof, between L2 textual organizational skills and linguistic skills. For example, the findings demonstrate that strong lexical skills are connected with effective skills to establish writer–reader interactions. The development of linguistic accuracy, however, lacks a clear connection with the development of meta-discourse skills.

6. Implications

The current study adds knowledge to our understanding of how L2 writers at different proficiencies employ metadiscourse features to shape their written discourses, as well as how various textual organizational performances relate to each other and correlate with linguistic performances.

This study has limitations that should be taken into consideration in future research in this area. The first limitation arises from the writing task used in the current study, i.e., descriptive writing with a single topic, which may have provided a limited context for applying certain metadiscourse features. For instance, the use of propositions and third-person pronouns is limited across the groups. Researchers may consider investigating whether other genres or topics may generate distinct outcomes with respect to these features. This line of research will deepen an understanding of the effects of genres and topics on textual organizations. The current findings also prove that several interactive and interactional metadiscourse markers successfully distinguish writing proficiencies. Multiple negative correlations also exist among various organizational features. These discriminative and correlational patterns will deserve additional investigations in different research contexts, such as different types of genres or other L2s. Second, the current study examined timed handwritten essays, a research condition that may have affected the composing process. Future research may explore whether type-written essays demonstrate differential textual organizational performances than handwritten ones. Finally, the current study examined only finished written products. Future studies may focus on process-oriented research to document the process of writing from beginning to completion. This research focus will help us know better the micro- and macro-level mechanisms that L2 writers go through to organize a text.

The findings also inform L2 writing pedagogy. First, they show that the development of L2 metadiscourse skills may follow specific complex patterns and may need to be nurtured in its own right. For instance, the results demonstrate that learners increase and decrease their use of particular organizational features with increased proficiency. Language instructors could consider providing more explicit guidance regarding how more or less use of specific metadiscourse features may boost coherence and organizational quality. Second, the current findings indicate that low-proficiency writers may not possess a capacity to engage the reader effectively. Language instructors may want to provide clear instructions to students from an early stage of learning on how to compose with an audience in mind. Third, the low-proficiency writers used few conditional/hypothetical and frame markers. Given that L2 learners may have learnt the linguistic items related to conditional/hypothetical markers or frame markers at lower levels of instruction, language instructors may consider designing writing

activities to guide learners to practice using a variety of logical operators to express logical meaning more effectively.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Post-Hoc Analysis Results for Organizational Measures

Table A1. Multiple comparisons with Bonferroni correction.

	(I) Groups	(J) Groups	Mean Difference (I–J)	Std. Error	*p*	95% CI	
						Lower Bound	Upper Bound
Continuative/ additive marker	low	mid	0.0002	0.0620	1.000	−0.1529	0.1525
		high	0.0250	0.0600	1.000	−0.1228	0.1727
	mid	low	0.0002	0.0620	1.000	−0.1525	0.1529
		high	0.0252	0.0591	1.000	−0.1205	0.1709
	high	low	−0.0250	0.0600	1.000	−0.1727	0.1228
		mid	−0.0252	0.0591	1.000	−0.1709	0.1205
Comparison/ contrast marker	low	mid	−0.0602	0.0730	1.000	−0.2401	0.1196
		high	−0.0179	0.0706	1.000	−0.1919	0.1562
	mid	low	0.0602	0.0730	1.000	−0.1196	0.2401
		high	0.0424	0.0697	1.000	−0.1293	0.2141
	high	low	0.0179	0.0706	1.000	−0.1562	0.1919
		mid	−0.0424	0.0697	1.000	−0.2141	0.1293
Causative marker	low	mid	0.0125	0.0665	1.000	−0.1514	0.1763
		high	−0.0709	0.0643	0.825	−0.2294	0.0877
	mid	low	−0.0125	0.0665	1.000	−0.1763	0.1514
		high	−0.0833	0.0635	0.582	−0.2397	0.0730
	high	low	0.0709	0.0643	0.825	−0.0877	0.2294
		mid	0.0833	0.0635	0.582	−0.0730	0.2397
Conditional/ hypothetical marker	low	mid	−0.0440	0.0257	0.278	−0.1073	0.0194
		high	−0.0613 *	0.0249	0.050	−0.1227	−0.00001
	mid	low	0.0440	0.0257	0.278	−0.0194	0.1073
		high	−0.0174	0.0245	1.000	−0.0778	0.0431
	high	low	0.0613 *	0.0249	0.050	0.00001	0.1227
		mid	0.0174	0.0245	1.000	−0.0431	0.0778
Misuse marker	low	mid	0.1683 *	0.0641	0.033	0.0104	0.3262
		high	0.1812 *	0.0620	0.015	0.0285	0.3340
	mid	low	−0.1683 *	0.0641	0.033	−0.3262	−0.0104
		high	0.0130	0.0612	1.000	−0.1377	0.1637
	high	low	−0.1812 *	0.0620	0.015	−0.3340	−0.0285
		mid	−0.0130	0.0612	1.000	−0.1637	0.1377
Preposition	low	mid	0.0009	0.0052	1.000	−0.0120	0.0137
		high	0.0002	0.0050	1.000	−0.0122	0.0126
	mid	low	−0.0009	0.0052	1.000	−0.0137	0.0120
		high	−0.0007	0.0050	1.000	−0.0129	0.0116
	high	low	−0.0002	0.0050	1.000	−0.0126	0.0122
		mid	0.0007	0.0050	1.000	−0.0116	0.0129
Frame marker	low	mid	−0.1816 *	0.0531	0.003	−0.3123	−0.0508
		high	−0.1614 *	0.0514	0.008	−0.2880	−0.0348
	mid	low	0.1816 *	0.0531	0.003	0.0508	0.3123
		high	0.0202	0.0507	1.000	−0.1047	0.1450
	high	low	0.1614 *	0.0514	0.008	0.0348	0.2880
		mid	−0.0202	0.0507	1.000	−0.1450	0.1047
Third-person pronoun	low	mid	−0.0007	0.0040	1.000	−0.0107	0.0093
		high	−0.0072	0.0039	0.210	−0.0169	0.0024
	mid	low	0.0007	0.0040	1.000	−0.0093	0.0107
		high	−0.0065	0.0039	0.289	−0.0160	0.0030
	high	low	0.0072	0.0039	0.210	−0.0024	0.0169
		mid	0.0065	0.0039	0.289	−0.0030	0.0160
Third-person pronoun/noun ratio	low	mid	−0.0028	0.0154	1.000	−0.0408	0.0353
		high	−0.0230	0.0149	0.387	−0.0598	0.0138
	mid	low	0.0028	0.0154	1.000	−0.0353	0.0408
		high	−0.0202	0.0147	0.524	−0.0566	0.0161
	high	low	0.0230	0.0149	0.387	−0.0138	0.0598
		mid	0.0202	0.0147	0.524	−0.0161	0.0566

Table A1. *Cont.*

	(I) Groups	(J) Groups	Mean Difference (I–J)	Std. Error	*p*	95% CI	
						Lower Bound	Upper Bound
Self-mention marker	low	mid	0.0649	0.0831	1.000	−0.1400	0.2697
		high	0.0723	0.0804	1.000	−0.1259	0.2706
	mid	low	−0.065	0.0831	1.000	−0.2697	0.1400
		high	0.0075	0.0793	1.000	−0.1880	0.2030
	high	low	−0.0723	0.0804	1.000	−0.2706	0.1259
		mid	−0.0075	0.0793	1.000	−0.2030	0.1880
Engagement marker	low	mid	−0.1701 *	0.0568	0.012	−0.3100	−0.0302
		high	−0.1341	0.0549	0.053	−0.2695	0.0013
	mid	low	0.1701 *	0.0568	0.012	0.0302	0.3100
		high	0.0360	0.0542	1.000	−0.0975	0.1695
	high	low	0.1341	0.0549	0.053	−0.0013	0.2695
		mid	−0.0360	0.0542	1.000	−0.1695	0.0975

* $p < 0.05$.

Appendix B. Post-Hoc Analysis Results for Linguistic Measures

Table A2. Linguistic measures descriptive statistics.

	Low (n = 19) *M (SD)*	Middle (n = 20) *M (SD)*	High (n = 23) *M (SD)*
Ratio of correct clauses	0.5209 (0.2397)	0.5510 (0.1196)	0.6979 (0.0852)
Lexical diversity	5.02 (0.90)	6.41 (0.89)	7.06 (0.79)
Clause length	5.53 (0.89)	6.51 (0.86)	7.73 (0.83)

Table A3. Linguistic measures tests of between-subjects effects.

Source	Dependent Variable	*Df*	*F*	*p*
Groups	Ratio of correct clauses	2	7.779 *	0.001
	Lexical diversity	2	30.13 **	<0.001
	Clause length	2	34.685 **	<0.001
Error	Ratio of correct clauses	59		
	Lexical diversity	59		
	Clause length	59		

** $p < 0.001$, * $p < 0.05$.

Table A4. Multiple comparisons with Bonferroni correction.

	(I) Groups	(J) Groups	Mean Difference (I–J)	Std. Error	*p*	95% CI	
						Lower Bound	Upper Bound
Ratio of correct clauses	low	mid	−0.0301	0.0505	1.000	−0.1546	0.0943
		high	−0.1770 *	0.0489	0.002	−0.2974	−0.0566
	mid	low	0.0301	0.0505	1.000	−0.0943	0.1546
		high	−0.1469 *	0.0482	0.010	−0.2657	−0.0281
	high	low	0.1770 *	0.0489	0.002	0.0566	0.2974
		mid	0.1469 *	0.0482	0.010	0.0281	0.2657
Lexical diversity	low	mid	−1.3891 **	0.2747	<0.001	−2.0661	−0.7121
		high	−2.0405 **	0.2659	<0.001	−2.6956	−1.3853
	mid	low	1.3891 **	0.2747	<0.001	0.7121	2.0661
		high	−0.6513 *	0.2622	0.048	−1.2974	−0.0052
	high	low	2.0405 **	0.2659	<0.001	1.3853	2.6956
		mid	0.6513 *	0.2622	0.048	0.0052	1.2974
Clause length	low	mid	−0.9734 *	0.2744	0.002	−1.6496	−0.2972
		high	−2.1962 **	0.2656	<0.001	−2.8506	−1.5419
	mid	low	0.9734 *	0.2744	0.002	0.2972	1.6496
		high	−1.2228 **	0.2619	<0.001	−1.8682	−0.5775
	high	low	2.1962 **	0.2656	<0.001	1.5419	2.8506
		mid	1.2228 **	0.2619	<0.001	0.5775	1.8682

** $p < 0.001$, * $p < 0.05$.

Appendix C. Rating Scale for the Essays

Table A5. Rating scale for the essays.

Scores	Proficiency Level	Score Criteria
1	Lower-Beginning	Limited content is presented. The meaning is difficult to understand. Limited formulaic language, such as familiar words or phrases, may be used. No discernible writing structure can be identified.
2	Higher-Beginning	Undeveloped content is presented. The meaning is generally comprehensible, but gaps in comprehension occurs. Formulaic language, such as familiar words or phrases, may be used. A very basic and undeveloped writing structure is available.
3	Lower-Intermediate	Simple and unsophisticated content is presented. A basic writing structure is available, but it lacks effective cohesion and coherence. The writing style resembles oral discourse and the writing communicates limited information to the audience.
4	Higher-Intermediate	Some variety of ideas is presented, but is often unsophisticated. A basic writing structure is available with some coherence and cohesion. The writing style resembles oral discourse and the writing communicates some basic information to the audience.
5	Lower-Advanced	A good variety of ideas is presented with some elaboration. An organized writing structure is presented with good coherence and cohesion. An introduction, elaboration, and conclusion on the topic are often presented. The writing communicates clear information to the audience.
6	Higher-Advanced	A good variety of well-developed ideas is presented. A clear and organized writing structure is evident with effective coherence and cohesion. An effective introduction, elaboration, and conclusion on the topic are presented. The writing communicates very clear information to the audience.

References

ACTFL. 2012. ACTFL Proficiency Guidelines. Available online: http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf (accessed on 2 January 2020).

Bardovi-Harlig, Kathleen. 1990. Pragmatic word order in English composition. In *Coherence in Writing: Research and Pedagogical Perspectives*. Edited by Ulla Connor and Ann M. Johns. Alexandria: TESOL, pp. 43–65.

Chen, Yu-Hua, and Paul Baker. 2016. Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics* 37: 849–80. [CrossRef]

Chiang, Steve. 2003. The importance of cohesive conditions to perceptions of writing quality at the early stages of foreign language learning. *System* 31: 471–84. [CrossRef]

Connor, Ulla. 1990. Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English* 24: 67–87.

Crossley, Scott A., and Danielle S. McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35: 115–35. [CrossRef]

Crossley, Scott A., Rod Roscoe, and Danielle S. McNamara. 2011. Predicting human scores of essay quality using computational indices of linguistic and textual features. In *Artificial Intelligence in Education: AIED 2011*. Edited by Gautam Biswas, Susan Bull, Judy Kay and Antonija Mitrovic. Berlin/Heidelberg: Springer, pp. 438–40.

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016a. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing* 32: 1–16. [CrossRef]

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016b. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48: 1227–37. [CrossRef]

Evola, Jill, Ellen Mamer, and Becky Lentz. 1980. Discrete point versus global scoring for cohesive devices. In *Research in Language Testing*. Edited by John W. Oller and Kyle Perkins. Rowley: Newbury House, pp. 177–81.

Ferris, Dana R. 1994. Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly* 28: 414–20. [CrossRef]

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36: 193–202. [CrossRef]

Guo, Liang, Scott A. Crossley, and Danielle S. McNamara. 2013. Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing* 18: 218–38. [CrossRef]

Halliday, Michael Alexander Kirkwood, and Ruqaiya Hasan. 1976. *Cohesion in English*. London: Longman.

Harman, Ruth. 2013. Literary intertextuality in genre-based pedagogies: Building lexical cohesion in fifth-grade L2 writing. *Journal of Second Language Writing* 22: 125–40. [CrossRef]

Hu, Guangwei, and Feng Cao. 2011. Hedging and boosting in abstracts of applied linguistics articles: A comparative study on English- and Chinese-medium journals. *Journal of Pragmatics* 43: 2795–809. [CrossRef]

Hyland, Ken. 2005. *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.

Jafarpur, Abdoljavad. 1991. Cohesiveness as a basis for evaluating compositions. *System* 19: 459–65. [CrossRef]

Kennedy, Chris, and Dilys Thorp. 2007. A corpus-based investigation of linguistic responses to an IELTS academic writing task. In *IELTS Collected Papers: Research in Speaking and Writing Assessment*. Edited by Lynda Taylor and Peter Falvey. Cambridge: Cambridge University Press, pp. 316–77.

Kintsch, Walter. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge: Cambridge University Press.

Kormos, Judit. 2011. Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing* 20: 148–61. [CrossRef]

Kyle, Kristopher, and Scott Crossley. 2017. Tool for the automatic analysis of cohesion. Available online: https://www.linguisticanalysistools.org/uploads/1/3/9/3/13935189/_taaco_1.5_user_manual_7-16-17.pdf (accessed on 5 January 2020).

Lee, Joseph J., and Lydia Deakin. 2016. Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing* 33: 21–34. [CrossRef]

Li, Shouji. 2014. The gap in the use of lexical cohesive devices in writing between native Chinese speakers and second language users. *Journal of the Chinese Language Teachers Association* 49: 25–47.

Li, Ting, and Sue Wharton. 2012. Metadiscourse repertoire of L1 Mandarin undergraduates writing in English: A cross-contextual, cross-disciplinary study. *Journal of English for Academic Purposes* 11: 345–56. [CrossRef]

Liu, Meihua, and George Braine. 2005. Cohesive features in argumentative writing produced by Chinese undergraduates. *System* 33: 623–36. [CrossRef]

Louwerse, Max. 2002. An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics* 12: 291–315. [CrossRef]

Mazgutova, Diana, and Judit Kormos. 2015. Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing* 29: 3–15. [CrossRef]

McCutchen, Deborah. 1996. A capacity theory of writing: Working memory in composition. *Educational Psychology Review* 8: 299–325. [CrossRef]

McNamara, Danielle S., Eileen Kintsch, Nancy Butler Songer, and Walter Kintsch. 1996. Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction* 14: 1–43. [CrossRef]

McNamara, Danielle S., Scott A. Crossley, and Philip M. McCarthy. 2010. Linguistic features of writing quality. *Written Communication* 27: 57–86. [CrossRef]

McNamara, Danielle S., Scott A. Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods* 45: 499–515. [CrossRef] [PubMed]

Neuner, Jerome L. 1987. Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English* 21: 92–105.

Polio, Charlene G. 1997. Measures of linguistic accuracy in second language writing research. *Language Learning* 47: 101–43. [CrossRef]

Polio, Charlene, and Mark C. Shea. 2014. An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing* 26: 10–27. [CrossRef]

Reid, Joy. 1992. A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing* 1: 79–107. [CrossRef]

Sanders, Ted J.M., Wilbert P.M. Spooren, and Leo G.M. Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes* 15: 1–35. [CrossRef]

Smith, Raoul N., and William J. Frawley. 1983. Conjunctive cohesion in four English genres. *Text* 3: 347–74. [CrossRef]

Van Dijk, Teun Adrianus, and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.

Yang, Chunsheng. 2013. Textual conjunctives and topic-fronting devices in CFL learners' written summaries. *Journal of the Chinese Language Teachers Association* 48: 71–89.

Yang, Wenxing, and Ying Sun. 2012. The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education* 23: 31–48. [CrossRef]

Zhao, Cecilia Guanfang. 2013. Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. *Language Testing* 30: 201–30. [CrossRef]

Zwaan, Rolf A., and Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123: 162–85. [CrossRef] [PubMed]

languages

MDPI

Article

Effects of Metacognitive Strategy Training on Chinese Listening Comprehension

Yanmei Liu

Chinese Department, Language Institute Foreign Language Center, Monterey, CA 93940, USA; yanmliu@gmail.com

Received: 8 February 2020; Accepted: 13 May 2020; Published: 15 May 2020

Abstract: In an exploration of solutions to improve Chinese second language learners' listening comprehension, this quantitative quasi-experimental study examines the effects of metacognitive strategy training (MST) on learners' metacognitive awareness, listening performance, and proficiency in an intensive language training program. In contrast with the extant research, this study designed a metacognitive learning cycle model, including self-diagnosis, planning, monitoring, evaluation, regulation, and reflection strategies, as the content of the MST. Six classes, comprised of a total of 80 participants, were assigned into three groups: Self-directed, teacher-led, and control groups. The Metacognitive Awareness Listening Questionnaire and a listening comprehension test were administered as pre- and post-tests, in addition to a proficiency test as a post-test only. Results demonstrate no significant differences in metacognitive awareness development, listening performance gains, and proficiency test results among the three groups. The results do indicate that the self-directed MST better enhances development of students' planning and evaluation awareness, and teacher-led MST workshops with special emphasis on the area of monitoring strategy will help students raise awareness. The findings of this study reveal that insufficient training time and MST without the integration of cognitive strategies do not yield significant effects. It is suggested that future MSTs should involve sufficient training time and effective follow-ups to ensure its positive effects. This study proposes that the effectiveness of MST could be improved by combining it with cognitive strategies training.

Keywords: Chinese; listening; metacognition; metacognitive strategy; strategy training

1. Introduction

1.1. Background

Listening is an essential skill for communication and learning in second language acquisition (SLA). However, listening instruction has been neglected in language curricula and practices for a significant period of time (Field 2008; Luo and Gao 2012; Richards and Rodgers 2001; Rubin 1994; Vandergrift 1999). The reason for this neglect is not that listening is unimportant, but most likely because it is difficult to learn and teach. Some researchers (Graham 2002, 2006; Prince 2013; Vandergrift 2004) have stated that listening is the most difficult skill for learners to develop in SLA, since listening comprehension is a complex, unidirectional, and unobservable process. Graham (2006) posits that "given this complexity and perhaps because the process is largely unobservable, it may be difficult for learners to have a clear understanding of how they go about listening in a foreign language, or, more importantly, how they might improve their performance" (p. 166). Listening is a unidirectional process in which listeners cannot interrupt the speaker for clarification or repetition in some contexts, such as radio broadcasts (Graham et al. 2014). Based on these reasons, "listening is probably the least explicit of the four language skills, making it the most difficult skill to learn" (Vandergrift 2004, p. 4).

Many learners have struggled to improve listening comprehension during their second language acquisition (Chang and Read 2006; Goh 2000; Graham 2006; Prince 2013; Vandergrift 2004). In the past decades, the core issues of second language listening research explored the factors attributed to listening skill development and methods to improve listening comprehension. Studies have shown that successful second language listeners used more metacognitive strategies in their listening tasks than less successful listeners (Goh 2000; Smidt and Hegelheimer 2004; Vandergrift 2002). Research has demonstrated that metacognitive awareness and strategies are crucial to learners' listening skill development. Providing metacognitive strategy training (MST) can help learners become more effective listeners (Goh 1999; Graham and Macaro 2008; Graham et al. 2008; Vandergrift 2003).

In order to assess second language listeners' metacognitive awareness and use of strategies, Vandergrift et al. (2006) developed the *Metacognitive Awareness Listening Questionnaire* (MALQ). The MALQ has been used to examine the effectiveness of MST in listening comprehension instruction in numerous studies (Altuwairesh 2013; Chang and Chang 2014; Coskun 2010; Gagen-Lanning 2015; Movahed 2014; O'Bryan and Hegelheimer 2009; Rasouli et al. 2013). Although these studies showed positive results for the use of MALQ, some doubts related to MST instructional methods and effects remain, because these studies mainly adopted teacher-led methods to provide MST; some even directly used the assessment tool, MALQ, to provide MST content (Coskun 2010; Rasouli et al. 2013). This use of the MALQ triggered more questions about MST, such as whether MST can be provided with a student-directed method and whether, with different training materials instead of using the assessment tool as the training material, MST can achieve the same positive results.

The purpose of this study is to examine the effects of MST when applying different methods—specifically self-directed and teacher-led methods—on Chinese as a second language (CSL) learners' metacognitive awareness, listening performance, and proficiency in an intensive language training program.

1.2. Literature Review

1.2.1. Metacognitive Strategy Training (MST)

The concept of metacognition was introduced by Flavell in 1976. Flavell (1976) advocated that metacognition plays an essential role in various cognitive tasks, including language acquisition, comprehension, learning, and self-instruction. He affirmed that it is beneficial and desirable to increase metacognitive knowledge and improve metacognitive skills by providing systematic training for learners (Flavell 1979). Listening comprehension is a complex cognitive process involving attention, perception, memory, information processing, problem-solving, language, and learning; based on Flavell's theory, metacognition may play an important role in listening comprehension.

Recognizing the importance of metacognition, some educators have researched how to incorporate MST into listening comprehension instruction. Goh (2000) proposed four ways to integrate MST into listening instruction: (a) Discussing problems and strategies, (b) encouraging thinking aloud, (c) using listening diaries to reflect, and (d) incorporating metacognitive activities in pre- and post-listening tasks. Vandergrift (2007) stated that the instructional tools capable of raising metacognitive awareness include questionnaires, listening diaries, and discussions, because these tools can activate listening reflection activities for both learners and teachers. Vandergrift (2007) also introduced a pedagogical cycle involving five stages with seven steps for MST in listening instruction.

When providing MST, the first question to be answered is what kind of metacognitive strategies should be taught therein. Flavell (1979) defined metacognition as a cognitive monitoring model and categorized it into four interacted components: Metacognitive knowledge, metacognitive experiences, tasks, and strategies. Metacognitive knowledge determines what strategies and actions a person adopts in a cognitive task, while metacognitive experiences occur to monitor the cognitive course, including planning beforehand, control in the process, and evaluation of the task afterwards. At least three strategies can be drawn from Flavell's description: Planning, controlling, and evaluating.

Wenden (1998) classified metacognitive strategies in SLA as planning, self-monitoring, self-evaluation, and self-reinforcement. Anderson (2002) proposed a model of metacognition that includes five components: "(a) Preparing and planning for learning, (b) selecting and using learning strategies, (c) monitoring strategy use, (d) orchestrating various strategies, and (e) evaluating strategy use and learning" (Anderson 2002, p. 2). Compared with Wenden's (1998) framework, Anderson's (2002) model lacks a component allowing learners to regulate after evaluation. Furthermore, there is an overlap between (b) "selecting and using learning strategies" and another two components: (a) Planning and (b) orchestrating. However, Anderson's inclusion of employed orchestration is noteworthy, as this concept has seldom been mentioned in other metacognition models. As Anderson stated: "The ability to coordinate, organize, and make associations among the various strategies available is a major distinction between strong and weak second language learners" (p. 4). Alongside the MALQ, Wenden's framework and Anderson's model have been widely adopted as training content in MST studies.

1.2.2. Effects of MST

MST has different effects on the learning of different subjects for students of different ages, for example, Diebold's (2011) and Prestwich's (2008) studies showed that MST had no significant effect on fourth grade students' reading, but there are a number of studies that have proven the positive effect of MST. With the greater attention to metacognitive strategies in the SLA literature, there have been increasingly more studies examining the effects of MST on SLA learners' reading and listening development over the past decade. Sterling's (2011) reading research did not find positive effects from MST, but other reading studies and almost all listening studies have found positive effects.

Among the SLA listening studies, participant samples vary in size, some as few as three (Gagen-Lanning 2015), and some around twenty (Altuwairesh 2013; Coskun 2010), and some more than forty (Abdelhafez 2006; Movahed 2014; Nosratinia et al. 2015; Chang and Chang 2014). Rasouli et al. (2013) conducted a MST study in a large-scale experiment. They examined the effectiveness of MST on 120 Iranian English as a second language (ESL) students' listening proficiency, and their results indicated that MST had a positive impact on the participants' English test results. Rasouli et al.'s (2013) study used the MALQ as an assessment. Unfortunately, Rasouli et al. did not mention the MALQ results and did not discuss the participants' metacognitive awareness changes before and after MST.

The MST treatments in these listening studies also differ in the length of time, training content, and instrument. In Abdelhafez's (2006) study, the experimental group received a 12-week training course comprised of three sets of metacognitive strategies. Coskun's (2010) study adopted the CALLA model proposed by Chamot and O'Malley (1994) and the MALQ (Vandergrift et al. 2006) as training materials for a 5-week MST treatment. Movahed's (2014) study used Vandergrift's (2007) pedagogical cycle to deliver MST, and the instruments included an anxiety scale, the MALQ, and the Test of English as a Foreign Language (TOEFL). Nosratinia et al. (2015) adopted Anderson's Model to provide MST in over 18 sessions.

It is worth noting that the MST methods utilized in the listening studies are also quite diverse. Altuwairesh (2013) employed a two-phase treatment: The MST with guided listening diaries to encourage self-reflection and deliberate practice. The study recognized the usefulness of both phases, but emphasized the significance of deliberate practice and claimed that MST was a necessary part of deliberate practice. Chang and Chang (2014) integrated MST with an online videotext listening activity to investigate their combined effects. Their study used self-dictation-generation (SDG) activities on YouTube to emphasize the use of metacognitive strategies in listening processes. In addition to researching the learners' achievement between pre- and post-listening tests, a questionnaire—the Strategy Inventory for Language Learning (SILL), developed by Oxford (1990)—and a focus-group interview were also implemented to examine listening strategy use. The study found that the online video-SDG activity facilitated the participants' development of a reflection strategy, which constitutes

an alternative instructional method to deliver MST. Similarly, Gagen-Lanning (2015) investigated the impact of MST on students' self-directed use of assisted technology in ESL listening. Gagen-Lanning delivered two 60-min sessions, including metacognitive strategies and TED Talk videos, to three participants, and then encouraged them to use self-directed TED Talk videos for their listening improvement. The study utilized the MALQ, a listening worksheet, screen casting software, and a follow-up survey to collect data. The screencast analysis showed what the participants actually did during the listening task. It is important to check what learners actually do after MST instead of only relying on questionnaire responses, but the sample size (three participants) was too small to draw a reliable conclusion from the findings. The study also found that MST could promote self-directed learning, but whether metacognitive strategies can be learned more effectively with the self-directed method in MST remains unanswered.

Although there are many differences in the MST treatments cited above, they have two points in common. One is that MST improved listening performance and metacognitive strategy use. The other is that the participants of these studies were all ESL learners. Goh (2008) posited that it is necessary to examine MST and the MALQ in various target language contexts.

1.2.3. Chinese Listening Research

Compared with ESL, CSL research is still developing, despite the fact that CSL instruction has a long history. There is little listening research in the CSL field. The reason could be either that listening is difficult to learn and research (Vandergrift 1997) or that listening just simply receives less attention. Oxford (1993) commented that language teachers frequently ignored listening as an essential skill in language learning. Although there have been studies (Chang 2010) investigating learners' reading metacognitive strategies in CSL, the number of listening strategies studies is very low, and the study of listening metacognitive strategies is rare.

The status of Chinese listening research is reflected in the following example from a well-known Chinese academic journal. *The Journal of Chinese Language Teachers Association* is an academic platform to exchange CSL studies and teaching ideas in the United States. It publishes three issues each year and has been in operation since 1966. Only seven listening studies were found in this journal from 1966 to 2016, whereas there were more studies in speaking (12), reading (36), and writing (20). Among the seven studies, four articles were reviews of listening textbooks or material, two discussed listening pedagogical issues, and only one, Cai's (2013) study, was a listening research paper.

Cai (2013) investigated four factors affecting Chinese listening proficiency and whether language heritage had an impact on the factors with 51 CSL learners. Cai's (2013) findings revealed that vocabulary and grammar knowledge are more critical for the development of listening proficiency than sound discrimination skill and metacognitive knowledge, the latter assessed by a revised version of the MALQ. Although Cai's study utilized MALQ, the correlation study is descriptive research and only reflects the relationship between metacognitive knowledge and the listening proficiency of Chinese heritage learners. Whether a cause–effect relationship exists between MST and the listening proficiency of Chinese non-heritage learners still needs to be researched. Further metacognitive research in the CSL field is needed.

1.3. Significance of the Study

The absence of studies concentrating on MST in the CSL context highlights the significance of the present study. This study is of importance for several reasons. Firstly, this study may fill gaps in the MST, MALQ, and even metacognition research in CSL listening literature given it examines the effects of MST with a different target language and population from those of existing literature. Chinese listening strategies have not been explored sufficiently in CSL, and further studies are necessary to develop the current knowledge. The results of this study may lead to a deeper understanding of the effects of MST in a CSL context. Secondly, this study presents a new perspective for MST research in the SLA domain. The existing literature on MST mainly focuses on investigating whether

MST positively impacts learners' listening and reading comprehensions. This study moves one step further: Establishing a metacognitive learning cycle (MLC) model to provide MST content, as well as an overview of examined effects of different MST methods. Thirdly, as there is controversy among existing studies surrounding the effects of MST, the findings of this empirical study could provide researchers a more comprehensive understanding of MST and present further evidence for educators and instructional leaders to enable decision making when adopting MST in world languages programs.

1.4. Research Questions

To achieve the purpose of this study, three research questions (RQ) are addressed:

RQ1: Are there significant differences in metacognitive awareness development among two experimental groups receiving MST (self-directed versus teacher-led) and a control group?

RQ2: Are there significant differences in Chinese listening performance gains among two experimental groups receiving MST (self-directed versus teacher-led) and a control group?

RQ3: Are there significant differences in the results of a Chinese listening proficiency test among two experimental groups receiving MST (self-directed versus teacher-led) and a control group?

2. Materials and Methods

2.1. Research Design

This study employed a quantitative quasi-experimental research method, including single factor within-subject and between-subjects designs. The within-subject design tested whether there were differences in metacognitive awareness and Chinese listening performance between the pre- and post-tests of experimental participants, as measured by the MALQ and classroom listening tests, respectively. In order to ensure the three experimental groups comprised of original classes were comparable, the between-subjects design examined whether there were differences in metacognitive awareness development and Chinese listening performance gains, that is, comparing the changes in the MALQ and the classroom listening pre- and post-tests, respectively, before and after the intervention, or MST. For the examination of final listening proficiency, this study used a statistical method to adjust for possible preexisting differences in the three groups, with the classroom listening pre-test as a covariate. The independent variable was the MST method, including three levels: Self-directed training, teacher-led training, and no training. The dependent variables were participants' metacognitive awareness, Chinese listening performance, and Chinese listening proficiency.

2.1.1. MST Content

In contrast with the extant research, this study designed a metacognitive learning cycle (MLC), including self-diagnosis, planning, monitoring, evaluation, regulation, and reflection strategies, to provide MST. In addition to the four essential metacognitive strategies (planning, monitoring, evaluation, and regulation) highlighted by early researchers, this study incorporated self-diagnosis and reflection into the MST. In the MLC, self-diagnosis is the starting point, and reflection the end point. In practice, the two strategies can be applied simultaneously to form a continuous metacognitive cycle, which means the self-diagnosis is a reflection on previous actions, and the reflection is also a self-diagnosis for the following task. This cycle is displayed in Figure 1.

The training material of the MST included an introduction and the MLC strategies content. The introduction presents what metacognition is, along with why and how to learn metacognitive strategies. The content of the self-diagnosis strategy contains how to self-diagnose listening problems and possible reasons based on a self-assessment with the Listening Self-Diagnosis Assistant (LSDA). The planning strategy consists of two parts: How to make plans and how to effectively manage time. The monitoring strategy states how to focus attention and process meaning during a listening activity. The evaluation strategy focuses on how to assess the quality of listening comprehension and the effectiveness of strategy use. The regulation strategy is about how to reinforce and adjust strategy use

after evaluation. The reflection strategy introduces how to analyze and summarize one's own listening performance and problems using answers to the guided questions in the training material. The content of each strategy covers three parts: What, why, and how.

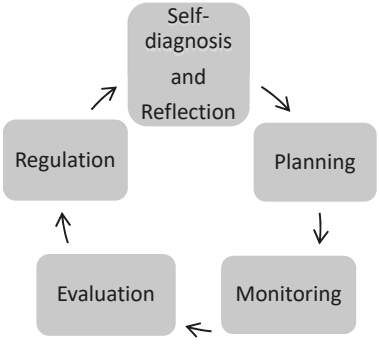


Figure 1. Metacognitive learning cycle.

2.1.2. MST Instruction

In this study, two experimental groups received the same content of the MLC as previously described. The designed learning time of the MLC was the same for both self-directed and teacher-led groups: 90 min. Both groups' participants were encouraged to apply the MLC strategies into daily listening activities and record their strategy use on provided worksheets. There was no preset time length for application activities. The difference between the two experimental groups lies in the instructional method, including training means, procedures, and activities.

The self-directed group attended a self-study session and the training lasted six weeks. The participants spent 15 min on learning one strategy from a handout prepared by the researcher each Monday, and then followed instructions on an application worksheet to practice and record their strategy use every day in that week. The application worksheet provided step-by-step instructions for the strategy use, and varied each week based on the strategy introduced. The application tasks on the worksheet were integrated into the participants' listening homework in this group. A sample page of the self-directed group's worksheets is in Appendix A.

By contrast, the teacher-led group received two 45-min workshops. In the workshops, the researcher introduced the MLC strategies with PowerPoint slides and explained why, when, and how to apply the strategies—the same content as in the handout for the self-directed group. The teacher-led group participants were encouraged to apply the strategies in their daily listening activities whenever and wherever needed after the workshops. They were also asked to record their strategy use on a provided worksheet. However, there were no step-by-step guided instructions for strategy use and no one-by-one focus for each week in the teacher-led group. The worksheet requested the participants to include the date, and to circle the strategy used and the activity applied. A sample of the teacher-led group worksheet is provided in Appendix B.

2.2. Participants and Sampling

The population of this study was the students learning Chinese as a second language in an intensive training program at a world languages center on the west coast in the United States. There were 30 classes, including approximately 360 students in the program when this study started. Among these classes, there were 10 classes in each of the three semesters in the year. As this quasi-experiment was conducted in a real teaching setting, some practical factors had to be considered, such as the appropriate timing of training, available assessment tools, integration with the current curriculum, comparability of different group participants, and so on. These factors led to the decision to adopt a

cluster sampling method for participant selection. According to Babbie (2007), "Cluster sampling is ideal when it is impossible or impractical to compile a list of the elements composing the population" (Creswell 2008, p. 148).

In the first step of cluster sampling, 10 classes of Semester Three became the targeted participating classes. Among the 10 classes, three would graduate in one month and therefore lacked sufficient time to complete the experiment. The remaining seven classes were all invited to voluntarily participate in the study. Six classes accepted the invitation and became participating classes. The students in the same classes were in the same group of the study, in an attempt to mitigate the threat of intervention diffusion among students. There were no heritage learners in the six classes. The students' level of Chinese listening proficiency is Intermediate High or Advanced Low in Semester Three.

For the second step of cluster sampling, the six participating classes were randomly assigned into three groups: Self-directed, teacher-led, and control groups, whereby each group was comprised of two classes. Eighty out of 86 students voluntarily participated in the study, and 72 completed both pre- and post-tests of the MALQ. Four classes, with 49 participants in total, were the experimental groups. Two classes of them, with 21 participants, received the self-directed MST, and the other two, with 28 participants, had the teacher-led MST. Another two classes, with 31 students in total, formed the control group. The participants' demographic information is provided in Table 1.

Table 1. Participants' gender, age, and education frequencies.

		Control Group		Teacher-Led Group		Self-Directed Group	
		N	**Percent**	**N**	**Percent**	**N**	**Percent**
Gender	Male	25	80.6	20	71.4	14	66.7
	Female	6	19.4	8	28.6	7	33.3
	Total	31	100.0	28	100.0	21	100.0
Age	19–20	2	15.4	5	41.7	5	23.9
	21–25	8	61.5	5	41.7	9	42.8
	26–30	3	23.1	2	16.6	4	19.0
	31					3	14.3
	Total	13 *	100.0	12 *	100.0	21	100.0
Education	High School	22	71.0	20	71.4	10	47.6
	Associate Degree			4	14.3	2	9.5
	Bachelor's	9	29.0	2	7.1	8	38.1
	Master's			2	7.1	1	4.8
	Total	31	100.0	28	100.0	21	100.0

* Note: One class was not asked to provide ages, but the participants' ages were 19–31 in the class.

2.3. Instrumentation

There were three dependent variables in this quantitative study. The first was listening metacognitive awareness, which was measured by the MALQ. The second, Chinese listening performance, was assessed by the Chinese Listening Comprehension Test (CLCT), a classroom test developed by a test team in the program. The third dependent variable was Chinese listening proficiency, represented by the Defense Language Proficiency Test (DLPT) listening score. The reasons for adopting the two listening test instruments are their convenience and standardization. The CLCT is not a standardized test, but it is convenient for checking changes in participants' listening performances before and after MST. As the DLPT is administered only once at the end of the program at this particular language center, it cannot show the participants' listening performance gains before and after MST; but it is a standardized proficiency test that has been validated.

2.3.1. MALQ

The MALQ (Vandergrift et al. 2006) was designed to assess second language listeners' metacognitive awareness. The reliability of the MALQ ranges from 0.68 to 0.78 according to Cronbach's alpha. During

its validation process, the MALQ developers used the questionnaire to assess French and English learners' listening metacognitive awareness, and a five-factor model emerged based on a confirmatory factor analysis.

The five factors were constituted by 21 items: Factor 1, problem-solving (PS: 5, 7, 9, 13, 17, 19), addresses inference and regulating strategies; Factor 2, planning and evaluation (PE: 1, 10, 14, 20, 21), assesses pre-listening preparation and while or after listening self-judgement; Factor 3, mental translation (MT: 4, 11, 18), contains three obstacles that listeners should overcome to become skilled while listening; Factor 4, directed attention (DA: 2, 6, 12, 16), refers to concentration strategies while listening; and Factor 5, person knowledge (PK: 3, 8, 15), represents listener's self-perception and self-efficacy on task difficulty. The responses to six items (3, 4, 8, 11, 16, 18) must be reverse coded because they are strategies for which lower scores are desirable. In order to avoid a neutral point caused by respondent's hedging, the questionnaire adopted a 6-point Likert scale from "strongly agree" to "strongly disagree". (Vandergrift et al. 2006).

The MALQ has been used for English learners from different countries, such as China, Singapore, Iran, and Saudi Arabia, but it has not been used for CSL non-heritage learners. The researcher received permission to use the MALQ in this study from a leading developer. This study used the original questionnaire without any revisions.

2.3.2. CLCT

The CLCT was developed by an ad hoc test team in the Chinese program at this language center for the purpose of providing a complete test sample to familiarize students with the DLPT format. The CLCT is a traditional paper and pencil instrument and is administered at the beginning and end of semester three in the program. The CLCT includes 40 authentic passages and 60 questions with four answer options. It is graded with raw scores, with a score range from 0 to 60 for the number of questions answered correctly. Each question has only one correct answer. The correct answer keys were provided to each rater, thereby ensuring the inter-rater reliability of different classes.

2.3.3. DLPT

The DLPT is a validated high-stakes proficiency test. It was designed to assess native English speakers' foreign language proficiency as defined by the Interagency Language Roundtable Skill Level Descriptions. The DLPT is used government- and military-wide in the U.S., and is comprised of two separate tests, listening and reading; it is available in different languages. The Chinese DLPT listening test consists of 40 authentic passages with 60 questions related to those passages. The Chinese DLPT is in a multiple-choice format, and each question is followed by four answer choices. The test is delivered on computers at an appointed testing center and graded at the same location. Test scores are in the format of level numbers, including 0+, 1, 1+, 2, 2+, and 3, based on the number of questions answered correctly.

2.3.4. Worksheets

Strategy application worksheets used for the two experimental groups' participants were an additional instrument in the study. Both groups' participants were encouraged to record their strategy applications on the worksheets. Both groups' worksheets were graded based on the same rubric, to quantify the quality and quantity of the experimental participants' actual strategy applications. The scoring rubric included two parts, quantity and quality scores; quantity was the number of uses for a strategy on each day, and quality represented how well the strategy was used each day, including none (0 point), fair (1 point), good (2 points), and excellent (3 points). The worksheet rubric is in Appendix C. The worksheet instrument not only helped the participants apply the strategies learned, but also provided a tool to observe the participants' actual strategy use in MST.

The four instruments formed a continuum of the actual strategy use (worksheets), self-reported strategy use (MALQ), developing listening performance (CLCT), and final listening proficiency (DLPT), which helped to methodically investigate the effects of MST on the experimental participants' listening comprehension development.

2.4. Procedure

All participants gave their informed consent for inclusion before they participated in this study which was approved by the ethics committees of the researcher's university and affiliation. Before the intervention, the MST, both experimental and control groups had taken the CLCT as a pre-test in the first week of Semester Three. The teachers graded the CLCT. Participants completed the MALQ the second week of Semester Three to assess their listening metacognitive awareness. The researcher administered the MALQ and collected the responses.

The MST started after the completion of the MALQ the third week of Semester Three. The self-directed group participants self-studied one strategy delivered on a handout for 15 min each Monday, and then followed a worksheet to practice and record the strategy use from Tuesday to Friday; the training lasted six weeks, and the total learning time of six MLC strategies was 90 min. The teacher-led group attended two 45-min workshops to learn the six MLC strategies by participating in the trainer's PowerPoint presentations. The first workshop introduced the first three MLC strategies in the third week and the second workshop finished the other three strategies in the fourth week. The participants were encouraged to use a worksheet to record their strategies use on a daily basis after the first workshop. The researcher was the only trainer for the two experimental groups. The researcher developed and provided all learning materials for the self-directed group and conducted MLC workshops for the teacher-led group.

After the MST that lasted 6 weeks, both experimental and control groups completed the MALQ again and the second set of CLCT in their ninth week of Semester Three, and then took the Chinese DLPT in the last week. The researcher collected all data and conducted the analysis.

2.5. Data Analysis

RQ1: Are there significant differences in metacognitive awareness development among the two experimental groups receiving MST (self-directed versus teacher-led) and the control group?

Among 80 participants, 72 took both pre- and post-tests of the MALQ. The data analysis of RQ1 was based on the 72 pairs of MALQ responses. The responses to six items (3, 4, 8, 11, 16, 18) were reverse coded according to the MALQ Scoring and Interpretation Guide. In order to accurately evaluate participants' metacognitive awareness development and reduce the bias of participants' possible differences before the MST, the difference for each item between the pre- and post-tests was calculated first by subtracting the pre-test value from the post-test value. The differences among the two experimental groups and the control group were compared with a one-way ANOVA. In order to look into each group's metacognitive awareness development, each group's pre- and post-test results were compared with a paired-samples *t*-test.

RQ2: Are there significant differences in Chinese listening performance gains among the two experimental groups receiving MST (self-directed versus teacher-led) and the control group?

All 80 participants completed the pre- and post-test of CLCT, and the data analysis was based on their scores in the two tests. A one-way ANOVA was used to examine participants' listening performance gains among three groups, that is, the difference between their pre- and post-test scores. A paired-samples *t*-test was performed, followed by a Pearson correlation analysis, to investigate each group's gains and the relationship with metacognitive awareness development.

RQ3: Are there significant differences in the results of a Chinese listening proficiency test among the two experimental groups receiving MST (self-directed versus teacher-led) and the control group?

All participants completed the Chinese DLPT at the end of the experiment. Their DLPT listening results varied from level 0+ to level 3, specifically 0+, 1, 1+, 2, 2+, and 3. The levels from 0+ to 3

were transformed into numbers from 0 to 5. As participants took the CLCT as a pre-test before the MST, this study made the pre-test scores a covariate when comparing the three groups' DLPT results. The RQ3 was answered with an analysis of covariance (ANCOVA), a general linear model that blends ANOVA and regression, followed by a Pearson correlation analysis. The ANCOVA was used to reduce the bias from participants' possible different performances before the MST and accurately evaluate the effectiveness of MST. The correlation analysis examined relationships between the DLPT results and the five factors identified in the MALQ.

3. Results

3.1. Metacognitive Awareness Development

After the MST, the two experimental groups demonstrated more awareness than the control group in the metacognitive factors. Table 2 presents the results of the MALQ pre- and post-tests. The overall means of the three participant groups increased from the pre-test to the post-test, with the self-directed group showing the greatest increase (4.48), followed by the teacher-led (3.19) and the control (2.21) groups.

Table 2. Metacognitive Awareness Listening Questionnaire (MALQ) descriptive statistics.

	Group	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
Pre-test	Control	24	85.04	7.40	1.51	68.00	98.00
	Teacher-led	27	86.07	9.84	1.86	68.00	107.00
	Self-directed	21	84.57	10.80	2.36	56.00	110.00
	Total	72	85.30	9.31	1.09	56.00	110.00
Post-test	Control	24	87.25	7.32	1.38	72.00	101.00
	Teacher-led	27	89.26	11.81	2.27	70.00	111.00
	Self-directed	21	89.05	8.56	1.87	75.00	103.00
	Total	72	88.46	9.39	1.08	70.00	111.00

There are 21 items constituting five distinct factors/subscales in the MALQ: Problem-solving (PS), planning and evaluation (PE), mental translation (MT), directed attention (DA), and person knowledge (PK). For each of the five factors, the difference between pre- and post-test scores was calculated; this difference represents "metacognitive awareness development." The descriptive statistics of the metacognitive awareness development for the five factors are shown in Table 3. The self-directed group's means of PE, DA, and PK factors were higher than those of the teacher-led and control groups, particularly on PE and PK factors, but lower on PS and MT factors, where the control group showed the greatest development.

Table 3. Metacognitive awareness development on the MALQ factors.

Group		N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
PS	Control	24	0.92	3.361	0.686	−6	10
	Teacher-led	27	0.41	2.707	0.521	−6	5
	Self-directed	21	−0.48	3.696	0.807	−7	8
	Total	72	0.32	3.241	0.382	−7	10
PE	Control	24	1.83	3.102	0.633	−2	8
	Teacher-led	27	2.41	4.236	0.815	−6	11
	Self-directed	21	3.52	6.080	1.327	−4	16
	Total	72	2.54	4.534	0.534	−6	16
MT	Control	24	0.38	2.183	0.446	−4	4
	Teacher-led	27	0.07	2.269	0.437	−5	6
	Self-directed	21	0.10	2.071	0.452	−4	3
	Total	72	0.18	2.158	0.254	−5	6

Table 3. *Cont.*

Group		N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
DA	Control	24	0.58	1.816	0.371	−3	5
	Teacher-led	27	0.93	2.385	0.459	−3	6
	Self-directed	21	1.00	2.720	0.594	−6	6
	Total	72	0.83	2.295	0.270	−6	6
PK	Control	24	−0.67	2.078	0.424	−5	4
	Teacher-led	27	−0.30	2.599	0.500	−6	6
	Self-directed	21	0.33	2.352	0.513	−4	5
	Total	72	−0.24	2.365	0.279	−6	6

Notes: PS = Problem-solving, PE = Planning and Evaluation, MT = Mental Translation, DA = Directed Attention, PK = Person Knowledge.

A one-way ANOVA was conducted to compare the three groups' metacognitive awareness development within the five factors. The results indicate that there were no statistically significant differences in metacognitive awareness development among the two experimental groups (self-directed and teacher-led) and the control group.

In order to examine metacognitive awareness development within groups, each group's pre- and post-scores for the five factors were compared using a paired-samples *t*-test. The *t*-test result presented in Table 4 shows that there were significant differences in PE factor development within the self-directed ($p = 0.015$), teacher-led ($p = 0.007$), and control ($p = 0.008$) groups. The different means indicate that the self-directed (M = 3.524) group showed the greatest difference between pre- and post-tests, followed by the teacher-led (M = 2.407) and control (M = 1.833) groups. Additionally, there was a significant difference in DA factor development within the teacher-led group ($p = 0.05$), although the self-directed group showed the greatest mean difference (M = 1.000). This indicates that all three groups showed a significant increase in metacognitive awareness development on the PE factor, regardless of MST intervention, but only the teacher-led group had a significant increase in the DA factor after MST. Table 4 confirmed the results from Tables 2 and 3. That is, except for PS and MT, two cognitive-focused skills, the other three metacognitive factors clearly demonstrated higher means for the experimental groups than for the control group.

Table 4. Tests of within-subject effects on metacognitive awareness development.

Group		Paired Differences					T	df	Sig. (2-Tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Control	PS	0.917	3.361	0.686	−0.503	2.336	1.336	23	0.195
	PE	1.833	3.102	0.633	0.523	3.143	2.895	23	0.008 *
	MT	0.375	2.183	0.446	−0.547	1.297	0.841	23	0.409
	DA	0.583	1.816	0.371	−0.183	1.350	1.574	23	0.129
	PK	−0.667	2.078	0.424	−1.54	0.211	−1.572	23	0.130
Teacher-led	PS	0.407	2.707	0.521	−0.663	1.478	0.782	26	0.441
	PE	2.407	4.236	0.815	0.732	4.083	2.953	26	0.007 *
	MT	0.074	2.269	0.437	−0.823	0.972	0.170	26	0.867
	DA	0.926	2.385	0.459	−0.017	1.869	2.018	26	0.054 *
	PK	−0.296	2.599	0.500	−1.32	0.732	−0.592	26	0.559
Self-directed	PS	−0.476	3.696	0.807	−2.16	1.206	−0.590	20	0.562
	PE	3.524	6.080	1.327	0.756	6.291	2.656	20	0.015 *
	MT	0.095	2.071	0.452	−0.848	1.038	0.211	20	0.835
	DA	1.000	2.720	0.594	−0.238	2.238	1.685	20	0.108
	PK	0.333	2.352	0.513	−0.737	1.404	0.649	20	0.523

* Correlation is significant at the 0.05 level (2-tailed). Notes: PS = Problem-solving, PE = Planning and Evaluation, MT = Mental Translation, DA = Directed Attention, PK = Person Knowledge.

The participants in the two experimental groups were asked to apply and record their metacognitive strategy use, and, although the participants did not record many uses, their application data was collected and graded according to a scoring rubric. The application scores were analyzed with a Pearson correlation to investigate relationships between strategy application and metacognitive awareness development on the five factors. There were no significant correlations between participants' strategy applications and metacognitive awareness development on the five factors ($p > 0.05$).

3.2. Listening Performance Gains

Table 5 shows descriptive statistics for participants' pre- and post-test results in the CLCT. Table 6 presents three groups' listening performance gains, which are the differences between their post- and pre-test scores. The results in Table 6 show that the control group had the highest gain (M = 9.19; SD = 4.64), followed by the teacher-led (M = 6.82; SD = 4.46) and self-directed (M = 6.00; SD = 5.99) groups.

Table 5. Descriptive statistics for listening performance.

Group		Mean	N	Std. Deviation	Std. Error Mean
Control	Pre	39.19	31	5.606	1.007
	Post	48.39	31	5.258	0.944
Teacher-led	Pre	41.61	28	7.445	1.407
	Post	48.43	28	6.839	1.292
Self-directed	Pre	40.19	21	8.424	1.838
	Post	46.19	21	6.638	1.449

Table 6. Descriptive statistics for listening performance gains.

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Control	31	9.1935	4.63623	0.83269	7.4930	10.8941	0.00	16.00
Teacher-led	28	6.8214	4.46429	0.84367	5.0904	8.5525	−4.00	14.00
Self-directed	21	6.0000	5.99166	1.30749	3.2726	8.7274	0.00	24.00
Total	80	7.5250	5.09399	0.56953	6.3914	8.6586	−4.00	24.00

A one-way ANOVA was performed to examine whether the differences among the three groups were statistically significant. The ANOVA results indicate that there were no significant differences ($F = 3.018$, $p = 0.055$) in listening performance gains among the three groups, although the control group showed the highest mean in the performance gain. Additionally, a Levene's test for the homogeneity of variance was employed, confirming that variances among the three groups were equal ($F = 0.700$, $p = 0.500$).

Before running the paired-samples *t*-test, the database was split by the MST variable. The paired-samples *t*-test results in Table 7 show that there were significant differences between the pre- and post-test scores in the CLCT for all three groups, including the self-directed ($p = 0.000$), teacher-led ($p = 0.000$), and control ($p = 0.000$) groups.

Table 7. Paired-samples *t*-test on listening performance.

Group		Paired Differences					T	df	Sig. (2-Tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Control	Pre-Post	−9.194	4.636	0.833	−10.894	−7.493	−11.041	30	0.000 *
Teacher-led	Pre-Post	−6.821	4.464	0.844	−8.552	−5.090	−8.085	27	0.000 *
Self-directed	Pre-Post	−6.000	5.992	1.307	−8.727	−3.273	−4.589	20	0.000 *

* Correlation is significant at the 0.05 level (2-tailed).

A Pearson correlation analysis was adopted to examine whether the significant listening performance gains in the three groups were related to participants' metacognitive awareness development. The results show that there were no significant correlations between listening performance gains and metacognitive awareness development on the five factors ($p > 0.05$). This means there were likely some other factors that resulted in participants' Chinese listening performance gains across the three groups.

3.3. DLPT Listening Results

The descriptive statistics of the three groups' DLPT listening results show that the teacher-led group's mean (M = 4.07; SD = 0.900) was slightly higher than the control (M = 3.87; SD = 0.846) and self-directed (M = 3.71; SD = 1.102) groups' means. This is similar to the groups' CLCT pre-test sequence, where the teacher-led group (M = 41.61) was slightly higher than the self-directed (M = 40.19) and control (M = 39.19) groups. The Levene's test for equality of variances demonstrates that error variances of the dependent variable were equal across groups (F (77) = 2.055, $p > 0.05$).

An ANCOVA was performed to answer the RQ3, with reduction in the bias from participants' possible different performances before the MST. In the ANCOVA analysis, the DLPT listening results were set as the dependent variable, the MST group as the independent variable, and the CLCT pre-test score as a covariate to compare the three groups' DLPT listening proficiency. The ANCOVA results presented in Table 8 shows that there was no overall statistically significant difference in Chinese DLPT listening results among the three groups after their means were adjusted using the CLCT pre-test scores [*p*-value (F = 0.733, $p = 0.484$)].

Table 8. Tests of between-subjects effects on Defense Language Proficiency Test (DLPT) listening proficiency.

Dependent Variable: DLPT Listening						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	22.139 [a]	3	7.380	11.917	0.000	0.320
Intercept	2.040	1	2.040	3.294	0.073	0.042
Performance Pretest	20.565	1	20.565	33.211	0.000	0.304
MST	0.907	2	0.454	0.733	0.484	0.019
Error	47.061	76	0.619			
Total	1286.000	80				
Corrected Total	69.200	79				

[a]: R Squared = 0.320 (Adjusted R Squared = 0.293).

After the ANCOVA analysis, a Pearson correlation analysis was performed in order to examine relationships between Chinese DLPT listening scores and the five metacognitive factors measured in the post-test of the MALQ. The results presented in Table 9 show that the DLPT listening results were significantly but negatively correlated with the PE factor (r = −0.235, $p = 0.041$) and the MT factor (r = −0.238, $p = 0.039$), and not significantly correlated with the other three factors. The negative correlations means that high DLPT scores were associated with low scores in the PE and MT factors.

Table 9. Correlations between DLPT listening proficiency and the MALQ factors.

		DLPT Listening	PS Post-Test	PE Post-Test	MT Post-Test	DA Post-Test	PK Post-Test
DLPT Listening	Pearson Correlation	1	0.090	−0.235 *	−0.238 *	−0.064	−0.052
	Sig. (2-tailed)		0.441	0.041	0.039	0.581	0.657
	Sum of Squares and Cross-products	69.200	21.658	−69.211	−46.816	−8.961	−8.368
	Covariance	0.876	0.289	−0.923	−0.624	−0.119	−0.112
	N	80	76	76	76	76	76

* Correlation is significant at the 0.05 level (2-tailed). Notes: PS = Problem-solving, PE = Planning and Evaluation, MT = Mental Translation, DA = Directed Attention, PK = Person Knowledge.

3.4. Results Summary

Overall results showed that there were no statistically significant differences in metacognitive awareness development, listening performance gains, and DLPT listening proficiency among the three groups. However, there were significant differences in PE factor development within the three groups, with the self-directed group showing the highest difference between pre- and post-tests, followed by the teacher-led and control groups. There was also a significant difference in DA factor development within the teacher-led group. Furthermore, the two experimental groups demonstrated more awareness than the control group in most metacognitive factors except for two cognitive-focused factors, PS and MT, in the MALQ descriptive data. The self-directed group showed higher scores on DA, and particularly on PE and PK, than the other two groups. The results indicate that the self-directed MST can bring the highest significant difference on the PE factor and the teacher-led MST will significantly improve DA awareness. Both MST training methods promoted the awareness of metacognitive factors except for the cognitive-focused factors. The different MST methods did not show significant differences in participants' Chinese listening performance gains nor in proficiency test results.

4. Discussion

4.1. RQ1 Findings

It was anticipated that there would be significant differences in metacognitive awareness development among the two experimental groups receiving MST (self-directed and teacher-led) and the control group. The differences were not supported by the statistics found in this study. In other words, there was no apparent development of metacognitive awareness after participating in MST. However, the *t*-test results indicate that the self-directed MST can bring the highest significant difference on the PE factor and the teacher-led MST can significantly improve DA awareness.

One possible reason for the non-significant overall results could be the differences between the instrument (MALQ) and the training content of MST (MLC). The MALQ is a questionnaire using self-report to assess metacognitive awareness represented by five factors: Problem-solving, planning and evaluation, mental translation, directed attention, and person knowledge. This study did not directly teach the five factors, but used the MLC model, consisting of self-diagnosis, planning, monitoring, evaluation, regulation, and reflection strategies, to provide MST. Although there were some overlaps (planning and evaluation) and similarities (DA factor and monitoring strategy, PK factor and self-diagnosis strategy) between the MALQ and the MLC, there were differences on other factors, such as problem-solving and mental translation in the MALQ, and regulation and reflection in the MLC. Some previous studies used the MALQ to provide MST content (Coskun 2010; Rasouli et al. 2013), which may then lead to a significant positive effect when answering the questionnaire. The MALQ is likely not the best instrument to assess what was taught in an experiment with different MST materials, such as the MLC model in this study.

The second possible explanation for the non-significant results could be attributed to insufficient training time and the teaching quality of the MST in this study. Although 90 min was sufficient for

introducing metacognitive strategies to the experiment participants, it may have been insufficient for a positive impact on the change of their metacognitive awareness. The participants might need more time to apply and internalize the strategies learnt, and the positive effects of MST might be reflected after more practice. In previous studies, treatment time was 45 to 50 min per week for five or more weeks (Birjandi and Rahimi 2012; Coskun 2010; Rasouli et al. 2013). For example, Birjandi and Rahimi's (2012) MST lasted six weeks and took 45 min once a week. Aside from the time insufficiency, the intensity of training might not have been enough; the researcher had concerns that the experiment participants would opt to withdraw from the MST if the training was too intense, because they already had a heavy learning workload and other duties on a daily basis. In addition, because participation was completely voluntary, the training lacked accountability, which means there were no follow-up actions if the participants did not practice, record, and submit their strategies application worksheets. The lack of accountability may have influenced the quality of the MST. As a result, the MST appears not to have had an overall impact on the experiment participants' listening metacognitive awareness development.

The *t*-test results showed that there were significant differences in planning and evaluation factor development for all three groups, including the control group. Planning before listening and evaluating after listening are two strategies commonly used in listening activities and may be taught in listening instruction or improved by learners themselves as they use metacognitive strategies. If participants did not spontaneously improve their planning and evaluation strategies as they developed their listening comprehension skill, then the two strategies were very likely taught to the control group during their routine listening instruction, which may have resulted in the significant gains in this group. The experimental groups' significant improvement could be caused by the MST interventions as well as by routine listening instruction. The two experimental groups, self-directed and teacher-led, both showed higher means than the control group, and the self-directed group showed the highest development on the planning and evaluation factor. Additionally, the planning and evaluation strategies were directly taught in the MLC model, and were measured by the MALQ as a merged factor. This may illustrate that the matching of training content and assessment tools could lead to significant differences between the participants' pre- and post-tests in the MALQ.

Furthermore, the *t*-test result showed that there was a significant difference on the development of the direct attention factor in the teacher-led group only. A possible explanation for this might be that the importance of the monitoring strategy was particularly emphasized in the teacher-led MST workshops. The monitoring strategy in the MLC is focused on two parts of the listening process: Attention focusing and meaning processing. The first, attention focusing, is very close to the DA factor assessed in the MALQ. When introducing the monitoring strategy, the researcher highlighted that this is the most important but most difficult metacognitive strategy for the listening skill. The special emphasis on the importance of this strategy might have attracted the teacher-led participants' attention to it. This also further speaks to the positive effects on MST that the matching of MST content and assessment tool could bring. On the other hand, compared with the PE factor, the difference on the DA factor between the teacher-led group and the other two groups may indicate that students cannot develop the monitoring strategy spontaneously or through self-directed learning, but teacher-led workshops with special emphasis on the strategy will help students raise awareness in this area.

The significant differences on PE and DA reflected in the *t*-test show that the interventions with the two MST methods, self-directed and teacher-led, have a certain effect on the metacognitive awareness development, but not on cognitive skills, such as the problem-solving factor in the MALQ, because it was not included in the MLC training content. Literature reviews indicate that promoting metacognitive awareness can improve listening performance and proficiency. However, since there were no significant differences in metacognitive awareness among the three participant groups, it may very well be that the three groups were actually not different with respect to the other dependent variables in this study, and, therefore, we would not expect any significant effects in RQ2 and RQ3.

4.2. RQ2 Findings

The ANOVA results of RQ2 indicated that there were no significant differences in listening performance gains among the three participant groups, although there were significant differences between pre- and post-tests across the three groups. A Pearson correlation analysis showed that there was no significant correlation between participants' listening performance gains and their metacognitive awareness development. This finding is consistent with those previous studies that did not discover significant effects of MST (Diebold 2011; Leary 1999; Prestwich 2008; Sterling 2011), but differs from the research showing the significant effects of MST (Coskun 2010; Fan 2009; Movahed 2014; Wang 2009).

Fan's (2009) and Wang's (2009) studies, for example, revealed significant effects when using MST combined with cognitive strategy training. Fan (2009) explored the impact of MST on ESL learners' reading comprehension with 143 first-year university students. Her study showed that the participants receiving MST performed better on a reading comprehension test than the students without MST. The three metacognitive strategies taught in her MST were think-aloud, text structure, and summarization. Rigorously speaking, text structure and summarization are not metacognitive strategies, but reading cognitive strategies. Similarly, Wang (2009) employed a sequential mixed-method research design to investigate the effects of MST on ESL high school students' reading comprehension, strategies awareness, and motivation. Wang's (2009) dissertation presented a strong research design, but the core content of her experimental treatment were basic reading strategies rather than the metacognitive strategies she originally proposed and presented in her title. Fan's (2009) and Wang's (2009) studies reminded the researcher that MST with a stress on cognitive strategies could result in more positive significant effects on participants' cognitive skills. However, the MST intervention in this present study focused on metacognitive strategies without integrating any cognitive strategies. This could reveal that a MST focused exclusively on metacognitive strategies without integrating cognitive strategies may have a limited impact on listening performance, because listening comprehension is a cognitive task, and therefore requires cognitive strategy application.

The provision of explicit instruction combining metacognitive and cognitive strategies for MST has been advocated by a number of scholars (Schraw 1998; Veenman et al. 2006; Schneider 2008). Leopold and Leutner (2015) conducted three experiments to compare the effects of cognitive-only strategy training and cognitive and metacognitive combined strategy training on students' learning using scientific texts. Their results showed that the combined training helped students improve performance; the cognitive-only strategy training was not effective. This combined MST effect was proven across their three experiments. Based on the previous studies and the findings of RQ2, this present study makes a new proposal for future research, that effectiveness in MST may be improved by combining it with cognitive strategy training. The cognitive strategies could include pre-listening prediction, skipping unknown parts while listening, and post-listening information compensation strategies. These cognitive strategies can be combined with corresponding metacognitive strategies, for example, prediction can be introduced into the planning strategy, the skipping strategy integrated with the monitoring strategy, and the compensation strategy included in the evaluation strategy.

Regarding the results of non-significant differences between the self-directed and teacher-led groups, this present study is different from Ball (1998) and Manning (2003), but echoes Leary's (1999), Prestwich's (2008), and Sterling's (2011). Sterling (2011) examined the effects of teacher-led and peer-led MST on community college students' achievement in English, finding that there was no statistical difference between the two methods.

4.3. RQ3 Findings

The RQ3 results showed that there was no significant difference in DLPT listening proficiency among the three participant groups. This indicates that the different MST methods did not produce significant differences in listening proficiency in this study.

As with the RQ1 and RQ2 findings, the results for RQ3 might be attributed to insufficient training time and the ineffectiveness of the MST. The RQ1 results showed that there were no significant

differences in metacognitive awareness development among the three participant groups. This means the three groups were not different when it comes to metacognitive awareness, and, therefore, one might expect no significant differences in the related dependent variables, including listening proficiency, among the three groups. As discussed in the RQ2 findings, the MST curriculum focused only on metacognitive strategy training and did not provide cognitive strategy training. The MST alone, without connecting corresponding cognitive strategy in training, could have a very limited impact on listening proficiency.

The DLPT, a measure of language proficiency, is not closely aligned with the MST in this study—the training of generic metacognitive strategies. Language proficiency is a comprehensive result of language learning and is influenced by a combination of factors. This quasi-experiment was conducted in the existing classes that had different students, teachers, teaching materials, and daily teaching schedules. These factors might be relevant to the participants' listening proficiency. The participants in the three groups might have different motivations for improving their listening proficiency and adopting varied listening practices. The teachers in the three groups may have different knowledge and teaching skills about listening instruction and metacognition. The participating classes had different teacher to student ratios and utilized different teaching materials and schedules during the experiment. These uncontrollable factors may have potentially influenced participants' listening performance gains and proficiency results. The MST was not necessarily the only factor that impacted the listening performance and proficiency results.

Interestingly, the correlation analysis of RQ3 showed that participants' DLPT listening proficiency negatively correlated with the PE and MT factors. This means that the participants who use the PE less or the MT more could obtain higher results in the DLPT listening test. Within the MALQ, the MT factor represents the translation strategies that skillful listeners must avoid (Vandergrift et al. 2006). In this study, all three items (4, 11, 18) representing the MT factor were reverse coded before data analysis; therefore, the negative correlation indicates that the more MT is used, the higher the listening proficiency, which is consistent with Goh and Hu (2014) findings. In their multiple regression analysis with 113 participants, they also found that there was a significant negative correlation between the MT factor and listening proficiency test results. Goh and Hu explained that this correlation might relate to their participants' lack of vocabulary and inability to recognize the sounds of words while listening. The repeated negative correlation results in this present study may suggest that the relationship between translation strategy and listening proficiency is not as simple as commonly thought when speaking of metacognition, and needs further research. This suggestion can be supported by Liu's (2008) study results, where the translation strategy showed variations between the most and least efficient listeners among advanced, intermediate, and elementary levels.

5. Conclusions

In this study, three research questions were not upheld by statistical results. The results indicate that MST does not necessarily show statistically significant effects on metacognitive awareness development, listening performance gains, and proficiency test results. The non-significant results may be attributed to the instruments, MST duration and intensity, and some uncontrollable factors related to the participating classes in a routine language course.

Despite the fact that the results were not as expected, this study answered some unresolved questions and may reveal important principles for future MST design. This study demonstrates that the MST using training materials different from the assessment tool, such as the MLC model in this study, may not necessarily achieve the same significant effects as those with the MALQ as training materials. The results do indicate that the self-directed MST enhances development of students' planning and evaluation awareness more than the teacher-led MST and non-training. It was also found that students could not develop their monitoring strategy spontaneously or through self-directed learning, but teacher-led MST workshops with special emphasis on the importance of monitoring strategy will help students evolve in this area. With respect to previous MST studies, the findings

of this study reveal that insufficient training time and the MST without the integration of cognitive strategies do not yield significant effects. Therefore, it is suggested that future MSTs should last for a longer period, with sufficient training time and effective follow-ups to ensure its positive effects. In addition, this study proposes a hypothesis for further research: MST effectiveness could be improved by combining it with cognitive strategy training.

Funding: This research received no external funding.

Acknowledgments: The author would like to sincerely thank the guest editor and anonymous reviewers for their constructive feedbacks and valuable inputs. All errors remain her own.

Conflicts of Interest: The author declares no conflict of interest.

Disclaimer: The content of this article is the sole responsibility of the author and is not necessarily the official views of, or endorsed by the author's affiliation.

Appendix A

Table A1. A sample of application worksheets for the self-directed group.

Week 1: Self-diagnosis Learning Objectives: 1. Identify your listening problems. 2. Find possible reasons of the problems. Tasks: Follow instructions of each day, complete assigned activities with 15 min daily.			
Instructions	Problems Identified	Possible Reasons	Notes
Day 1: Review the LSDA. Identify and summarize your problems and possible reasons with the LSDA. Fill out the right columns.			
Day 2: Listen to a text in homework. After listening, fill out the right columns. Pay attention to linguistic problems.			
Day 3: Listen to a text in homework. After listening, fill out the right columns. Pay attention to your listening strategies.			
Day 4: Listen to a text in homework. After listening, fill out the right columns. Pay attention to your management strategies, anxieties, and nervousness.			
Day 5: Listen to a text in homework. Check the problems identified and summarize patterns.			

Appendix B

Table A2. Application worksheet for the teacher-led group.

Date Length	What Activity (circle one or more as actual use)	Strategies Used (circle one or more as actual use)	Notes
MM/DD mins	Homework Self-study Preview Review Others	Self-diagnosis Planning Time-management Monitoring Evaluation Regulation	

Appendix C

This rubric was used for quantifying experimental participants' strategy applications recorded on their worksheets in the MST. The strategy applications were scored based on the quantity and quality of strategies used on each day. Quantity is the number of the strategies used on each day. Quality represents how well the strategy is used. Since each day is represented by one row on the recording worksheets, the following rubric is the scoring standard for each row.

Table A3. Scoring rubric for strategy use on the application worksheets.

Quantity	None (0 point)	Use (1 point)			Max. points for each row
	No application record in any column on each row.	Shows one application record in any one column on each row.			1
Quality	None (0 point)	Fair (1 point)	Good (2 points)	Excellent (3 points)	Max. points for each row
	No application record in any column on each row.	Shows an application record in any one column on each row.	Shows application records in any two columns on each row. Or shows an application record in any one column on each row with detailed description or multiple circles.	Shows application records in any three columns on each row. Or shows records in any two columns on each row with detailed description or multiple circles.	3
Total points	0	2	3	4	4

References

Abdelhafez, Ahmed M. M. 2006. The effect of a suggested training program in some metacognitive language learning strategies on developing listening and reading comprehension of university EFL students. In *ERIC Online Submission*. Available online: https://eric.ed.gov/?id=ED498262 (accessed on 16 April 2016).

Altuwairesh, Nasrin. 2013. Expertise in L2 Listening: Metacognitive Instruction and Deliberate Practice in a Saudi University Context. Ph.D. dissertation, University of Leeds, Leeds, UK.

Anderson, Neil J. 2002. The role of metacognition in second language teaching and learning. In *ERIC Digest*; ERIC Identifier: ED463659. Available online: files.eric.ed.gov/fulltext/ED463659.pdf (accessed on 15 May 2020).

Babbie, Earl. 2007. *The Practice of Social Research*, 11th ed. Belmont: Wadsworth/Thomson.

Ball, Marjann K. 1998. The Effects of Thinking Maps on Reading Scores of Traditional and Nontraditional College Students. Ph.D. dissertation, University of Southern Mississippi, Hattiesburg, MP, USA.

Birjandi, Parviz, and Amir Hossein Rahimi. 2012. The effect of metacognitive strategy instruction on the listening performance of EFL students. *International Journal of Linguistics* 4: 495–17. [CrossRef]

Cai, Wei. 2013. Investigating second language listening: Factors affecting Chinese listening and the effect of language heritage. *Journal of Chinese Language Teachers Association* 48: 67–97.

Chamot, Anna Uhl, and J. Michael O'Malley. 1994. *The CALLA Handbook: Implementing the Cognitive Academic Language Learning Approach*. New York: Addison-Wesley Publishing Company.

Chang, Cecilia. 2010. See how they read: An investigation into the cognitive and metacognitive strategies of nonnative readers of Chinese. In *Research among Learners of Chinese as a Foreign Language*. Edited by Michael Everson and Helen Shen. Honolulu: University of Hawaii, National Foreign Language Resource Center (NFLRC), pp. 93–116.

Chang, Ching, and Chih-Kai Chang. 2014. Developing students' listening metacognitive strategies using online videotext self-dictation-generation learning activity. *The EuroCALL Review* 22: 3–19. [CrossRef]

Chang, Anna C., and John Read. 2006. The effects of listening support on the listening performance of EFL learners. *TESOL Quarterly* 40: 375–97. [CrossRef]

Coskun, Abdullah. 2010. The effect of metacognitive strategy training on the listening performance of beginner students. *Novitas-ROYAL (Research on Youth and Language)* 4: 35–50.

Creswell, John W. 2008. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 3rd ed. Newbury Park: Sage publications, p. 148.

Diebold, Tamara W. 2011. Relationship between Metacognitive Strategy Instruction and Reading Comprehension in At-risk Fourth Grade Students. Master's thesis, Walden University, Minneapolis, MN, USA.

Fan, Hsiu-Chiao S. 2009. The Effectiveness of Metacognitive Strategies in Facilitating Taiwanese University Learners in EFL Reading Comprehension. Ph.D. dissertation, University of Kansas, Lawrence, KS, USA.

Field, John. 2008. *Listening in the Language Classroom*. Cambridge: Cambridge University Press.

Flavell, John H. 1976. Metacognitive aspects of problem-solving. In *The Nature of Intelligence*. Edited by Lauren B Resnick. Hillsdale: Erlbaum, pp. 231–36.

Flavell, John H. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist* 34: 906–11. [CrossRef]

Gagen-Lanning, Kelsey. 2015. The Effects of Metacognitive Strategy Training on ESL Learners' Self-directed Use of TED Talk Videos for Second Language Listening. Master's thesis, Iowa State University, Ames, IA, USA.

Goh, Christine. 1999. How much do learners know about the factors that influence their listening comprehension? *Hong Kong Journal of Applied Linguistics* 4: 17–40.

Goh, Christine. 2000. A cognitive perspective on language learners' listening comprehension problems. *System* 28: 55–75. [CrossRef]

Goh, Christine. 2008. Metacognitive instruction for second language listening development: Theory, practice, and research implications. *RELC Journal* 39: 188–13. [CrossRef]

Goh, Christine, and Guangwei Hu. 2014. Exploring the relationship between metacognitive awareness and listening performance with questionnaire data. *Language Awareness* 23: 255–74. [CrossRef]

Graham, Suzanne. 2002. Experiences of learning French: A snapshot at Years 11, 12 and 13. *Language Learning Journal* 1: 15–20. [CrossRef]

Graham, Suzanne. 2006. Listening comprehension: The learners' perspective. *System* 34: 165–82. [CrossRef]

Graham, Suzanne, and Ernesto Macaro. 2008. Strategy instruction in listening for lower-intermediate learners of French. *Language Learning* 58: 747–83. [CrossRef]

Graham, Suzanne, Denise Santos, and Robert Vanderplank. 2008. Listening comprehension and strategy use: A longitudinal exploration. *System* 36: 52–68. [CrossRef]

Graham, Suzanne, Denise Santos, and Ellie Francis-Brophy. 2014. Teacher beliefs about listening in a foreign language. *Teaching and Teacher Education* 40: 44–60. [CrossRef]

Leary, Samuel Ferebee, Jr. 1999. The Effect of Thinking Maps Instruction on the Achievement of Fourth-grade Student. Ph.D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Leopold, Claudia, and Detlev Leutner. 2015. Improving students' science text comprehension through metacognitive self-regulation when applying learning strategies. *Metacognition and Learning* 10: 313–46. [CrossRef]

Liu, Hsueh-Jui. 2008. A study of the interrelationship between listening strategy use, listening proficiency levels, and learning style. *Annual Review of Education, Communication & Language Sciences* 5: 84–104.

Luo, Xiaorong, and Jian Gao. 2012. On the existing status in listening teaching and some suggestions for it. *Theory and Practice in Language Studies* 6: 1270. [CrossRef]

Manning, Cynthia. 2003. Improving Reading Comprehension through Visual Tools. Master's thesis, Eastern Nazarene College, Quincy, MA, USA.

Movahed, Roya. 2014. The effect of metacognitive strategy instruction on listening performance, metaconitive awareness and listening anxiety of beginner Iranian EFL students. *International Journal of English Linguistics* 4: 88–99. [CrossRef]

Nosratinia, Mania, Samira Ghavidel, and Alireza Zaker. 2015. Teaching metacognitive strategies through Anderson's model: Does it affect EFL learners' listening comprehension? *Theory and Practice in Language Studies* 6: 1233–43. [CrossRef]

O'Bryan, Anne, and Volker Hegelheimer. 2009. Using a mixed methods approach to explore strategies, metacognitive awareness and the effects of task design on listening development. *Canadian Journal of Applied Linguistics/Revue Canadienne de Linguistique Appliquée* 12: 9–38.

Oxford, Rebecca L. 1990. *Language Learning Strategies: What Every Teacher Should Know*. New York: Newbury House.

Oxford, Rebecca L. 1993. Research update on teaching L2 listening. *System* 21: 205–11. [CrossRef]

Prestwich, Dorothy L. 2008. Effects of Linguistic or Non-linguistic Cognitive Maps on Fourth Grade Students' Reading Comprehension. Ph.D. dissertation, University of Mississippi, Mississippi, MS, USA.

Prince, Peter. 2013. Listening, remembering, writing: Exploring the dictogloss task. *Language Teaching Research* 17: 486–500. [CrossRef]

Rasouli, Maliheh, Kambiz Mollakhan, and Alireza Karbalaei. 2013. The effect of metacognitive listening strategy training on listening comprehension in Iranian EFL context. *European Online Journal of Natural and Social Sciences* 2: 115–28.

Richards, Jack C., and Theodore S. Rodgers. 2001. *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.

Rubin, Joan. 1994. A review of second language listening comprehension research. *The Modem Language Journal* 2: 199–221. [CrossRef]

Schneider, Wolfgang. 2008. The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education* 2: 114–21. [CrossRef]

Schraw, Gregory. 1998. Promoting general metacognitive awareness. *Instructional Science* 26: 113–25. [CrossRef]

Smidt, Esther, and Volker Hegelheimer. 2004. Effects of online academic lectures on ESL listening comprehension, incidental vocabulary acquisition, and strategy use. *Computer Assisted Language Learning* 17: 517–56. [CrossRef]

Sterling, Ra Shaunda Vernee. 2011. The Effect of Metacognitive Strategy Instruction on Student Achievement in a Hybrid Developmental English Course. Ph.D. dissertation, University of South Alabama, Mobile, AL, USA.

Vandergrift, Larry. 1997. The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Language Annals* 30: 387–409. [CrossRef]

Vandergrift, Larry. 1999. Facilitating second language listening comprehension: Acquiring successful strategies. *ELT Journal* 53: 168–76. [CrossRef]

Vandergrift, Larry. 2002. 'It was nice to see that our predictions were right': Developing metacognition in L2 listening comprehension. *Canadian Modern Language Review* 58: 555–75. [CrossRef]

Vandergrift, Larry. 2003. Orchestrating strategy use: Toward a model of the skilled second language listener. *Language Learning* 53: 463–96. [CrossRef]

Vandergrift, Larry. 2004. Learning to listen or listening to learn? *Annual Review of Applied Linguistics* 24: 3–25. [CrossRef]

Vandergrift, Larry. 2007. Recent developments in second and foreign language listening comprehension research. *Language Teaching* 40: 191–210. [CrossRef]

Vandergrift, Larry, Christine C. M. Goh, Catherine J. Mareschal, and Marzieh H. Tafaghodtari. 2006. The metacognitive awareness listening questionnaire: Development and validation. *Language Learning* 56: 431–62. [CrossRef]

Veenman, Marcel V. J., Bernadette HAM Van Hout-Wolters, and Peter Afflerbach. 2006. Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning* 1: 3–14. [CrossRef]

Wang, Min-Tzu. 2009. Effects of Metacognitive Reading Strategy Instruction on EFL High School Students' Reading Comprehension, Reading Strategies Awareness, and Reading Motivation. Ph.D. dissertation, University of Florida, Gainesville, FL, USA.

Wenden, Anita L. 1998. Metacognitive knowledge and language learning. *Applied Linguistics* 19: 515–37. [CrossRef]

languages

MDPI

Article

A Longitudinal Study of the Acquisition of the Polysemous Verb 打 *dǎ* in Mandarin Chinese

Jidong Chen * and Xinchun Wang

Department of Linguistics, California State University, Fresno, CA 93720, USA; xinw@mail.fresnostate.edu
* Correspondence: jchen@csufresno.edu

Received: 16 February 2020; Accepted: 13 May 2020; Published: 18 May 2020

Abstract: Lexical ambiguity abounds in languages and multiple one-to-many form-function mappings create challenges for language learners. This study extends the theoretical approaches to the acquisition of polysemy to the Mandarin verb 打 *dǎ*, which is highly polysemous and among the earliest verbs in child speech. It analyzes longitudinal naturalistic data of nine children (1;05–3;10) from two Mandarin child speech corpora to explore the developmental trajectory of different senses of 打 *dǎ* and the role of input. The results support a continuous derivational and restricted monosemy approach: children initially extract a core feature of 打 *dǎ*, but only apply it in a restricted way, reflected in a small number of senses in a limited set of semantic domains and syntactic frames, revealing an early preference for initial unambiguous form-meaning mappings. Mandarin-speaking children's production mirrors the semantic and syntactic distribution of the input, supporting the usage-based approach to the acquisition of polysemy that meaning is derived from the confluence of lexical and syntactic cues in the usage patterns in the input. Our research is the first longitudinal study of the emergence and development of polysemous verbs in Mandarin and has pedagogical implications for teaching Mandarin as a second language.

Keywords: Mandarin; child language acquisition; polysemy; verb semantics

1. Introduction

Lexical ambiguity abounds in languages and multiple one-to-many form-function mappings create challenges for child language learners (e.g., Clark 1993). Lexical ambiguity can arise from polysemy or homonymy (e.g., *male* and *mail*). Polysemy is characterized as a single word associated with multiple related sense in contrast to homonymy, a single form associated with multiple unrelated meanings (e.g., Vicente and Falkum 2017), as illustrated by examples from child language (1) (cited from Tomasello 1992) and (2) (extracted from the Tong corpus, Deng and Yip 2018) below.[1]

[1] CHI = child. The age of the children is conventionally notated as years;months. Utterances are transcribed in Chinese characters and Pinyin, the official Romanized transcription of Chinese.

1. a. CHI (1;04): *Get* it. ('obtain')
   b. CHI (1;05): *Get* out. ('move')

2. a. CHI (1;07): *打* *球*. ('play')
   dǎ *qiú*
   play ball
   'Play with the ball.'
   b. CHI (2;10): *小熊* *打死* *它*. ('kill')
   xiǎoxióng *dǎsǐ* *tā*
   little.bear kill-die it
   'Little bear killed it.'

In (1), the English verb *get* expresses two related senses, meaning 'to obtain' (moving objects towards the prospective possessor in 1a) and 'to move' (moving objects away from a location or possessor in 1b), respectively; and in (2), the Mandarin verb *打 dǎ* means 'to use hand to play' (2a) or 'use hand to destroy" (2b).

Polysemy is pervasive in natural languages, and has attracted much attention in linguistics theoretically and empirically regarding the mental representation, access, and storage of multiple senses in adult language. Cognitive linguistic approaches (e.g., Lakoff 1987) argue for a network representation of multiple senses, where related meanings are connected to a core (prototypical) sense and each meaning extension is motivated in some cognitively natural fashion (e.g., Langacker 1987). This polysemy approach has been criticized as resulting in an over-proliferation of distinct senses that may have implausible correspondence in the speaker's mind (e.g., Sandra and Rice 1995). An alternative approach to polysemy acknowledges the context-dependence of word meanings and argues that multiple senses are contextual variations or elaborations of a single core sense, i.e., computed based on the context (e.g., Allwood 2003; Evans 2005; Tuggy 1993). This approach is in line with the constructional grammar approach to verb semantics and argument structure. For example, Goldberg (1995) argued that verb meanings are often associated with the constructions that they frequently occur in. Other researchers (e.g., Nerlich et al. 2003) suggested that both the lexical-semantic and the grammatical constructional approaches should be combined.

1.1. Acquisition of Polysemy in Child Language

In child language acquisition, it remains a question as to how children arrive at adult-like mental representation of multiple meanings of polysemous words. Clark (1993) argued that children must solve the "mapping problem" in learning the words of a language, i.e., establishing a mapping between a phonological forms and its meaning. Multiple meaning to the same form mapping creates potential problems for child learners. Children have been found to prefer to apply only one meaning to a lexical item (e.g., nouns) at the beginning (Clark 1993), but they also use many highly frequent polysemous verbs freely from as young as two years old (e.g., Clark 1996). Studies of the acquisition of polysemous verbs show that children do not acquire the full range of senses of verbs until late primary school ages (e.g., mental verb *know* in (Booth and Hall 1995)). The acquisition of polysemous verbs thus poses a particular challenge to the mapping problem.

A number of studies of the acquisition of polysemy have examined the developmental trajectory of different senses of a target relational word (e.g., verb or preposition). Nerlich et al. (2003) conducted a cross-sectional study of the semantic knowledge of the polysemous verb *get* in children at the ages of 4, 7, 8, and 10, using elicited production and sense ranking tasks. They found that the semantic development started with the core sense of 'obtain' and moves onto the most distant sense of *get* as 'understand' in overlapping stages. Younger children produced fewer senses than the older children focusing on the senses of 'obtain', 'receive', 'have', and 'fetch'. The semantic knowledge of the 4-year-olds showed general and abstract core meanings and the 10-year-olds had good knowledge of prototypical versus non-prototypical meanings.

McKercher (2001) studied the acquisition of the multiple senses of the preposition *with* (e.g., 'attribute', 'nominal', 'instrument', 'accompaniment', 'manner', etc.) and proposed two different developmental approaches, the monosemy approach and the multiple-meanings (polysemy) approach. The monosemy approach predicts that children would start with an underspecified (or core) representation and use a range of different senses simultaneously from the beginning, and the multiple-meanings approach predicts that children would acquire each sense of *with* item-by-item (e.g., $with_{instrument}$, $with_{accompaniment}$, etc.) and different senses would emerge at a different time. McKercher (2001) analyzed longitudinal data of six English-learning children with varied length of data between the ages of 1;3 and 4;10 from the Child Language Data Exchange System (CHILDES) (MacWhinney 2000). He found that the children produced a range of semantic roles in their early speech and showed a more general meaning of *with* (i.e., 'having') than each of the specific senses, which supported the monosemy approach. Kidd and Cameron-Faulkner (2008) analyzed a dense (one hour five days per week from 2;0.12 to 3;1.30) longitudinal corpus of one English-learning child, and argued that children initially extracted a core feature of *with* (i.e., spatial proximity or co-location) but only use it in a restricted way (i.e., using the core meaning for some time before gradually extending to other senses), which suggested children's preference for an initial one-to-one form-meaning. Kidd and Cameron-Faulkner (2008) also examined in detail the input of the target child and showed that the relative frequency of senses was similar in the child speech and the input and argued that the input offered reliable cues for the uses of different senses, including the semantics of the verbs and the construction in which *with* occurred most often. A similar argument was made in the study of children's acquisition of various forms of the verb *go* longitudinally, where a good predictor of children's usage was the input frequency of *go* in different structures and the specific meanings with particular forms of *go* (Theakston et al. 2002).

1.2. Semantics of the Mandarin Verb 打 dǎ and Its Acquisition

Our study aims to extend the theoretical approaches to the acquisition of polysemy in Mandarin Chinese (henceforth Mandarin), focusing on the Mandarin verb *打 dǎ*. The Mandarin verb *打 dǎ is* 'hit/beat' is one of the most frequently used verbs in Mandarin (e.g., Gao 2001) and is highly polysemous. *The Contemporty Dictionary of Chinese* (2016, 7th edition) lists 24 senses of *打 dǎ*, e.g., *打门 dǎmén* 'knock on door', *打架 dǎjià* 'fight', *打家具 dǎjiājù* 'make furniture'. The basic meaning of *打 dǎ* refers to a physical action of the hand and can extend from its basic prototypical meaning to a wide range of actions or events involving a hand or instrument as well as to events that are metaphorically hand-involving (e.g., *打折 dǎzhé* 'discount', *打听 dǎtīng* 'inquire'). The Chinese Wordnet (Huang et al. 2010) lists 121 senses of *打 dǎ* based on a detailed lexical semantic analysis ranging from concrete actions involving hand manipulation (e.g., *打桌子 dǎzhuōzi* 'knock on table') to metaphorically extended senses such as *打天下 dǎtiānxià* 'establish power', *打票 dǎpiào* 'buy tickets', *打光 dǎguāng* 'polish'. These dictionary-based depictions provide a descriptive picture of the senses of *打 dǎ*, but do not offer a systematic categorization and characterization of the many senses and the derivational relations between them.

Gao's (2001) study of the physical action verbs in Chinese (including *打 dǎ*) filled this gap. It provided a comprehensive analysis of the lexical semantics of *打 dǎ* in a cognitive linguistics framework that grounded verbal semantics and argument structures in the nature of human bodies and their interactions with the physical, social, and cultural environment (cf. Lakoff and John 1999). Chinese physical actions verbs could be characterized in terms of typical bodily experiences including, e.g., the body parts involved (e.g., hand, foot, head, mouth, etc.), physical contact, motion, and intention. Gao (2001) analyzed the distribution of different senses of *打 dǎ* in two large corpora of Mandarin (the Academia Sinica Balanced Corpus and the Beida Institute of Computational Linguistics corpus). She found 152 distinct senses of *打 dǎ* in 27 semantic domains (e.g., game playing, physical punishment, open, fastening, construction, covering, etc.) in five broad semantic representations (categories) that set up "a linkage of all the sub-fields from the descriptions of the most prototypical actions if *打 dǎ*

with hand contact as the focus down to the metaphorical uses with very vague or even non indications of hand actions of any type" (Gao 2001, p. 166). Table 1 presents the five semantic representations and the distinct senses in each semantic category, adapted from Gao (2001, pp. 163–65). Gao also found that the two semantic categories defeat and physical punishment are the most frequent in both corpora, despite the high degree of polysemy.

Table 1. Semantic representations and senses of the target verb *打 dǎ* (Gao 2001).

Semantic Representations	Specific Senses
Type 1: Physical action focusing on hand contact	Defeat (play games, battle, fight, kill, attack, break, smash) Physical punishment (beat, hit, punch, spank, whip, slap, strike) Open (turn on, take out, unpack) Fastening (pack up, knot, tie)
Type 2: Hand action (mostly with instrument)	Construction (make, build, dig, drill, burrow) Launching (shoot, fire, send, report, pump, post, set up, signal) Cover (dress up, pretend, spray, wax, powder, plaster, clout, drench, label, paint, polish, wrap) Insert (inject, nail, hammer, knock) Possess (fetch, catch, have, hunt, ladle, hold) Mark (type, work out, engrave, press, label, print, stamp) Sound source (drum, knock, cap, pound, crow, flap, ring, tap, whistle) Motion (stir, return, move, mix) Reflection (flash, reflect) Collection (gather, reap, sweep, get in, get) Removal (prune, peel, knock out, get rid of, rob, thresh) Engagement (work)
Type 3: Physical action with physical contact unspecified	Physiological reaction (shiver, yawn, doze off, snore, sneeze, cheer up, hiccup, nod, stupefy) Gymnastic feat (roll, tumble, loop) Posture (bare, remain, zazen)
Type 4: Metaphorical uses (hand action traceable)	Social interaction (call, contact with, gesticulate) Business deal (discount, buy, invest) Authoritative conduct (issue, score) Legal activity (go to court)
Type 5: Metaphorical uses (hand action untraceable)	Mental activity (plan, seek, make, think of, calculate, disturb, consider, decide, draw analogy, concern, estimate) Verbalization (ask about, greet, bet, draw, interrupt, talk, chat, cry out, discuss, question, speak) Opposition (defend, protest) Visual contact (meet, bump into, look at)

The verb *打 dǎ* has been found among the earliest verbs in child speech and input frequency and children's physical development, growth environment, and cognitive understanding of the actions correlate directly with multiple senses (Gao 2015). Little research has been conducted to examine monolingual Mandarin-learning children's semantic acquisition of polysemous verbs. Two recent studies were most relevant to our inquiry of the acquisition of polysemous verbs in Mandarin. Sak and Gao (2016) examined the semantic knowledge of preschool bilingual Mandarin-English children in Singapore using elicited descriptions of pictures or videos depicting *打 dǎ* actions. They analyzed 31 senses in the data and found that the semantic domains of "social interactions" (e.g., *打电话 dǎdiànhuà* 'call') and "physical punishment" (e.g., *打耳光 dǎěrguāng* 'slap in the face') are the most frequent, whereas semantic domains such as "fastening" and "possession" are least used. English was found to negatively affect the bilingual children's use of *打 dǎ*. It is argued that children's exposure to action events and the competition of near synonyms of *打 dǎ* accounted for the observed usage.

Zhang et al. (2010) examined the acquisition of eight polysemous words, among which three were polysemous verbs (i.e., *看 kàn* 'look', *走 zǒu* 'walk', and *给 gěi* 'give'), in a longitudinal corpus (weekly one-hour recordings) of a young monolingual Mandarin-learning child from age 1;6 to 3;0. They proposed three possible approaches for learning polysemous words: (1) a continuous approach, where multiple senses could be derived via metaphor or metonymy; (2) an independent approach, where each sense was acquired independently without derivational relations; and (3) a mixed strategy using both (1) and (2). Their analyses of the longitudinal emergence of the different senses of the three verbs supported the continuous derivational approach, and the derivational routes could proceed in three

possible ways, i.e., radiationally (i.e., multiple senses derived from one single sense simultaneously), serially (i.e., new senses derived from the previously acquired senses), or both radiationally and serially. The results also suggest that derivational routes could vary from verb to verb, subject to usage factors such as input frequency and functional needs: the verb 看 *kàn* 'look' showed a serial route where the basic sense emerged first and different derived senses appeared longitudinally, whereas the other two verbs, 走 *zǒu* 'walk' and 给 *gěi* 'give', showed an independent route, where multiple senses appeared simultaneously around a similar time, and the derived sense could appear early (e.g., the sense 'leaving' for 走 *zǒu* appeared before the prototypical sense 'walk'). Since 打 *dǎ* was not examined, it remains an empirical question if and how its acquisition fits in the proposed developmental routes in Zhang et al. (2010).

1.3. Research Questions

The current study aims to examine monolingual Mandarin-learning children's semantic acquisition of the polysemous verb 打 *dǎ*. Based on prior research, the acquisition of the multiple senses of 打 *dǎ* could potentially proceed in three different routes, following the monosemy approach (McKercher 2001), the restricted monosemy approach (Kidd and Cameron-Faulkner 2008), or the polysemy approach (similar to the independent approach proposed in Zhang et al. 2010). The continuous strategy in Zhang et al. (2010) shares with the monosemy approach and the restricted monosemy approach the emphasis on the derivational relations between all senses, but they differ in the claim about the starting point of the sense derivation in acquisition: the basic core or functionally most salient and frequent sense (Zhang et al. 2010), the basic core sense with derived senses fully accessible simultaneously (McKercher 2001), or partially accessible with restricted uses and slow extension to other senses (Kidd and Cameron-Faulkner 2008). It is thus interesting to investigate empirically how multiple senses of 打 *dǎ* develop longitudinally to evaluate the different approaches. This study aims to answer the following research questions:

1. What is the developmental trajectory (i.e., emergent order) of different senses of 打 *dǎ* in Mandarin?
2. How do Mandarin-learning children proceed in acquiring different senses of 打 *dǎ* in a multiple-meanings (polysemy) approach, a monosemy approach, or the restricted multiple-meanings (polysemy) approach?
3. How does input, including syntactic, semantic, and contextual cues, contribute to the acquisition of different senses of 打 *dǎ*?

2. Materials and Methods

We analyzed longitudinal naturalistic corpus data of 9 children (age range 1;05–3;10) in two Mandarin child corpora, one child from the Tong corpus (Deng and Yip 2018; MacWhinney 2000) and eight children from the Taiwan Corpus of Mandarin Chinese (TCMC) (MacWhinney 2000). The Tong corpus contains hour-long monthly recordings of interactions between the target Mandarin-learning child Tong and his caregivers (mostly mother, father and occasionally grandparents) from age 1;07 to 3;4, including a total of 22 transcripts (see Table 2), the largest dataset among the 9 children. The TCMC contains monthly naturalistic data from 10 children, age range from 1;05–4;03 with varied ages of the start and the end of data collection, and length of data recordings (see Table 1). The data from eight of the children were included in the analysis due to their early age (between 1;05 and 2;07) at the start of the data collection to explore the early emergence of the use of 打 *dǎ* and 2 children who were at 3;01 and 3;6 at the start of the data collection were excluded. A total of 375 tokens of 打 *dǎ* were produced in the child speech and a total of 809 tokens of 打 *dǎ* in the adult speech.

All utterances containing the target verb 打 *dǎ* in the children's speech were extracted with the Computerized Language Analysis (CLAN) program (MacWhinney 2000), and sample early utterances that contain 打 *dǎ* are shown in Table 3. At least 3 utterances above and below the target utterance were also extracted with the CLAN program to provide contextual information to determine the specific

senses. The verb 打 *dǎ* in the target utterances were coded for (1) the type of semantic representations and (2) the specific sense based on Gao's (2001) lexical semantic analysis of 打 *dǎ* (cf. Table 1).

Table 2. Information about the target children in the longitudinal naturalistic corpora.

	Child	Age Range	# Files	# Utterances	Types	Tokens	Token Freq. of 打 *dǎ* in child Speech	Token Freq. of 打 *dǎ* in Input
1	Tong	1;07–3;04	22	9110	5240	30,798	175	294
2	Chou	2;01–3;04	16	4991	1482	14,025	33	44
3	JC	2;02–3;05	14	4955	1183	10,710	22	90
4	Pan	1;07–3;09	19	2661	942	7634	40	129
5	Wang	2;05–3;04	12	3149	1092	10,133	19	44
6	Wu	1;07–2;01	12	2785	922	7030	15	92
7	Wuys	2;07–3;10	10	1396	758	5223	23	40
8	Xu	1;06–2;05	11	2700	651	4315	14	7
9	Yang	1;05–2;09	13	1929	733	4787	34	69
						Total	375	809

Table 3. Sample utterances containing the target verb 打 *dǎ* in child speech[2].

Child Age	Samples	Pinyin	Glossing	Translation
1;07.18	打开啦。	*dǎkāla*	hit open SFP	'Opened!'
1;07.18	打球。	*dǎqiú*	play ball	'Play with the ball.'
1;08.22	打牌。	*dǎpái*	play card	'Play cards.'
1;11.21	打妈妈。	*dǎmāma*	hit mommy	'Hit mommy.'
2;00.19	打锣了。	*dǎluó*	play gong SFP	'Play the gong.'
2;00.19	妈妈给打开。	*māmagěi dǎkāi*	mommy give hit-open	'Mommy open (for me).'
2;01.17	我们来打牌。	*wǒménlái dǎ pái*	we come play card	'Let's play cards.'

To explore the relationship between the child's output and input, we further extracted and coded the semantics of the target verb 打 *dǎ* in the caregivers' speech in the Tong corpus, the biggest dataset among the 9 children. We also coded all the target utterances in the speech of both Tong and his caregivers for the syntactic contexts or frames in which 打 *dǎ* occurs, e.g., VV (verb compound), VNP (verb followed by an object noun phrase), NPVNP (subject noun phrase followed by the verb and an object noun phrase), to see if the emergence of the multiple senses of 打 *dǎ* are closely tied to certain syntactic frames or constructions in the child and the caregivers' speech. The rationale for this additional construction-based analysis comes from the findings that children are sensitive to the syntactic frames in which a verb occurs and use them to infer verb meanings (e.g., Gleitman 1990; Lee and Naigles 2005), and surrounding linguistic context plays a role in the sense distribution of verbs (e.g., Theakston et al. 2002).

Two native Chinese-speaking authors coded the semantics of all the target utterances in the child speech and the input independently. Both authors checked each other's coding and any discrepancies were resolved and agreed upon on a case-by-case basis. Where utterances were ambiguous, linguistic context of the target utterances (i.e., preceding and following utterances) was used to determine the meaning in the corresponding transcripts. The intercoder agreement was high, with 98% agreement.

3. Results

We analyzed the semantic distribution of the total of 375 tokens of 打 *dǎ* in the child speech and the total of 809 tokens of 打 *dǎ* in the adult speech. All the children were found to have started to produce 打 *dǎ* at a young age, around the first or second data session of each individual child, between the ages of 1;06 and 2;07. The verb 打 *dǎ* is among the top 10 most frequent verbs in the children's speech. In the sections below, we present the overall distribution of the different senses of 打 *dǎ*, followed by

[2] SFP = sentence final particle.

an examination of the earliest emergent uses (i.e., the first 10 tokens) and the longitudinal emergence of different senses across the nine children. Next, we focus on the comparison of the distributional patterns between the child Tong and his input.

3.1. Distribution and Emergence of Different Semantic Categories and Senses of 打 dǎ

The token frequencies of *打 dǎ* were calculated by the semantic representation types. Figure 1 presents the overall proportions of the different semantic categories of *打 dǎ* across all nine children. The predominant type of semantic representation for all the children is Type 1 (physical action focusing on hand contact) with a mean proportion of 93% (ranging from 75% to 100%). Four out of the nine children (JC, Pan, Wu, and Xu) used only the Type 1 meaning (100%). Type 2 (hand action mostly with instrument) and Type 4 (metaphoric uses with hand action traceable) were used with a low mean proportion of 1.78% (Type 2) and 4.6% (Type 4), with the latter type being mostly produced by one child, Yang (24% of his usage), to refer to only one specific event of calling by phone. Types 3 (physical action with physical contact unspecified) and 5 (metaphoric uses with hand action untraceable) meanings were used minimally across the children with a mean proportion of 0.06% (Type 3) and 0.13% (Type 5), respectively. This suggests that the core meaning of *打 dǎ* as a physical action involving hand contact is acquired early and used prevalently among the multiple senses, and metaphoric uses with hand action traceable seem to emerge before untraceable hand action. Less concrete meanings seem to be produced less and at later ages.

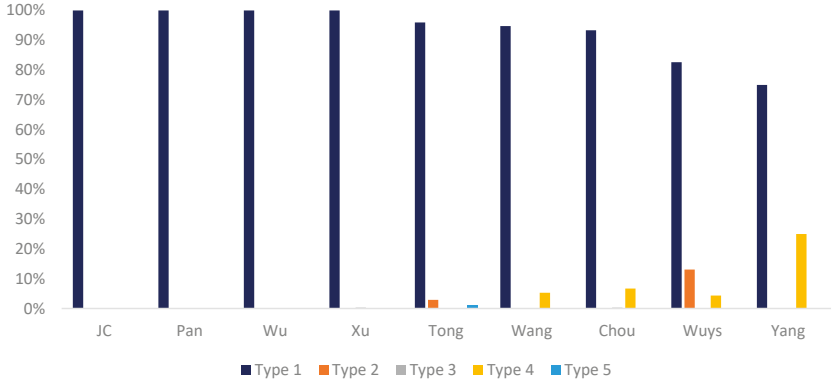


Figure 1. Proportions of different types of semantic categories of *打 dǎ* by individual child. Type 1: Physical action focusing on hand contact; Type 2: hand action (mostly with instrument); Type 3: physical action with physical contact unspecified; Type 4: metaphorical uses (hand action traceable); Type 5: metaphorical uses (hand action untraceable).

We also tallied the token frequencies of *打 dǎ* by different senses and calculated the overall proportions of the senses that were used more than once (i.e., a minimum of 2% usage in each child's speech). Figure 2 shows the proportions of different senses by individual child. For the ease of interpreting the distribution, color coding is used to indicate the semantic categories in which the senses belong to: various shades of blue indicate Type 1 senses; different shades of green indicate Type 2 senses, and light yellow indicates senses in Type 4. As shown in Figure 2, the majority of the different senses (75–100%) involves Type 1 physical action with hand contact ('hit', 'spank', 'fight', 'break', and 'open'), among which the sense open dominates (mean proportion 41%). Type 2 senses, hand action with instrument ('play', 'shoot', and 'hold/possess'), account for a small portion of usage (mean proportion 2.7%) and Type 4 sense, call (metaphoric uses with traceable hand action), is used by four children (mean proportion 4.5%), with the most frequent uses from one child, Yang (24%).

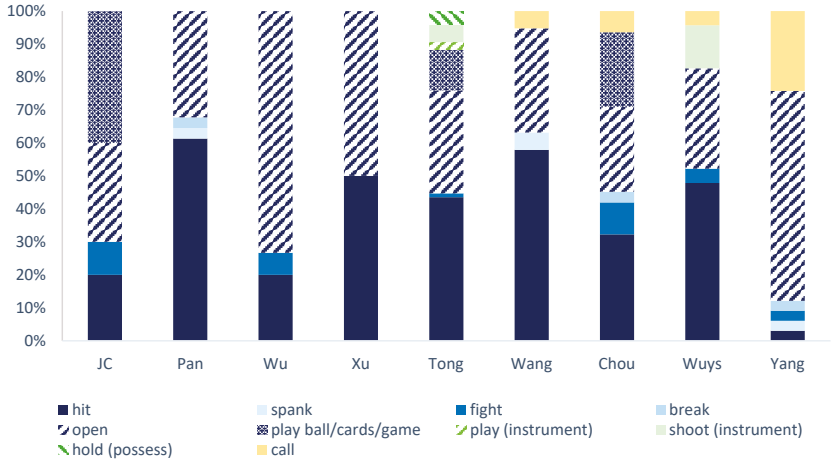


Figure 2. Proportions of different senses of *打 dǎ* by individual child[3].

To examine the earliest uses of *打 dǎ* (emergent categories and senses), the first 10 tokens from each child were analyzed. The distribution is shown in Figure 3. Similar to the overall distribution, the most frequent senses in the earliest tokens of *打 dǎ* involve physical action of hand contact, i.e., 'open' (40%) and 'hit' (31%), followed by metaphorical use of traceable hand action (i.e., 'call', 7.8%) and three senses ('fight', 'play games', and 'spank') that also involve physical action of hand contact. A further examination of the two most frequent senses 'open' and 'hit' shows that both senses occur in specific syntactic and semantic contexts: 100% of all the utterances of the 'open' sense occur in a resultative verb compound 打开 dǎkāi 'hand.action-open' with an animate agent and an inanimate patient referents, and 100% of the utterances of the 'hit' sense occur in a transitive sentence frame with animate agent and patient referents. (e.g., 我打妈妈 wǒ dǎ māma 'I hit mom'). A similar pattern is observed in the sense 'call', which occurs 100% with an object NP 电话 diànhuà 'phone' and an animate agent (caller). The sense 'fight' occurs only in a compound verb 打架 dǎjià; and the sense 'play games' is also used only with inanimate objects (e.g., cards, ball) and an animate agent.

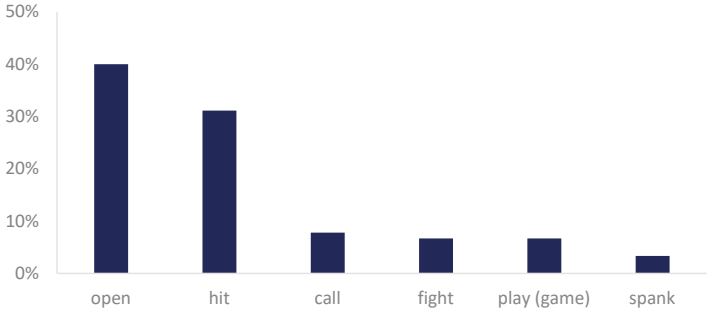


Figure 3. Proportions of the senses of *打 dǎ* in the first 10 tokens by all the children.

[3] Examples illustrating the different senses in the child speech: *hit: 打人 dǎrén 'hit a person'; spank: 打屁股 dǎpìgǔ; open: 打开 dǎkāi; fight: 打架 dǎjià; play (ball): 打球 dǎqiú; play (game): 打牌 dǎpái; play (instrument): 打鼓 'play the drum'; shoot: 打枪; hold: 打伞 dǎsǎn 'hold an umbrella'; call: 打电话; mark: 打勾 'mark with a tick'; knot: 打结 dǎjié 'make a knot'; thunder strike: 打雷 dǎléi; type: 打字, 'type'; doze: 打盹 dǎdǔn 'doze off'; sneeze: 打喷嚏 dǎpēnti 'sneeze'; launch: 打到 dǎdào 'launch (to)'.*

3.2. Longitudinal Development of Different Senses of 打 dǎ

We further examined the longitudinal development of different senses of *打 dǎ*. We first present the results from the speech of the child Tong, whose dataset is the largest among the nine children (cf Table 2). Tong produced *打 dǎ* in 15 out of the 22 transcript files. Table 4 summarizes the token frequencies of the different senses by age. A total of 10 different senses show up in Tong's speech and most involve the physical action of hand contact. The top five senses, 'hit', 'open', 'play (games)', 'shoot', and 'launch', account for 94% of the uses of *打 dǎ*. The earliest production of *打 dǎ* at 1;07 includes two senses involving physical action of hand contact, 'open' and 'play (games)', suggesting that multiple senses can emerge at the same time. Three new senses, 'hit', 'launch', and 'shoot', emerged simultaneously at 1;11, among which 'hit' was the most frequent sense (37%) in Tong's speech from 1;07 to 3;04. The sense 'open' is the second most frequent produced sense (30%) overall. The production of the top two senses is consistent with the general pattern observed across the other children that 'open' and 'hit' are the most dominant senses in early Mandarin-learning children's speech. Newly emerged senses are also used with previously used senses, e.g., 'shoot', 'hit', and 'launch' emerged with the previously used sense *open* at 1;11. Overall, Tong used at least two or more different senses in 13 out of the 15 transcripts, indicating that the majority of his uses of *打 dǎ* are polysemous (80%).

Table 4. Token frequencies of different senses of *打 dǎ* in the child Tong's speech by age.

Senses\Age	1;07	1;08	1;11	2;00	2;01	2;02	2;03	2;05	2;06	2;07	2;09	2;10	3;01	3;03	3;04	Total	prop.
hit (person/object)	0	0	7	2	4	0	6	5	3	4	1	31	0	0	1	64	37%
open	1	0	4	1	1	1	0	0	0	0	0	22	13	6	3	52	30%
play (games)	2	2	0	0	11	0	8	0	0	0	0	0	0	0	2	25	14%
shoot	0	0	4	0	0	0	0	1	1	0	0	0	0	0	4	10	6%
launch	0	0	2	0	0	1	4	0	2	0	0	0	0	0	0	9	5%
hold	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	7	4%
fight	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	1%
play (instrument)	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	2	1%
knot	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	1%
thunder strike	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1%
mark	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1%

Do other children show similar developmental trajectories? Table 5 presents longitudinally the different senses produced by Tong and the eight children in the TCMC corpus. Compared with Tong, fewer senses were produced, which is probably due to the lower frequency of data collection and fewer tokens of *打 dǎ* in each child's speech in the TCMC corpus (see Table 2). The early production (1;06–2;01) of *打 dǎ* is limited to a single sense for the children in the TCMC corpus, and two senses, 'open' and 'hit', are the most frequent. The sense 'sneeze' appeared once for the child Xu at age of 1;06, and the sense call appeared three times in the speech of one child Yang at 1;10. Multiple senses were used simultaneously from age 2;02 and a limited number of new senses emerged gradually, including 'play games' (2;02), 'break' (2;02), 'spank' (2;02), 'type' (2;02), 'fight' (2;03), 'doze' (2;04), 'turn on' (2;09), 'hold' (2;10), 'shoot' (3;0), and 'sweep' (3;03). The most frequent senses are 'open' and 'hit' through ages 2;0–3;10. The developmental pattern in the eight children from the TCMC corpus is thus similar to that in the speech of Tong. Individual difference is also observed in the specific senses produced, which could tie to the particular contexts of the speech produced (e.g., 'sneeze', 'call', or 'type').

Table 5. Longitudinal emergence of different senses of 打 *dǎ* in all children's speech, (numbers in parentheses indicates token frequency; *play* = *play games*; grey areas indicate gaps in data collection or no production of 打 *dǎ*).

Age/Child	Tong	Yang	Xu	Pan	Wu	Chou	JC	Wang	Wuys
1;06			sneeze (1)						
1;07	open (1), play (2)								
1;08	play (2)		open (2)						
1;09				open (4)	open (1)				
1;10		call (3)	hit (3)						
1;11	hit (7), open (4), shoot (4), launch (2)								
2;0	hit (2), open (1), play instrument (2),	open (4)		hit (1)	open (1)				
2;01	hit (4), open (1), play (11), hold (7), thunder strike (1)			hit (1)		open (1)			
2;02	open (1), launch (1),	break (1), call (1)	type (1)	hit (1), spank (1)			open (1), play (1)		
2;03	hit (6), play (8), launch (4)		open (1)			fight (1), call (2), open (2)	play (1)		
2;04		fight (1), hit (1)			hit (2)		open (3)		
2;05	hit (5), shoot (1)				open (3)	open (1)		hit (1)	
2;06	hit (3), shoot (1), launch (2), mark (1)	open (15)		hit (2)	fight (1), open (1)	hit (2)	open (1)	hit (2)	
2;07	hit (4), fight (2)			hit (1)	open (4)	open (2)	hit (1)	hit (3)	
2;08		call (4), open (2), spank (1)		hit (2)				open (1), spank (1), hit (2), doze (1)	open (3), hit (6), fight (1)
2;09	hit (1)					turn on (1)			
2;10	hit (31), open (22),				hit (1)	hit (1), fight (2)			call (1), hit (2), hold (1)
3;0				hit (7), open (1)		hit (2), play (4), break (1), open (1), sneeze (1)		hit (1)	shoot (1)
3;01	open (13)						fight (2)		
3;02	knot (2)			break (2), hit (2)			open (1), play (1)	call (1), hit (1)	
3;03				open (4), hit (1)		sweep (1)	hit (3)	open (5)	shoot (3)
3;04	hit (1), open (3), play (2), shoot (4)			open (1)		play (4), hit (4)	sweep (1), play (6)		
3;09									hit (2), open (1)
3;10									open (3)

To summarize, the earliest production of 打 *dǎ* emerges around age 1;06. A small set of different senses of 打 *dǎ* are produced by the children between 1;06 and 3;10, centered around the core meaning of concrete physical action with hand contact. Metaphorical senses are overall very infrequent in terms of both types and tokens. Multiple senses emerge and are often used simultaneously with the senses produced earlier. The small set of senses of 打 *dǎ* typically involves specific contexts that are frequent in a child's daily interactions and bodily experience (e.g., calling, playing games, fighting, sneezing), which usually occur in specific syntactic and semantic contexts that are inherent to the meaning of the specific senses (e.g., an open event typically involves an animate agent and an inanimate patient).

3.3. Comparision between Tong and His Input

To explore the role of input on the semantic development, we compared the production of 打 *dǎ* in the speech of the child Tong and his input. The Tong corpus is selected because of the higher tokens of 打 *dǎ* in both the child and the adults' speech due to a relatively longer period of data collection (cf. Table 2). As shown in Figure 4, Tong is similar to his caregivers in the overall distributions of the semantic representations of 打 *dǎ*. He used 打 *dǎ* as meaning physical action with hand contact (Type 1) most frequently and 95% of his 打 *dǎ* belong to this semantic category, suggesting that the core meaning of hand action is a prototypical usage. In addition to the dominant Type 1, Tong's caregivers show a wider range of semantic categories, including metaphorical uses of traceable hand action (Type 4, 10%) and hand action with instrument (Type 2, 7%). Physical action with unspecified physical contact (Type 3, 2.8%) and metaphorical uses with untraceable hand action (Type 5, 2%) also show minimal uses.

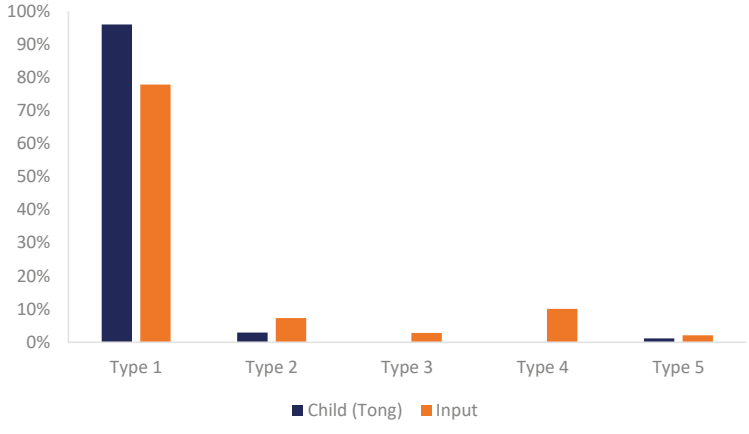


Figure 4. Proportions of the semantic representations of 打 *dǎ* in Tong and his input. Type 1: Physical action focusing on hand contact; Type 2: hand action (mostly with instrument); Type 3: physical action with physical contact unspecified; Type 4: metaphorical uses (hand action traceable); Type 5: metaphorical uses (hand action untraceable).

We further compared the distributions of different senses of 打 *dǎ*. As shown in Figure 5, both Tong and his caregivers produced a variety of different senses. The most frequent sense in Tong's speech is 'hit' (37%), followed by 'open' (30%), 'play (games)' (14%), and 'shoot' (6%), whereas in the caregivers' speech, those senses are also frequent but more evenly distributed, 'open' (17%), 'hit' (17%), 'shoot' (15%), and 'play (games)' (14%). The senses 'launch' and 'hold/possess' show similar proportions of usage in both Tong and his caregivers' speech (4–5%). Overall Tong's use of the different senses of 打 *dǎ* reflects the distributional pattern in his caregivers' speech.

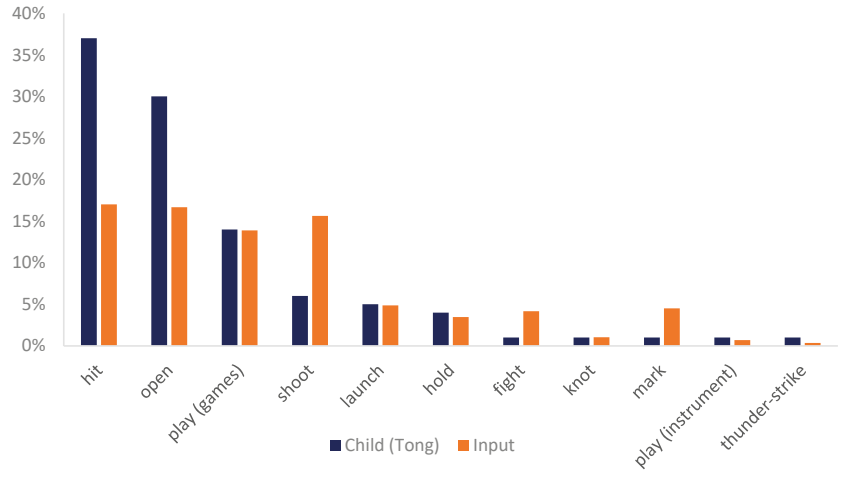


Figure 5. Proportions of different senses of *打 dǎ* in Tong and his input.

To investigate if the development of the multiple senses of *打 dǎ* is closely tied to certain syntactic contexts, we analyzed the syntactic frames or constructions that *打 dǎ* occurs in the speech of Tong and his caregivers. Figure 6 presents the proportions of different syntactic frames. A total of 30 different syntactic frames were identified in Tong's use of *打 dǎ*. As shown in Figure 6, the most frequent type of syntactic frame is *打 dǎ* used with another verb in the form of a verb compound as the predicate of a sentence (50%), followed by transitive frames with an overt object NP (23.84%) and with overt subject and object NPs (9.88%), and locative construction (VPP, 5.23%). Within the sentences containing a verb compound with *打 dǎ*, a variety of syntactic constructions were produced. The most frequent type is the basic transitive frame (38%), followed by negative constructions (30%), modal verb constructions (12%), *ba* constructions (7%), and bare verb compound (7%).[4] Despite the varied syntactic frames, the verb compounds with *打 dǎ* show very limited types—only four different verb compounds were produced, among which *打开 dǎ-kāi* 'hand.action-open' (open) and *打死 dǎ-sǐ* 'hit-be.dead' (kill) account for 93% of the compound predicates. The adults used a wider range of syntactic frames with *打 dǎ* than the child, i.e., a total of 50 different syntactic frames. As shown in Figure 6, the most frequent syntactic frame is *打 dǎ* in the form of a verb compound as the predicate of a sentence (28%), followed by transitive frames with an overt object NP (25.17%) and with overt subject and object NPs (15.52%), and negative construction (13.1%). The adults and the child Tong are thus similar in the overall distribution of the syntactic frames, and verb compounds with *打 dǎ* and the transitive frames account for the majority of the sentence forms.

[4] Examples of the syntactic constructions are shown below. Basic transitive frame: e.g., *打死了七个蚊子 dǎsǐ le qi ge wénzi* 'kill LE seven mosquitos' (I killed seven mosquitos); negative construction: e.g., *没打开 méi dǎ kāi 'not hand.action-open'* (it did not open); modal verb constructions: e.g., *你需要打开 nǐ xūyào dǎkāi 'you need open'* (you need to open this); ba constructions: *把那打开 bǎ nà dǎ-kāi 'ba that hand.action-open'* (open that); bare verb compound: *打死 dǎ-sǐ 'hit-be.dead'* (kill).

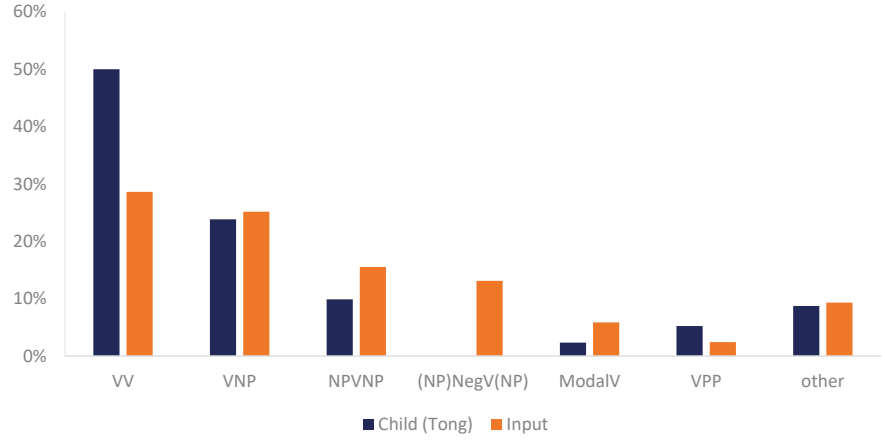


Figure 6. Proportions of syntactic frames of *打 dǎ* in Tong's and his caregiver's speech. (VV = verb compound; VNP = verb followed by an object noun phrase; NPVNP = subject noun phrase followed by the verb and an object noun phrase; VPP = verb followed by a prepositional phrase)[5].

4. Discussion

This study investigates a fundamental issue in child language acquisition, i.e., how do children figure out the mapping between the forms and the meanings in the ambient language, in the case of polysemy. When multiple meanings are available for the same form, how is the form-meaning mapping correctly established? We focus on the semantic development of the polysemous verb *打 dǎ* based on the analyses of longitudinal corpus data from nine Mandarin-learning children from the age of 1;05 to 3;10. The verb *打 dǎ* is highly polysemous in Mandarin and it is also among the first verbs used by all nine young Mandarin children.

We examined the developmental trajectory (i.e., emergent order) of different senses of *打 dǎ*. The results show that young Mandarin-learning children produce a small set of different senses of *打 dǎ* between 1;05 and 3;10, centered around the core meaning of concrete physical action with hand contact (e.g., 'hit', 'open', and 'play games'), and metaphorical senses are overall very infrequent in terms of both types and tokens. Physical actions through an instrument or with unspecified physical contact or metaphorical uses with untraceable hand action are infrequent. This emergent order of concrete before abstract or metaphorical meanings is congruent with findings in prior research that the most prototypical meaning tends to be acquired before the more metaphorically and metonymically motivated meanings (e.g., Nerlich et al. 2003). Booth and Hall (1995) investigated children's (3-, 6-, 9-, and 12-year-olds) understanding of the polysemous cognitive verb *know* and found that children's development of the verb meanings showed an effect of the abstractness and conceptual difficulty hierarchy—low levels of meaning (e.g., perception, recognition, recall, understanding) developed earlier and faster than high levels of meaning (e.g., metacognition, evaluation). Concrete and conceptually less difficulty senses therefore emerge earlier than more abstract and conceptually different senses. In the case of the acquisition of the Mandarin verb *打 dǎ*, we could also see the effect of the general abstractness and conceptual difficulty hierarchy.

[5] Examples of the syntactic frames are shown below. VV: *打死 dǎ-sǐ 'hit-be.dead'* (kill); VNP: *打怪物 dǎ guàiwù '(I) shot the monster'*; NPVNP: *我打妈妈 wǒ dǎ māma 'I hit mom'*; VPP: *打到上面了 'launch to the top'*.

Furthermore, within the concrete physical action with hand contact, three different senses, 'hit', 'open', and 'play (games)', are the most frequent, suggesting the influence of typical hand actions such as hitting, opening, or playing games in young children's daily events. This pattern corroborates prior findings that giving physical punishment emerges as one of the earliest senses of *打 dǎ* in monolingual children (1;9–2;3) (Gao 2015) and bilingual Mandarin children (Sak and Gao 2016). The emergence of the concrete core senses of *打 dǎ* is thus in line with the cognitive linguistics approaches that regard meaning as perceptually grounded—the primary experiential scene in which *打 dǎ* is embedded in events where somebody uses their hand to act on something, and from this experience derives many senses originated from hand actions such as playing games, defeating someone, opening, fastening, construction, covering, possession, marking, launching, insertion, collection, removal, and working.

How do children acquire the different senses of *打 dǎ*? The developmental trajectory supports the continuous derivational approach (Zhang et al. 2010) and the restricted monosemy approach (Kidd and Cameron-Faulkner 2008)—children initially extract a core feature of *打 dǎ*, i.e., volitional physical action focusing on hand contact, but only use it in a restricted way. The knowledge of the core feature of *打 dǎ* is revealed in children's simultaneous production of different senses (e.g., 'open', 'hit', 'spank', 'fight') in the first uses. The limited productivity is reflected in the small number of related concrete senses of *打 dǎ* and the slow increase in the sense types across the sampled ages, 1;06 to 3;10. The dominant senses tend to center around events involving concrete hand motion ('hit' and 'open'). New senses do not emerge in large numbers and are closely tied to the immediate contexts of the children's social interactions. The limited productivity is further seen in the limited set of syntactic frames that *打 dǎ* occurs in, including verb compounds and transitive sentences.

How do children go beyond the limited productivity? Children are likely to be able to recognize the multiple senses (mostly Type 1 senses) as being related in a network model consciously or subconsciously due to the concrete nature of the hand action feature in these related senses. The more abstract metaphorically derived senses (e.g., Type 5 senses such as *打算 dǎsuàn 'plan'*), on the other hand, may be acquired in an independent approach (cf. Zhang et al. 2010) due to the less visible derivational relations to the concrete core sense, and may be treated as unrelated to the core meaning at the beginning. Even in adult language, some meanings may be so different from the core that they may be stored and learned separately (cf. Theakston et al. 2002). Theakston et al. (2002) analyzed the acquisition of the highly versatile English verbs "go", and found that children acquire different constructions in different contexts without evidence that these uses are initially related. It is possible to explain these results in the sense that the various usage patterns of these multifunctional verbs are not linked to one another, but initially represent different syntactic frames. Our current data only contain early speech of Mandarin-learning children (1;05–3;10) and show limited uses of abstract metaphorical senses. Future research should explore whether abstract metaphorical senses are acquired in an independent approach and how they are integrated with the basic core senses to build a complete semantic network with data from older children.

How does input, including syntactic, semantic, and contextual cues, contribute to the acquisition of senses of *打 dǎ*? Our results from the comparison of the child Tong and his input show that Tong's uses of *打 dǎ* reflect the semantic and syntactic distributional patterns in the input, as both Tong and his caregivers use multiple senses of *打 dǎ* from the beginning and both use the physical action involving hand contact most frequently. The most frequently produced senses are also used similarly, even though Tong tends to use the prototypical senses dominantly and the adults' usage is more evenly distributed across the frequent senses. The result provides further support for the usage-based learning that the most frequent senses tend to be the earliest senses in child speech and children's knowledge of the multiple meanings correlates with parental uses (Adricula and Pielke 2019; Booth et al. 1997; Theakston et al. 2002). Computational modeling, utilizing actual child-directed data, show that distributional models are sufficient to reasonably distinguish verb senses, but further information is needed to better predict children's learning (Parisien and Stevenson 2009).

5. Conclusions

To conclude, we could summarize the acquisition of the polysemy of 打 *dǎ* as the following. Children start with the concrete prototypical meaning of 打 *dǎ* as a physical action verb that bears the semantic features [+hand, +volition, +contact, +force] and [-instrument] and gradually extend to physical actions that may involve an instrument ([+instrument]), or without hand contact ([–contact]), and finally to metaphorical senses that share the physical action and volition features with the prototypical sense of 打 *dǎ*, but lack the contact or instrument features ([–contact, –instrument]). Traceable hand action senses may be learned before the more abstract senses with untraceable hand action. The conceptual difficulty in terms of levels of abstractness of the meanings plays a role in children's emergent order of different senses of 打 *dǎ*. The development trajectory also supports the restricted monosemy approach in the early acquisition of polysemy and future research needs to examine further how children develop the full range of the diverse meanings of 打 *dǎ* beyond age 3;10. Furthermore, the inherent lexical semantics of 打 *dǎ* is not the only factor that determines the acquisition. The learning process is also modulated by the distributional patterns of 打 *dǎ* in the input, i.e., the semantic and the syntactic contexts of 打 *dǎ*, which suggests that statistical learning may play a role in the acquisition of different senses (e.g., Adricula and Pielke 2019; Kidd 2012). Note that the order and the derivational route of acquisition of different senses of polysemous verbs may vary, subject to usage-based factors such as input distributional patterns and functional saliency and needs (cf. Zhang et al. 2010). Future research should examine other highly frequent polysemous verbs in Mandarin child speech to map out similarities and differences in development and provide a comprehensive account for early semantic development.

Our research is the first longitudinal study of the emergence and development of the polysemous verb 打 *dǎ* in the speech of monolingual Mandarin-learning children and has pedagogical implications for the teaching and learning of Mandarin as a second/foreign language (L2). Zhang et al. (2011) showed that even intermediate and advanced L2 learners of Mandarin had not acquired the wide range of meanings of the verb 打 *dǎ* in a meaning matching (comprehension) task and in their analysis of interlanguage writing (written production) corpus data. They found that the dominant senses that the learners showed better knowledge of centered around a very limited number of senses, 'hit' and 'play (games)' in both the comprehension and production data. Following the development trajectory of child Mandarin learners, L2 learners of Mandarin should be exposed to a small set of senses derived from the concrete core/prototypical senses of 打 *dǎ* before the more abstract senses that are derived metaphorically. Functionally salient and frequent senses should also be introduced before less frequently used senses. Recent studies of L2 acquisition of Mandarin polysemous verbs also revealed that prototypicality (e.g., prototypical senses) significantly predicts the learning outcomes (e.g., Liang 2014). Furthermore, large number of exemplars of different senses should be provided to facilitate the inference of multiple senses of 打 *dǎ* in communicative contexts (similar to what a child learner experiences in naturalistic language learning situations). Explicit instructions on the semantic features of 打 *dǎ* and the derivational relations may also promote the understanding, generalization, and productive uses of the complicated meanings of 打 *dǎ*. Future research should be conducted to explore the appropriate methods and assessments of these pedagogical applications.

Author Contributions: Conceptualization, J.C. and X.W.; methodology, J.C. and X.W.; data analysis, J.C. and X.W.; writing—original draft preparation, J.C.; writing—review and editing, J.C. and X.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsor had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Adricula, Norielle, and Megan Pielke. 2019. A child's acquisition of polysemy: Of, with, and by in child English. In *Proceedings of the 43rd Boston University Conference on Language Development*. Edited by Megan Brown and Brady Dailey. Somerville: Cascadilla Press, pp. 27–41.

Allwood, Jens. 2003. Meaning potentials and context: Some consequences for the analysis of variation in meaning. In *Cognitive Approaches to Lexical Semantics*. Edited by Hubert Cuyckens, René Dirven and John R. Taylor. Berlin: Mouton de Gruyter, pp. 29–66.

Booth, James R., and William S. Hall. 1995. Development of the understanding of the polysemous meanings of the mental-state verb know. *Cognitive Development* 10: 529–49. [CrossRef]

Booth, James R., William S. Hall, Gregory C. Robison, and Su Yeong Kim. 1997. Acquisition of the Mental State Verb Know by 2- to 5-Year-Old Children. *Journal of Psycholinguistic Research* 26: 581–603. [CrossRef] [PubMed]

Clark, Eve V. 1993. *The Lexicon in Acquisition*. Cambridge: Cambridge University Press.

Clark, Eve V. 1996. Early verbs, event-types, and inflections. In *Children's Language*. Edited by C. E. Johnson and J. H. V. Gilbert. Hillsdale: Lawrence Erlbaum, pp. 61–73.

Deng, Xiangjun, and Virginia Yip. 2018. A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics* 46: 69–92.

Evans, Vyvyan. 2005. The meaning of time: polysemy, the lexicon and conceptual structure. *Journal of Linguistics* 41: 33–75. [CrossRef]

Gao, Hong. 2001. The Physical Foundation of the Patterning of Physical Action Verbs. Ph.D. dissertation, Lund University Press, Sweden.

Gao, Hong. 2015. Children's early production of physical action verbs in Chinese. In *The Oxford Handbook of Chinese Linguistics*. Edited by William Shi-Yuan Wang and Chaofen Sun. Oxford: Oxford University Press, pp. 654–65.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language Acquisition* 1: 3–55. [CrossRef]

Goldberg, Adele. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Huang, Chu-Ren, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I.-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing* 24: 14–23.

Kidd, Evan. 2012. Individual differences in syntactic priming in language acquisition. *Applied Psycholinguistics* 33: 393–418. [CrossRef]

Kidd, Evan, and Thea Cameron-Faulkner. 2008. The acquisition of the multiple senses of with. *Linguistics* 46: 33–61. [CrossRef]

Lakoff, George. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: The University of Chicago Press.

Lakoff, George, and Mark John. 1999. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. New York: Basic Books.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford: Stanford University Press.

Lee, Joanne N., and Letitia R. Naigles. 2005. The input to verb learning in Mandarin Chinese: A role for syntactic bootstrapping. *Developmental Psychology* 41: 529–40. [CrossRef]

Liang, Haiyan. 2014. Factors accounting for acquisition of polysemous shàng 'to go up'-phrases in Chinese as a second language (CSL). *Chinese as a Second Language Research* 3: 201–25. [CrossRef]

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum.

McKercher, David A. 2001. The Polysemy of with in First Language Acquisition. Ph.D. dissertation, Stanford University, Stanford, CA, USA.

Nerlich, Brigitte, Zazie Todd, and David D. Clarke. 2003. Emerging patterns and evolving polysemies: The acquisition of get between four and ten years. In *Polysemy: Flexible Patterns of Meaning in Mind and Language*. Edited by Brigitte Nerlich, Zazie Todd, Vimala Herman and David D. Clarke. Berlin: Mouton de Gruyter, pp. 333–57.

Parisien, Christopher, and Suzanne Stevenson. 2009. Modelling the acquisition of verb polysemy in children. In *Proceedings of the CogSci2009 Workshop on Distributional Semantics beyond Concrete Concepts*. Edited by Niels Taatgen and Hedderik van Rijn. Austin: Cognitive Science Society, pp. 17–22.

Sak, Hui Er, and Helena Hong Gao. 2016. The Polysemy of the Chinese Action Verb "Dǎ" and Its Implications in Child Language Acquisition. In *Chinese Lexical Semantics: CLSW 2016*. Edited by Minghui Dong, Jingxia Lin and Xuri Tang. Cham: Springer, pp. 524–33.

Sandra, Dominiek, and Sally Rice. 1995. Network analyses of prepositional meaning: Mirroring whose mind—The linguist's or the language user's? *Cognitive Linguistics* 6: 89–130. [CrossRef]

中国社会科学院词典编辑室 (Chinese Academy of Social Science Institute of Linguistics), ed. 2016. *The Contemporty Dictionary of Chinese*, 7th ed. (现代汉语词典 (第7版)). Beijing (北京): The Commercial Press (商务印书馆).

Theakston, Anna L., Elena V. M. Lieven, Julian M. Pine, and Caroline F. Rowland. 2002. Going, going, gone: The acquisition of the verb 'go'. *Journal of Child Language* 29: 783–811. [CrossRef] [PubMed]

Tomasello, Michael. 1992. *First Verbs: A Case Study of Early Grammatical Development*. Cambridge: Cambridge University Press.

Tuggy, David. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics* 4: 273–90. [CrossRef]

Vicente, Agustín, and Ingrid L. Falkum. 2017. Polysemy. In *Oxford Research Encyclopedia of Linguistics*. Oxford: Oxford University Press.

Zhang, Yunqiu, Jianshe Zhou, and Jing Fu. 2010. The Acquisition Strategies of Polysemous Verbs in Early Mandarin-Learning Child Speech: A Case Study of a Child Learning the Beijing Mandarin (早期汉语儿童多义词的习得策略—一个北京话儿童的个案研究). *Studies of the Chinese Language (*中国语文*)* 334: 34–43.

Zhang, Jiangli, Dehong Meng, and Weihong Liu. 2011. The Depth of Monosyllabic Polyseme Acquisition of Learners Who Speak Chinese as the Second Language: A Case Study of the Verb Dǎ (汉语第二语言学习者单音多义词习得深度研究一动词"'打"为例). *Applied Linguistics (*语言文字应用*)* 2: 112–21.

languages

MDPI

Article

The Acquisition of Mandarin Consonants by English Learners: The Relationship between Perception and Production

Xinchun Wang * and Jidong Chen

Department of Linguistics, College of Arts and Humanities, California State University, Fresno, CA 93740, USA; jchen@csufresno.edu

* Correspondence: xinw@csufresno.edu

Received: 15 February 2020; Accepted: 9 May 2020; Published: 13 May 2020

Abstract: This study investigates native English CFL (Chinese as a Foreign Language) learners' difficulties with Mandarin consonants at the initial stage of learning and explores the relationship between second language (L2) speech perception and production. Twenty-five native English CFL learners read the eight Mandarin consonants (j/tɕ/, q /tɕ^h/, x /ɕ/, zh /tʂ/, ch /tʂ^h/, sh /ʂ/, z /ts/, and c /ts^h/) in sentences and identified the target sounds in a forced-choice identification task. Native Mandarin listeners identified the consonants produced by the learners and rated the quality of each sound they identified along a scale of 1 (poor) to 7 (good). The learners' mean percentage accuracy scores ranged from 29% to 80% for perception and 25% to 88% for production. Moderate correlations between the perception and production scores were found for two of the eight target sounds. The Mandarin retroflex, palatal, and dental fricatives and affricates, though all lack counterparts in English, pose different problems to the English CFL learners. The misperceived retroflex and palatal sounds were substituted with each other in perception but mis-produced palatal sounds were substituted with each other, not with retroflex sounds. The relationship between perception and production of L2 consonants is not straightforward. The findings are discussed in terms of current speech learning models.

Keywords: Mandarin consonants; English CFL learners; L2 perception and production

1. Introduction

1.1. L2 Speech Perception and Perception Models

Adult second language (L2) speakers' problems with the perception and production of non-native speech sounds are closely related to their first language (L1) experience (Flege 1995). Research has shown that infants are language-general perceivers of speech sounds at the phonetic level. This universal perceptual pattern undergoes a profound change due to increased experience with their first language in the later half of the first year in life (Best 1994; Polka and Bohn 1996; Strange 1995; Werker 1994; Werker and Polka 1993). Adult monolinguals are language-specific perceivers of speech sounds. Perceptual studies using synthesized stimuli found that adult speakers identified stop consonants along a VOT (Voice Onset Time) continuum according to their L1 stop inventories (Lisker and Abramson 1964, 1970). Similar studies on L2 vowel perception using a synthesized vowel continuum also indicated that listeners labeled vowel sounds according to their L1 vowel categories (Rochet 1995). To a large extent, the language-specific nature of adult monolinguals' speech perception underlies the difficulties adult learners face in L2 speech learning.

Evidence from cross-linguistic speech perception studies in which listeners map L2 sounds onto their L1 sound system suggest that the phonetic distances between learners' L1 and L2 sound systems

play an important role in the degree of success in L2 speech perception (Flege and Wayland 2019; Guion et al. 2000; Wang and Chen 2019). For example, in a cross-linguistic perceptual study assessing the phonetic distances between Japanese and English consonants, monolingual Japanese speakers identified English consonants using Japanese consonant categories. The subsequent experiment found that phonetic distances between Japanese and English consonants, as established by the cross-linguistic direct mapping experiment, predicted the discrimination patterns of English consonants by Japanese learners of English with different L2 experience (Guion et al. 2000). Wang and Chen (2019) also found that English CFL learners' perception problems with Mandarin consonants were closely related to the L2 to L1 assimilation patterns.

In searching for the nature of such cross linguistic influence in L2 phonetic learning, researchers have come up with different L2 speech perception models. The Perceptual Assimilation Model (PAM) (Best 1994; Best et al. 2001) assumes that several pairwise assimilation types are possible when two non-native phones are mapped onto the L2 sound system. The pair of L2 phones may be assimilated to two different L1 phones, the Two Category (TC) type, or to a single L1 category equally poorly or well, the Single Category type (SC). The two L2 sounds can also be assimilated to a single native category but one can be a better fit than the other, the Category Goodness type (CG). The PAM model also predicts the degree of difficulties in discriminations of L2 sounds from the most to the least: SC > CG >TC (Best et al. 2001).

Flege (1995, 2007) Speech Learning Model (SLM) states that a learner's L1 and L2 sound systems interact and exist in a common phonological space. Learners will establish an L2 sound category if they perceive the phonetic differences between the L2 sound from the nearest L1 sound or the closest L2 sound. In contrast, "equivalence classification" of an L2 sound with the nearest L1 category blocks the formation of a new phonetic category. Flege claims that learners' ability to establish new phonetic categories remains intact throughout their life span and increases with their L2 experience. Perceptual learning will eventually lead to better production, although the alignment between perception and product may be partial only (Flege 1999). Therefore, the SLM is a dynamic model that emphasizes learners' L2 experiences with the target language.

1.2. The Relationship between Perception and Production

As both the PAM and SLM models place more emphasis on the perceptual assimilation or dissimilation of the L2 sound categories to the L1 sounds, the question arises about the relationship between L2 speech perception and production. Previous research on L2 speech perception and production has led to different conclusions. For example, Rochet (1995) found that native Portuguese speakers produced French /y/ as /i/ while native English speakers produced /y/ as /u/, although both English and Portuguese have /i/ and /u/ in their vowel systems. The subsequent perceptual test using synthesized high vowel continuum revealed that Portuguese listeners assimilated /y/ to /i/ while English listeners assimilated /y/ to /u/ (Rochet 1995). Similarly, Mandarin speakers' production problem with French voiced stops was related to their faulty perception of the voiced stops that do not exist in the Mandarin sound system (Rochet 1995). In a study on English front vowels /i ɪ eɪ ɛæ/, Wang (1997) found that Mandarin speakers had problems with both the perception and production of English lax vowels /ɪ ɛæ/, but they performed better in perception than in production on these three vowels. In contrast, they performed better in production than in perception on English /i eɪ/ categories. Such performance discrepancies between the perception and production on the English front vowels suggest that native Mandarin ESL (English as a Second Language) learners may have used different cues or strategies in their perception and production of English vowels. Flege (1999) also reported a series of studies that showed partial alignment between L2 perception and production. In a more recent study on native Arabic speakers' acquisition of British English vowels and consonants, Evans and Alshangiti (2018) found a link between perception and production, as the better perceivers of English vowels were also the better producers.

L2 phonetic training studies have also examined the relationship between perception and production when assessing the effects of training in both modes. In a recent review study applying the meta-analysis method analyzing 30 perception training studies on L2 segments conducted in the past 25 years, Sakai and Moorman (2018) found that perception training only led to small-sized gains in productions of the target sounds. Their subsequent statistical analysis based on 18 out of the 30 studies led to the conclusion that the production gains were larger on obstruents than on sonorants and vowels. Correlation tests suggested there was a small to medium-sized but statistically nonsignificant relationship between gains in perception and production.

1.3. Studies on L2 Mandarin Consonants

While L2 Learners' problems with non-native speech sounds are well documented on consonants (Bradlow et al. 1997; Guion et al. 2000; Munro et al. 2015), and on vowels (Evans and Alshangiti 2018; Munro and Derwing 2008; Wang 1997; Wang and Munro 2004), as well as on lexical tones (Wang 2006, 2008, 2013), parallel studies on perception and production of L2 speech sounds, particularly on CFL learners' difficulties with Mandarin consonants are still very limited. Several studies on the perception or production of Mandarin consonants reported in the past two decades are summarized in the following.

Lai (2009) investigated learners' perception difficulties with the six Mandarin affricates z /ts/, c /ts^h/, zh /tʂ/, ch /tʂ^h/ and j /tɕ/, and q /tɕ^h/ by native Malay and Burmese speakers residing in Taiwan. The learners and a control group of native Taiwan Mandarin speakers took the same/different discrimination test followed immediately by the identification test on the target affricates paired across different place and manner of articulations. Both learner groups were more accurate in identifying unaspirated affricates than the aspirated counterparts. They also had more problems identifying the dental-retroflex z /ts/-zh /tʂ/ and c /ts^h/-ch /tʂ^h/ contrasts than the palatal affricates. Lai (2009) concluded that there was a merge of dental and retroflex affricates and the dentalization of the retroflex sounds was better explained by the Markedness theory than the learners' first language inference. In fact, the native Mandarin control group demonstrated exactly the same perceptual merge pattern as they had the same rate of errors (around 67%) as the two learner groups on their z /ts/-zh /tʂ/ and c /ts^h/-ch /tʂ^h/ identifications. The findings were not surprising as both L2 groups were learning Mandarin in Taiwan and the dental and retroflex fricative/affricate merge is common in many Mandarin dialects spoken in Southern China as well as in Taiwan (Zhu 2012; Chuang et al. 2019).

Hao (2012) investigated how the learners' L2 to L1 sound mapping patterns and the amount of L2 experience affect the perception of Mandarin sounds. Three groups of native English CFL learners with different length of Mandarin learning experience: Ex group (5.6 years), Inex group (1.5 years) and Noex (No experience) took the perceptual tests. Hao (2012) found that phonetic context and L2 Mandarin experience affect the learners' L2 to L1 sound mapping patterns. More experienced learners gave more consistent responses in Mandarin to English sound classifications and were less affected by phonetic contexts than less experienced learners. The Noex group assimilated Mandarin /s/ to English /z/ more often than to /s/ while both learner groups identified /s/ as English /s/. All three groups assimilated Mandarin /ʂ/ and /ɕ/ to English /ʃ/ mostly except that the Noex group split the classification of /ɕ/ to /s/ and /ʃ/ equally when /ɕ/ was followed by an unrounded vowel /i/. While both Mandarin /ʂ/ and /ɕ/ were assimilated to the English /ʃ/, /ʂ/ was a better fit than /ɕ/ as indicated by both the identification accuracy rate and the higher goodness rating score, a Category Goodness type of assimilation according to the PAM model. In the identification test, the Mandarin /ʂ-ɕ/ contrast was found difficult for the learners and more so for the Inex group than the Ex group. All three groups performed equally well in discriminating the /ʂ u-su/ and /ʂi-si/ contrasts in the discrimination test. The author concluded that the L2 to L1 assimilation patterns failed to predict discrimination accuracy of Mandarin contrasts in most cases.

In a more recent cross-linguistic perception study on Mandarin consonants (Wang and Chen 2019), native English listeners with no Mandarin learning experience identified 10 Mandarin consonants

in syllables (z /tsa/, c /ts^{h}a/, s /sa/, j /tɕ^{j}a/, q /tɕ^{hj}a/, x /ɕ^{j}a/, zh /tʂ a/, ch /tʂ^{h}a/, sh /ʂa/, and r /ʐa/) using the closest English sounds in a ten-way forced choice task followed by a goodness rating task along a scale of 1 (poor) to 7 (good). L2 to L1 sound mapping fitting indexes (identification score x rating score) were calculated to assess the phonetic distances between Mandarin and English consonants. Wang and Chen (2019) found there was a range of phonetic distances between the L1 and L2 sounds based on the fit indexes (range from 1.0 to 6.3 out of 7). The "poor" matching categories were x /ɕ/, c /ts^h/, q /tɕ^h/, zh /tʂ/, and j /tɕ/ whose fit indexes were below the mean (3.7, s.d.=1.7). The "fair" fitting categories were ch /tʂ^h/, s /s/, and z /ts/ whose fit indexes were at the mean. The "good" matching sounds were r /ʐ/, and sh /ʂ/ whose fit indexes were 1s.d. above the mean (Wang and Chen 2019). In a subsequent study on the identification of Mandarin consonants by English CFL learners at two different proficiency levels, the learners' perception scores of Mandarin consonants were found to be closely related to the L2 to L1 assimilation patterns. Results showed that zh /tʂ/, q /tɕ^h/, c /ts^h/, and x /ɕ/ (the poor fitting sounds) received the lowest % identification scores among the 10 sounds by the beginning level learners. The intermediate group outperformed the beginning group on zh /tʂ/, q /tɕ^h/, and c /ts^h/. These findings suggest that the perceived phonetic distances between L1 and L2 consonants predicted the English CFL learners' L2 Mandarin consonant identification problems and increased L2 experience improved perceptual learning. No production data were reported in this study.

In a production study, Liu and Jongman (2012) investigated both the temporal and spectral features of Mandarin dental affricates z /ts/, and c /ts^h/ produced by native English CFL learners with different proficiency levels. The authors found that both the novice and more experienced learner groups acquired the durational differences for the /ts/, and /ts^h/ contrast but only the more advanced learners acquired the spectral (center of gravity) contrast between the target sound pair. It was not clear what weight the temporal and spectral cue each carries to the perceptual accuracy of the target contrast as no perception test was conducted to measure the accuracy of the learners' productions. This study dealt with only one pair of Mandarin affricate contrast at dental place of articulation.

In a similar study involving more Mandarin affricate contrasts, Yang and Yu (2019) investigated the perception and production of six Mandarin affricates z /ts/, c /ts^h/, zh /tʂ/, ch /tʂ^h/, j /tɕ/, and q /tɕ^h/ by native English CFL learners at beginning and intermediate levels. Both learner groups matched the native Mandarin group in perception accuracy scores in discriminating but not in identifying the target sounds. The effect of place of articulation and aspiration were significant but not uniform across the board. For example, the unaspirated palatal j /tɕ/ was significantly better identified than the aspirated palatal counterpart q /tɕ^h/ but the aspirated retroflex ch /tʂ^h/ was better identified than the unaspirated counterpart zh /tʂ/. The authors concluded that different affricates pose different learning difficulties for English CFL learners. In the production test, the intermediate group outperformed the beginning group in approximating the native speakers in the production of some but not all the acoustical features under investigation, indicating the learners did not acquire the affricates completely. Their data suggest that the distinction between palatal and retroflex affricates is more difficult for learners due to the assimilation of both classes to the same English post-alveolar affricates, the two to one type of (SC) of assimilation, according to the PAM model.

To summarize the findings of the above studies, the difficulties with the perception accuracy of Mandarin consonants by native English CFL learners are related to their L2 to L1 perceptual assimilation patterns (Hao 2012; Wang and Chen 2019; Yang and Yu 2019). In general, Mandarin retroflex and palatal contrasts pose more difficulties to the English CFL learners than other place contrasts (Hao 2012; Yang and Yu 2019), while dental and retroflex contrasts were more difficult for Malay and Burmese learners (Lai 2009). The effect of L2 experience did not appear to affect the learners' discrimination accuracy but did influence their identification accuracy of the Mandarin consonants (Hao 2012; Lai 2009; Wang and Chen 2019; Yang and Yu 2019). These results confirmed earlier findings of the advantage of identification over discrimination task in L2 phonetic test and training because the former help the learners focus more on the key phonetic/acoustic features that distinguish the target sound contrasts (Wang and Munro 2004). In production, the effect of L2 experience was more evident

as the more experienced learners outperformed less experienced learners in approximating the native speakers in the production of some but not all the acoustical features of the target sound contrasts (Liu and Jongman 2012; Yang and Yu 2019).

1.4. The Current Study

Several studies summarized in Section 1.3 (Hao 2012; Lai 2009; Wang and Chen 2019) investigated CFL learners' perception problems with Mandarin consonants but did not examine their production problems. Two production studies on Mandarin consonants (Liu and Jongman 2012; Yang and Yu 2019) compared the acoustic properties of the learners' productions with those of the native speakers but did not include the direct assessment of the intelligibility of the L2 speech. Parallel studies that compare the CFL learners' perception and production performance with Mandarin consonants are extremely rare. This study aims to fill this gap by investigating native English CFL learners' difficulties with the Mandarin consonants in both perception and production at initial stage of learning. An additional goal is to examine the relationship between L2 speech perception and production. The research questions are:

1. Which Mandarin consonants are difficult to identify and produce for native English CFL learners at early stage of learning?
2. How do the phonetic differences and distances between Mandarin and English consonants, as perceived by English listeners in an earlier study (Wang and Chen 2019), affect the perception and production of Mandarin consonants?
3. What are the learners' performance differences between their perception and production of Mandarin consonants?

1.5. Mandarin Consonants

Table 1 presents the 22 Mandarin consonants in IPA. The sounds in bold are the eight target Mandarin consonants under investigation in the current study: z /ts/, c /ts^h/, j /tɕ/, q /tɕ^h/, x /ɕ/, zh /tʂ/, ch /tʂ^h/, sh /ʂ/. They form the fricative/affricate groups at dental, retroflex, and alveolo-palatal (also commonly referred to as palatal) places reported to be difficult for English CFL learners to acquire as these sounds do not have corresponding counterparts in English (Lin 2005; Wang and Chen 2019).

Table 1. Mandarin Consonants.

	Labial	**Dental**	**Retroflex**	**Palatal**	**Velar**
Stop	p p^h	t t^h			k k^h
Affricate		**ts ts^h**	**tʂ tʂ^h**	**tɕtɕ^h**	
Fricative	f	s	ʂ ʐ	ɕ	x
Nasal	m	n			ŋ
Liquid		l			

2. Experiment 1: Perception of Mandarin Consonants

2.1. Participants

The participants were 25 native English speaking (15 male, 10 female, mean age = 19.6) beginning level CFL learners enrolled in a first semester Chinese course in a public university in the U.S. All participants reported speaking English as their native language. Twenty of them were born and raised in the United States and five were born in foreign countries but moved to the U.S. between the ages of 2 and 5. Some participants reported speaking another language along with English as their first languages. They were four English/Spanish, four English/Hmong, two English/Tagalog, and one English/Vietnamese early bilinguals. At the point of data collection, the participants were about three months into the 16-week semester and all had learned and practiced Chinese consonants by then.

2.2. Material

The perceptual identification test initially included 10 Mandarin consonants in syllables (z /tsa/, c /ts^{h}a/, s /sa/, j /tɕ^{j}a/, q /tɕ^{hj}a/, x /ɕ^{j}a/, zh /tʂa/, ch /tʂ^{h}a/, sh /ʂa/, and r /ʐa/) that were produced by two native Mandarin Speakers, one male and one female, through a reading task. The target words were produced in a carrier sentence *wo shuo* ___ *zi* (我说 —字). "I say — word". The recordings were made on a MacBook Pro computer using the Praat software. The target syllables were separated from the sentences using waveform editing, normalized for peak volume, and saved as wave form for presentations. Eight of the 10 sounds (z /ts/, c /ts^h/, j /tɕ/, q /tɕ^h/, x /ɕ/, zh /tʂ/, ch /tʂ^h/, sh /ʂ/) were analyzed for this perception experiment to pair exactly with the eight target consonants in the production test.

2.3. Procedure

All participants gave their informed consent for inclusion before they participated in this study which was approved by the ethics committee of the researchers' university. Individual perception identification tasks were carried out in a sound booth on a MacBook Pro computer using Praat ExperimentMFC identification test design. A total of 60 stimuli (10 sounds 2 speakers 3 repetitions) were randomized and presented in a 10-way forced choice task. The labels for choices were the 10 consonants in pinyin (the official Romanized transcription of Mandarin Chinese) displayed on the computer screen during the test. The listeners were instructed to listen carefully for the initial consonant in each stimulus and identify the sound they heard by clicking on the corresponding consonant on the screen. During the test, they could choose to replay each stimulus twice in the case of uncertainty. To familiarize the learners with the task, before the real test began, each participant had a trial session using the stimuli not included for analyses. The software automatically recorded the test data to be exported for analysis.

2.4. Results

Individual participants' correct identifications of each target consonant were converted to percentage accuracy scores. Their misidentified target sounds were also converted to percentage error rate and were tallied for substitution patterns. The group mean percentage correct identification scores of the eight consonants ranged from 29% zh /tʂ/ to 80% ch /tʂ^h/. To investigate the learners' perceptual substitution patterns of the misidentified consonants, a confusion matrix was created and is presented along with the percentage correct identification scores in Table 2.

Table 2. Mean percentage correct identification scores (in bold) and confusion matrix of Mandarin consonants by native English CFL (Chinese as a Foreign Language) learners (N = 25).

Target	**Identified**									
	zh /tʂ/	**ch /tʂ^h/**	**sh /ʂ/**	**j /tɕ/**	**q /tɕ^h/**	**x /ɕ/**	**z /ts/**	**c /ts^h/**	**s /s/**	**r / ʐ/**
zh /tʂ/	**29**	22	5	32	4	1	3	2	1	1
ch /tʂ^h/	4	**80**	4		5		1	3	1	1
sh /ʂ/	1	5	**70**		1	15		3	5	
j /tɕ/	8	11		**61**	11	3	4	1		
q /tɕ^h/	5	43	3	3	**31**	4	3	4	3	1
x /ɕ/	2	3	31	3	5	**46**		3	5	1
z /ts/	7	1		1		3	**56**	11	20	
c/ts^h/	5	11	2		7	1	9	**53**	11	

A One-Way repeated measures ANOVA was conducted to compare the differences between the perception scores on the consonants (8 levels). There was a significant effect of consonant (Wilk's Lambda = 0.102, F (7, 25) = 22.526, p = 0.000). The subsequent post hoc Bonferroni tests adjusted for multiple comparisons revealed that a series of pairwise comparisons were significant. The results are presented in Table 3.

Table 3. Pairwise comparisons of differences between the consonants in perception (** $p < 0.01$, * $p < 0.05$).

	c /ts^h/	**ch /ʈʂ^h/**	**j /tɕ/**	**q /tɕ^h/**	**sh /ʂ/**	**x/ɕ/**	**z /ts/**	**zh /ʈʂ/**
c /ts^h/	-	*						*
ch /ʈʂ^h/	*	-		**		**		**
j /tɕ/			-	**				**
q /tɕ^h/		**	**	-	**		**	
sh /ʂ/				**	-			**
x /ɕ/		**				-		
z /ts/				**			-	**
zh /ʈʂ/	*	**	**		**		**	-

2.5. Discussion

The perception test results showed that Mandarin zh /ʈʂ/ (29%), q /tɕ^h/ (31%), and x /ɕ/ (46), were the worst identified sounds by native English CFL learners. The results of the pairwise comparisons confirmed that the perception scores of zh /ʈʂ/ (29%), q /tɕ^h/ (31%), and x /ɕ/ (46) were significantly different from all the other five sounds but not different from each other (See Table 3). The findings were similar to the Wang and Chen (2019) study in which the same three sounds were also among the four most difficult consonants for the beginning level learners. These three sounds were also among the poorest fitting categories to the learners' L1 English sounds as established by the native English listeners in a cross-linguistic identification test found in the Wang and Chen (2019) study. The current findings support the Wang and Chen (2019) findings that phonetic distances and differences between L1 and L2 consonants predicted native English CFL learner's perception problems with Mandarin consonants at initial stage of learning.

An inspection of the confusion matrix of the misidentified sounds led to the observation that misidentified retroflex and palatal sounds are mostly confused with each other. For example, the retroflex zh /ʈʂ/ was heard as palatal j /tɕ/ 32% of times. The confusion score exceeded the correct % identifications of zh /ʈʂ/ of only 29%. Similarly but to a less degree, the highest % of misidentified sh /ʂ/ was heard as x /ɕ/ 15%. Misidentified palatal sounds were heard mostly as retroflex sounds as well. The palatal sounds q /tɕ^h/ and x /ɕ/ were misidentified as the retroflex sounds ch /ʈʂ^h/ and sh /ʂ/ 43% and 31% of times, respectively. Therefore, the retroflex and palatal fricatives and affricates are both difficult for English CFL learners to identify.

The dental affricates z /ts/ and c/ts^h/ were also poorly perceived by the learners. The most misidentified unaspirated dental z /ts/ sound was heard as /s/ 20% of times. The aspirated dental affricate c/ts^h/ sound was misidentified as ch /ʈʂ^h/ and /s/ 11% each.

3. Experiment 2: Production of Mandarin Consonants

3.1. Participants

The participants were the same 25 beginning level CFL learners who took the perception test in Experiment 1. They provided the production data through a reading task that took place immediately before the perception tests.

3.2. Material and Procedure

The reading list consisted of 20 target words in pinyin embedded in a carrier sentence *wo shuo* ___ *zi* (我说 —字). "I say — word". Each of the 20 target words was repeated once, yielding two versions of the same target sounds. The participants were given a few minutes to prepare for the reading task. Any questions about the pronunciation of any sounds were answered during the preparation time. The participants were told to read the list at normal speed. The recordings were then made on a MacBook Pro computer using the Praat software. The words containing the eight target consonants

(z/tsa/, c /ts^{h}a/, j /tɕ^{j}a/, q /tɕ^{hj}a/, x /ɕ^{j}a/, zh /tʂa/, ch /tʂ^{h}a/, and sh /ʂa/) were separated from the sentences using waveform editing, normalized for peak volume, and saved as wave form for presentations.

3.3. Assessment

Three phonetically trained native Mandarin speakers, all have taught Mandarin Chinese courses in North America, assessed the participants' productions in an eight-way forced choice identification task followed by the goodness rating task along a scale of 1 (poor) to 7 (good). The participants' productions of the eight target sounds were blocked by groups of five speakers, yielding 80 tokens in each session (8 words 5 speakers 2 repetitions). The eight-way forced choice task and the subsequent rating task were created using Praat ExperimentMFC identification test design. Individual identification tasks were carried out on a MacBook Pro computer in a quiet room. The native Mandarin speakers listened to each Mandarin stimulus and identified the initial consonant by clicking on the corresponding label in pinyin on the computer screen. Immediately after the identification of each sound, the listeners rated the fitness of the sound they identified by choosing a number along the scale of 1 (poor) to 7 (good). In the cases of uncertainty, the listener could replay the stimulus up to three times before the choice was made. In addition to the eight target sounds in pinyin, /s/, /t/ and /k/ sounds were also included in the labels for identifications. Based on a screening test by the first author, these three sounds provided additional options for the listeners to choose for the mis-produced target sounds. The listeners all had a trial session to learn the test procedure before the real judgement test began. They each then completed 6 sessions with mandatory breaks in between sessions. The data of one session were excluded from analysis as those five participants were not native English speakers.

3.4. Results

To assess interrater variability, a reliability test was carried out and a high degree of agreement was found among the three raters. The average measures Intraclass Correlation was 0.821 with a 95% confidence interval from 0.788 to 0.849 (F (399,798) = 5.605, $p < 0.001$). Therefore, the mean group production score for each consonant was calculated by taking the average of the three listeners' identification scores. To further explore the production substitution patterns of each mis-produced consonant, a confusion matrix was created. Table 4 summarizes the mean percentage correct production scores (in bold) and the confusion matrix of the mis-produced eight target consonants. The mean goodness rating scores of each target consonant were also calculated and presented (in italic) by the mean correct identification scores in Table 4.

Table 4. Mean percentage correct production scores (in bold), mean rating scores (in italic), and confusion matrix of mis-produced Mandarin consonants by native English CFL learners (N = 25).

Target	**Identified**											
	zh /tʂ/	**ch /tʂ^h/**	**sh /ʂ/**	**j /tɕ/**	**q /tɕ^h/**	**x /ɕ/**	**z /ts/**	**c/ts^h/**	**/s/**	**/t/**	**/k/**
zh /tʂ/	**25** (*4.8*)	1	6	37	6	1	19	2	1		
ch /tʂ^h/	5	**47** (*5*)	1	2	35	1	1	7		2	
sh /ʂ/	4	8	**52** (*5*)	2	7	22	1	2	1		
j /tɕ/	13	1		**68** (*5.6*)	3	2	12			1	
q /tɕ^h/	3	9	1	25	**43** (*4.7*)	7	2	5			4
x /ɕ/	1		7	11	5	**64** (*5.5*)	4	2	6	1	
z /ts/	3			2			**88** (*5.1*)		6		
c /ts^h/	1	4	1	5	1		27	**25** (*4.1*)	8	7	23

Overall, the percentage correct production scores of the eight target sounds, as identified by the three native listeners, ranged from 25% (zh /tʂ/ and c/ts^h/) to 88% (z /ts/). A One-Way ANOVA on the percentage identification scores revealed a significant effect of consonant (Wilk's Lambda = 0.288, F (7, 143) = 50.486, $p = 0.000$). The subsequent post hoc Bonferroni tests adjusted for multiple comparisons revealed that a series of pairwise comparisons were significant. The results are presented in Table 5.

Table 5. Pairwise comparisons of differences between the consonants in production (** $p < 0.01$, * $p < 0.05$).

	c /ts^h/	**ch /ʈʂ^h/**	**j /tɕ/**	**q /tɕ^h/**	**sh /ʂ/**	**x/ɕ/**	**z /ts/**	**zh /ʈʂ/**
c /ts^h/	-	**	**	*	**	**	**	
ch /ʈʂ^h/	**	-	*				**	**
j /tɕ/	**	*	-	**			*	**
q /tɕ^h/	*		**	-		**	**	*
sh /ʂ/	**				-		**	**
x /ɕ/	**			**		-	**	**
z /ts/	**	**	**	**	**	**	-	**
zh /ʈʂ/		**	**	*	**	**	**	-

In addition to the percentage correct identification scores, the assessment of the learners' production performance also included the rating scores (along a scale of 1 to 7) for each correctly identified consonant. The mean rating scores of each correctly identified sound ranged from 4.1 to 5.5 out of 7. The goodness rating task provided the listeners with a choice among a range of "fitness" of the learner's production to the native norm of the target sound, even if the intended sound was correctly identified. Therefore, taking into consideration of the rating scores, the "adjusted" production score for each consonant was calculated by multiplying the percentage correct identification score by the rating score. The results of the adjusted production scores for the eight target consonants are summarized in Table 6.

Table 6. Mean adjusted production scores (% production x rating score) of the target consonants produced by English CFL learners (N = 25).

Sounds	**Production Score**	**Rating Score**	**Adjusted Score**
z /ts/	88	5.1	4.5
j /tɕ/	68	5.6	3.8
x /ɕ/	64	5.5	3.5
sh /ʂ/	52	5	2.6
ch /ʈʂ^h/	47	5	2.4
q /tɕ^h/	43	4.7	2.0
zh /ʈʂ/	25	4.8	1.2
c /ts^h/	25	4.1	1.0
Mean	52 (21.5)	5 (.5)	2.6 (1.2)

3.5. Discussion

The two most difficult consonants for the learners to produce were c /ts^h/ and zh/ʈʂ/, each with a low percentage production score of 25% only, which was significantly different from all the other sounds under investigation. Mandarin c /ts^h/ and zh/ʈʂ/ also received the lowest adjusted production scores (c /ts^h/ (1.0) and zh/ʈʂ/ (1.2), when the rating scores were taken into consideration. On the other hand, the adjusted production scores for z /ts/ and j /tɕ/ were the highest among the eight sounds. The mean rating score of the eight target consonants was 5.0, with a range from 4.1 to 5.6. The difference between the best and worst rated sound was 1.5. These rating scores suggest that the listeners did not use the full range of the rating scale, especially at the lower end, once a target sound was correctly identified. While the range of the rating scores was relatively small comparing to the widely different percentage correct identification scores, those three poorly produced sounds with the lowest percentage correct identification scores c /ts^h/ (25%), zh/ʈʂ/ (25%), and q/tɕ^h/ (35%) also received the rating scores below 5. All the other five sounds had a rating score of 5 and above.

The substitution patterns in production were similar to those in perception for the retroflex sounds. The mis-produced retroflex sounds zh/ʈʂ/, ch/ʈʂ^h/, and sh/ʂ/ were overwhelmingly heard as the palatal sounds j/tɕ/, (37%), q/tɕ^h/ (35%), and x/ɕ/ (22%) by the native Mandarin listeners. However, the mis-produced palatal sounds q/tɕ^h/ and x/ɕ/ were mostly heard by the native Mandarin listeners

as the unaspirated palatal affricate j /tɕ/. For the aspirated dental affricate c /ts^h/, 23% was heard as the /k/ sound. This unexpected substitution pattern was more likely caused by the pinyin spelling of "ca" being mistaken as the English orthography. Obviously, these speakers have not learned the c /ts^h/ sound, or, at least have not associated the pinyin c with the Mandarin sound /ts^h/.

4. Perception and Production Comparisons

Visual inspection of Figure 1, which compares the mean percentage correct perception and production scores of the eight consonants, shows the patterns of higher scores in perception than in production for retroflex sounds zh/tʂ/, ch/tʂ^h/, and sh/ʂ/ but vice versa for palatal sounds j/tɕ/, q/tɕ^h/, and x/ɕ/. The dental affricates were mixed as c /ts^h/, the aspirated dental affricate received the lowest production score of 25% while the unaspirated counterpart z/ts/ had the highest production score of 88%.

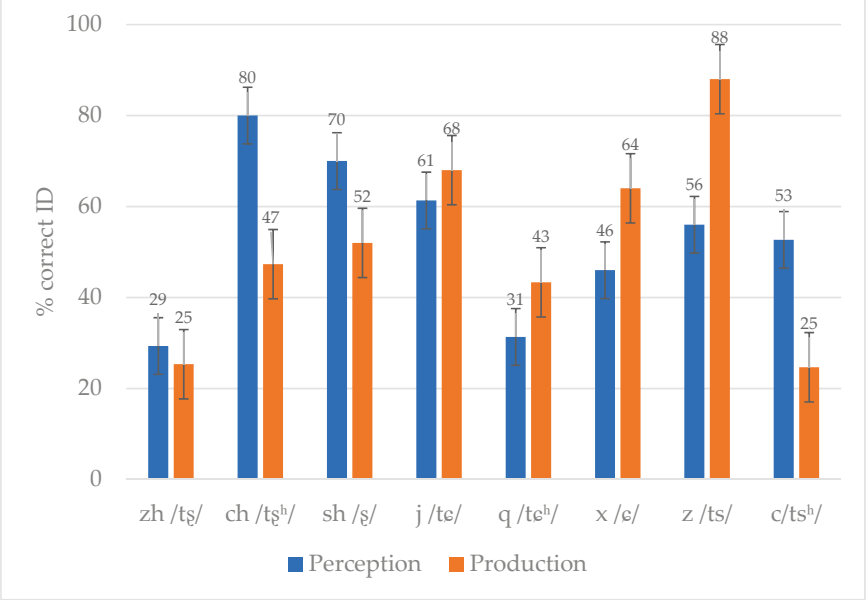


Figure 1. Mean % perception and production scores of Mandarin consonants by English CFL Learners (N = 25).

To investigate the relationship between the perception and production accuracies by the participants, Pearson Coefficients Correlation tests were performed on the percentage correct perception and production scores. The strength of the correlation test results, along with the mean percentage perception and production scores and standard deviations are presented in Table 7.

Results of the Pearson coefficients correlation tests (2-tailed) revealed a moderate size of correlations between the perception and production scores for two of the eight consonants. They were c/ts^h/, r = 0.619, $p < 0.01$, and x/ɕ/, r = 0.508, $p < 0.01$. No significant correlations between the perception and production scores were found for the remaining six consonants. The Pearson's r ranged from (r = 0.028, $p = 0.896$) for q /tɕ^h/ to (r = 0.077, $p = 0.715$) for j /tɕ/.

Table 7. Pearson coefficients correlation tests between the % correct perception and production scores (** $p < 0.01$).

Sounds.	Perception	Production	r =	*p* =
zh /tʂ/	29 (29)	25 (34)	0.061	0.773
ch /tʂ^h/	80 (25)	47 (34)	0.039	0.854
sh /ʂ/	70 (25)	52 (37)	0.030	0.888
j /tɕ/	61 (34)	68 (38)	0.077	0.715
q /tɕ^h/	31 (21)	43 (35)	0.028	0.896
x /ɕ/	46 (29)	64 (37)	0.508 **	0.010
z /ts/	56 (28)	88 (24)	0.162	0.439
c /ts^h/	53 (36)	25 (30)	0.619 **	0.001

5. General Discussion and Conclusions

To answer research question 1 which asked which Mandarin consonants pose difficulties for native English CFL learners, results of Experiment 1 showed that the learners had different degrees of difficulties in identification of the eight Mandarin consonants under investigation. The most difficult sounds were zh /tʂ/ (29%), q /tɕ^h/ (31%), and x /ɕ/ (46%). The findings were consistent with an earlier study by Wang and Chen (2019) in which zh /tʂ/, q /tɕ^h/, and x /ɕ/ were also among the four most difficult categories identified by the low level CFL learners. Similarly, Yang and Yu (2019) also found that among the six Mandarin affricates they investigated, zh /tʂ/ and q /tɕ^h/ were more difficult than their counterparts ch /tʂ^h/ and j /tɕ/ for the native English CFL learners.

The confusion matrix of misidentified sounds showed the English CFL learners substituted the retroflex and palatal sounds with each other mostly and such confusion patterns suggest that the learners have not established separate categories for these sounds. The English CFL learners' problems with the retroflex and palatal sounds, also reported in Hao (2012) and Yang and Yu (2019) studies, are closely related to the phonetic differences between their L1 and L2 sound systems. Wang and Chen (2019) found the native English listeners mapped both Mandarin retroflex affricates zh/tʂ/, ch /tʂ^h/, and the aspirated palatal affricate q/tɕ^h/ onto the English /tʃ/ sound, though the degree of "fitness" was different, as ch /tʂ^h/ was identified as a much better fit to /tʃ/ (4.4) than zh/tʂ/ (3.3), and q/tɕ^h/ (2.3), indicated by their "fit indexes": (% identification x goodness rating score). The three-to-one perceptual assimilation pattern, to a large extent, underlies the native English CFL learners' perception problems with zh /tʂ/, ch /tʂ^h/, and q /tɕ^h/. The better fitting category ch /tʂ^h/ was identified with an accuracy score of 80%, as compared with 31% for q/tɕ^h/ and 29% for zh/tʂ/ in the current study. The findings support the Category Goodness (CG) type of assimilation of the Perceptual Assimilation Model (PAM), which states that two sounds are assimilated to a single native category resulting in a better fit for one than the other (Best et al. 2001). The current findings suggest that the two-to-one CG type of assimilation can be expanded to three-to-one assimilation. More such three-to-one, and two-to-one, as well as one-to-two mappings of Mandarin consonants onto English categories found in the Wang and Chen (2019) study are presented in Figure 2. These cross-linguistic assimilation patterns shed light on the difficulties native English CFL learners demonstrated in their perception and production of the eight target consonants in the current findings. (See the original study for detailed analysis of the assimilation patterns).

L2 to L1 assimilation patterns can also explain the English CFL learners' difficulties with Mandarin palatal sound x/ɕ/. Both x/ɕ/ and sh/ʂ/ were mapped onto English /ʃ/ but sh/ʂ/ was a better fit than x/ɕ/, a CG type of assimilation. Hao (2012) reported the same CG assimilation pattern for x/ɕ/ and sh/ʂ/ and similar learning results on x/ɕ/ in her study. Mandarin x/ɕ/ was difficult for the English learners also because it was assimilated to both English /z/ and /ʃ/, a one-to-two "split" match to the native categories, a "revised" Single Category (SC) type of assimilation (see Figure 2).

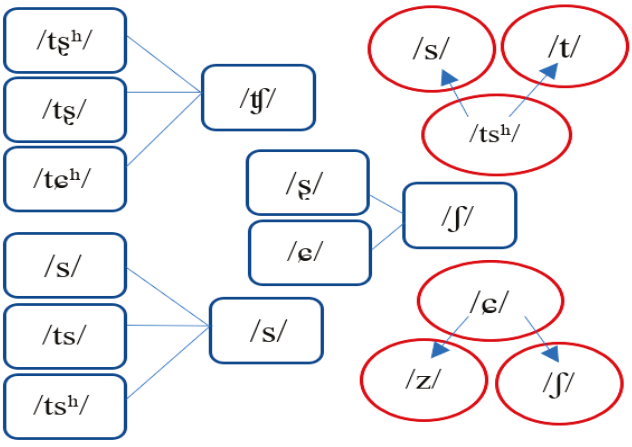


Figure 2. Mandarin to English sound mapping patterns by English listeners; 3 to 1 and 2 to 1 mappings are in blue squares and 1 to 2 mappings are in red circles.

The Mandarin dental affricates z /ts/ and c/ts^h/ were also poorly perceived by the learners in the current study. As seen in Figure 2, the aspirated affricate /ts^h/ was assimilated to both English /s/ and /t/, also a "revised" Single Category (SC) type of assimilation. The unaspirated dental affricate z /ts/, together with s /s/ and c /ts^h/, were mapped onto English /s/, causing difficulties for the poor fitting categories z /ts/ and c /ts^h/. While both z /ts/ and c /ts^h/ exist in the English word finals "reads" and" boots", these novice learners did not appear to have made the associations in identifying the target sounds. One explanation may be that Mandarin z /ts/ and c /ts^h/ are stand-alone phonemes and are more prominent at word initial positions than the morphological word endings of /dz/ and /ts/ in English.

Overall, to answer research question 2, the current data suggest that the perceived phonetic differences and distances between Mandarin and English consonants predicted the learners' perceptual difficulties with the L2 Mandarin consonants. The perception data also support the PAM model.

Flege's Speech Learning Model (SLM) may also provide explanations for the current findings. The learners' phonetic spaces for L1 and L2 consonants need to be reorganized to establish new phonetic categories for the Mandarin retroflex, palatal and dental sounds. For example, learners need to distinguish the differences between c /ts^h/, x /ɕ/, z /ts/, q /tɕ^h/ and others in order to establish these categories. On the other hand, "equivalence classification" of the SLM may be at work for ch /tʂ^h/ to be identified as English /tʃ/. While ch /tʂ^h/ (80%) was the best identified category among the eight target sounds by the learners, its production score (47%) was much lower. Therefore, even if some of the L2 categories seemed to have been established by the majority of the listeners, "equivalence classifications" may have prevented them from forming the native-like perception category.

The results of Experiment 2 showed the percentage correct production scores of the eight target sounds ranged from 25% (zh /tʂ/ and c/ts^h/) to 88% (z /ts/). Pairwise comparisons data shown in Table 4 indicated the native English CFL learners had the most production difficulties with the Mandarin c/ts^h/ and zh /tʂ/, followed by q/tɕ^h/ sounds. The pattern of substitutions in production was similar to that of perception for the retroflex sounds zh /tʂ/, ch /tʂ^h/, and sh /ʂ/, which were substituted with palatals j/tɕ/, q/tɕ^h/, and x/ɕ/. However, the mis-produced palatal sounds q/tɕ^h/, x/ɕ/ were not confused with the retroflex sounds but were mostly heard by the native Mandarin listeners as the unaspirated palatal affricate j /tɕ/. These substitution patterns suggest that the retroflex sounds were more difficult for the English CFL learners to produce.

Comparing the results of the two experiments, the learners had the tendency of better performance on the retroflex sounds zh /tʂ/, ch /tʂ^h/, and sh /ʂ/ in perception than in production but vice versa on

palatal sounds j /tɕ/, q /tɕ^h/, and x /ɕ/. The results on dental sounds z /ts/, and c /ts^h/ were mixed across the two domains. Mandarin retroflex and palatal fricatives and affricates, though both lack counterparts in English, pose different problems to the English CFL learners in perception and production. The results of the correlation tests comparing the perception and production scores showed only two of the eight target consonants, x /ɕ/ and c /ts^h/ were moderately correlated. The lack of correlations in the learners' perception and production scores for the majority of the sounds under investigation suggest the relationship between L2 speech perception and production is not straightforward.

One possible explanation for such misalignment in L2 speech perception and production might be that perception does not always lead production. For example, different mechanisms or strategies may be involved in perception and production of the retroflex sound ch /tʂ^h/ by the beginning level English CFL learners. The learners' better perception of ch /tʂ^h/ (80%) may be explained by their closest match of the L2 Mandarin ch /tʂ^h/ to their L1 English /tʃ/ sound. In perception identification tasks, anything that is close enough to the nearest English /tʃ/ can be labeled as ch /tʂ^h/. However, the same strategy would not work for the production, if the learners have not established native-like retroflex ch /tʂ^h/ category. The key phonetic gestures in producing the correct retroflex affricates in the Mandarin ch /tʂ^h/ sound cannot be effectively replaced by the gestures of English /tʃ/. The intended ch /tʂ^h/ would not be heard as the target ch /tʂ^h/ sound but as the q /tɕ^h/ sound by the native Mandarin listeners. The same patterns seem to hold true for the other retroflex sounds zh /tʂ/ and sh /ʂ/ that were identified by the native Mandarin listeners as the palatal sounds j /tɕ/ and q/tɕ^h/. These substitution patterns suggest the cues for the retroflex sounds were absent in these non-native productions. Therefore, it is very likely the misalignment between perception and production is partly due to the different mechanisms the learners attended to in perception and production of the target consonants. Acoustic analyses of the learners' productions of these sounds, along with perceptual test using synthesized stimuli manipulating the key acoustic cues differentiating the target categories are needed to draw a firm conclusion in future studies.

The current data also show that partial alignment between perception and production of Mandarin consonants does exist. The listeners' perception and production scores of both x /ɕ/ and c /ts^h/ were moderately but significantly correlated, indicating the link between perception and production. Past studies have come to the same conclusions of such partial alignment in perception and production of L2 sounds (Flege 1999; Rochet 1995).

L2 phonetic training studies have also examined the relationship between perception and production when assessing the outcomes of the training. There is evidence that perceptual training only led to improvement in both perception and production of L2 consonants (Bradlow et al. 1997) and lexical tones (Wang 2008, 2012, 2013; Wang et al. 2003), and production gains are larger on obstruents than on sonorants and vowels (Sakai and Moorman 2018). There is also evidence that trainees' perceptual learning did not lead to better productions on L2 vowel contrasts (Wang 2002). These findings suggest that the relationship between perception and production of L2 speech sounds can be further complicated by the different sound classes.

In conclusion, the current data showed partial alignment but more discrepancies on native English CFL learners' perception and production of the eight Mandarin consonants. Different phonetic mechanisms and strategies may be involved in the L2 speech sound perception and production. Perception may not always lead production. Future studies need to carry out more detailed acoustic analysis of the native and non-native productions of the target consonants to investigate specific problems that learners have in perception and production of Mandarin consonants.

Finally, one limitation of the current study was the exclusion of the s /s/ and r /ʐ̩a/ sounds in the analyses. It would have been better to include at least the s /s/ sound to have a nice set of dental fricative and affricates, making it parallel to the retroflex and palatal sets.

Author Contributions: Conceptualization, X.W. and J.C.; methodology, X.W. and J.C.; data collection and analysis, X.W. and J.C.; writing—original draft preparation, X.W.; writing—review and editing, X.W. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Research, Scholarly and Creative Activities (RSCA) award to the first author (2018–2019) by California State University, Fresno.

Acknowledgments: For data collection, we are grateful to Li Mann, Xinping Yu, and Pei Chen Chang for their support and their permission for us to recruit their students. We also thank all the participants for their participation.

Conflicts of Interest: The authors declare no conflict of interest.

References

Best, Catherine. 1994. The emergence of native-language phonological influences in infants: A perceptual assimilation model. In *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words*. Edited by Judith C. Goodman and Howard C. Nusbaum. Cambridge: MIT Press, pp. 167–244.

Best, Catherine T., Gerald W. McRoberts, and Elizabeth Goodell. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America* 109: 775–94. [CrossRef]

Bradlow, Ann R., David B. Pisoni, Reiko Akahane-Yamada, and Yoh'ichi Tohkura. 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America* 101: 2299–310. [CrossRef]

Chuang, Yu-Ying, Ching-Chu Sun, Janice Fon, and R. Harald Baayen. 2019. Geographical Variation of the Merging between Dental and Retroflex Sibilants in Taiwan Mandarin. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Edited by Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren. Canberra: Australasian Speech Science and Technology Association Inc., pp. 472–76.

Evans, Bronwen G., and Wafaa Alshangiti. 2018. The perception and production of British English vowels and consonants by Arabic learners of English. *Journal of Phonetics* 68: 15–31. [CrossRef]

Flege, James. 1995. Second language speech learning: Theory, findings, and problems. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Baltimore: York Press, pp. 233–77.

Flege, James. 1999. The relation between L2 production and perception. In *Proceedings of the XIVth International Congress of Phonetics Sciences*. Edited by John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granveille and Ashlee C. Bailey. College Park: American Institute of Physics, pp. 1273–76.

Flege, James. 2007. Language contacts in bilingualism: Phonetic systems interactions. *Laboratory Phonology* 9: 353–80.

Flege, James E., and Ratree Wayland. 2019. The role of input in native Spanish Late learners' production and perception of English phonetic segments. *Journal of Second Language Studies* 2: 1–44. [CrossRef]

Guion, Susan G., James E. Flege, Reiko Akahane-Yamada, and Jesica C. Pruitt. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America* 107: 2711–24. [CrossRef] [PubMed]

Hao, Yen-Chen. 2012. The Effect of L2 Experience on Second Language Acquisition of Mandarin Consonants, Vowels, and Tones. Ph.D. dissertation, Indiana University, Bloomington, IN, USA.

Lai, Yi-hsiu. 2009. Asymmetry in Mandarin affricate perception by learners of Mandarin Chinese. *Language and Cognitive Processes* 24: 1265–85. [CrossRef]

Lin, Hua. 2005. Understanding problems in learning Mandarin consonants by monolingual speakers of English. *Journal of Canadian Teachers of Chinese as a Second Language* 1: 1–18.

Lisker, Leigh, and Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurement. *Word* 20: 384–422. [CrossRef]

Lisker, Leigh, and Arthur S. Abramson. 1970. The voicing dimension: Some experiments on comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences*. Prague: Academia, vol. 563, pp. 563–77.

Liu, Jiang, and Allard Jongman. 2012. American Chinese learners' acquisition of L2 Chinese affricates/ts/and/tsh/. In *Proceedings of Meetings on Acoustics 164ASA*. Melville: Acoustical Society of America, vol. 18, p. 060005.

Munro, Murray, and Tracey M. Derwing. 2008. Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning* 58: 479–502. [CrossRef]

Munro, Murray, Tracey M. Derwing, and Ron I. Thomson. 2015. Setting segmental priorities for English learners: Evidence from a longitudinal study. *International Review of Applied Linguistics in Language Teaching* 53: 39–60. [CrossRef]

Polka, Linda, and Ocke-Schwen Bohn. 1996. A cross-language comparison of vowel perception in English-learning and German-learning infants. *The Journal of the Acoustical Society of America* 100: 577–92. [CrossRef] [PubMed]

Rochet, Bernard. 1995. Perception and production of second-language speech sounds by adults. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Baltimore: York Press, pp. 379–410.

Sakai, Mari, and Colleen Moorman. 2018. Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics* 39: 187–224. [CrossRef]

Strange, Winifred. 1995. Cross-language studies of speech perception: A historical review. In *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Edited by Winifred Strange. Baltimore: York Press, pp. 3–45.

Wang, Xinchun. 1997. The Acquisition of English Vowels by Mandarin ESL Learners: A Study of Production and Perception. Master's thesis, Simon Fraser University, Burnaby, BC, Canada.

Wang, Xinchun. 2002. Training Mandarin and Cantonese Speakers to Identify English Vowel Contrasts: Long-Term Retention and Effects on Production. Doctoral dissertation, Simon Fraser University, Burnaby, BC, Canada.

Wang, Xinchun. 2006. Perception of L2 tones: L1 lexical tone experience may not help. Paper presented at Third International Conference on Speech Prosody, Dresden, Germany, May 2–5; pp. 85–88.

Wang, Xinchun. 2008. Training for learning Mandarin tones. In *Handbook of Research on Computer-Enhanced Language Acquisition and Learning*. Edited by Felicia Zhang and Beth Barber. Hershey: IGI Global, pp. 259–74.

Wang, Xinchun. 2012. Auditory and Visual Training on Mandarin Tones: A Pilot Study on Phrases and Sentences. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 2: 16–29. [CrossRef]

Wang, Xinchun. 2013. Perception of Mandarin Tones: The Effect of L1 Background and Training. *The Modern Language Journal* 97: 144–160. [CrossRef]

Wang, Xinchun, and Jidong Chen. 2019. English Speakers' Perception of Mandarin Consonants: The Effect of Phonetic Distances and L2 Experience. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Edited by Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren. Canberra: Australasian Speech Science and Technology Association Inc., pp. 250–54.

Wang, Xinchun, and Murray Munro. 2004. Computer-based training for learning English vowel contrasts. *System* 32: 539–52. [CrossRef]

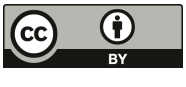
Wang, Yue, Allard Jongman, and Joan A. Sereno. 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after training. *The Journal of the Acoustical Society of America* 113: 1033–43. [CrossRef] [PubMed]

Werker, Janet. 1994. Cross-language speech perception: Developmental change does not involve loss. In *The Development of Speech Perception: The Transition from Speech sounds to Spoken Words*. Edited by Judith C. Goodman and Howard C. Nusbaum. Cambridge: The MIT Press, pp. 93–120.

Werker, Janet, and Linda Polka. 1993. Developmental changes in speech perception: New challenges and new directions. *Journal of Phonetics* 21: 83–101. [CrossRef]

Yang, Chunsheng, and Alan C. L. Yu. 2019. The acquisition of Mandarin affricates by American L2 learners. *Taiwan Journal of Linguistics* 17: 91–122.

Zhu, Liyi. 2012. Retroflex and Non-Retroflex Merger in Shanghai Accented Mandarin. Master's thesis, University of Washington, Seattle, WA, USA.



languages

MDPI

Article

Relating Lexical Access and Second Language Speaking Performance

Yu Liu

Department of Asian and Near Eastern Languages, Brigham Young University, Provo, UT 84602, USA; rachelyuliu@byu.edu

Received: 11 February 2020; Accepted: 9 April 2020; Published: 13 April 2020

Abstract: Vocabulary plays a key role in speech production, affecting multiple stages of language processing. This pilot study investigates the relationships between second language (L2) learners' lexical access and their speaking fluency, speaking accuracy, and speaking complexity. Fifteen L2 learners of Chinese participated in the experiment. A task-specific, native-referenced vocabulary test was used to measure learners' vocabulary size and lexical retrieval speed. Learners' speaking performance was measured by thirteen variables. The results showed that lexical access was significantly correlated with learners' speech rate, lexical accuracy, syntactic accuracy, and lexical complexity. Vocabulary size and lexical retrieval speed were significant predictors of speech rate. However, vocabulary size and lexical retrieval speed each affected learners' speaking performance differently. Learners' speaking fluency, accuracy, and complexity were all affected by vocabulary size. No significant correlation was found between lexical retrieval speed and syntactic complexity. Findings in this study support the Model of Bilingual Speech Production, revealing the significant role lexical access plays in L2 speech production.

Keywords: lexical access; second language speech; fluency; accuracy; complexity

1. Introduction

Vocabulary is fundamental for second language learning. Vocabulary knowledge has been identified as a good indicator of language proficiency (Milton 2013; Nation 2001). Research shows that second language (L2) learners with large or better developed vocabularies have better performance in reading (e.g., Albrechtsen et al. 2008; Jeon and Yamashita 2014; Laufer 1992; Qian 1999; Stæhr 2008), listening (e.g., Stæhr 2008, 2009), writing (e.g., Engber 1995; Grant and Ginther 2000; Schoonen et al. 2003; Stæhr 2008), and speaking (e.g., De Jong et al. 2013; Uchihara and Clenton 2018; Uchihara and Saito 2019).

According to Levelt's (1989, 1999) model of speech production, lexical access plays a key role in speech production, affecting multiple stages of language processing. Different examples of evidence supporting this claim have been found in empirical studies of L2 speaking:

(1) Vocabulary size and depth are associated with the overall scores of L2 learners' speaking proficiency (De Jong et al. 2012; Milton 2010);
(2) Both receptive and productive vocabulary knowledge can predict L2 speaking performance, especially fluency (Uchihara and Clenton 2018; Uchihara and Saito 2019);
(3) The speed and efficiency of lexical access affect L2 speaking fluency, allowing it to be used as a measure of L2 cognitive fluency (De Jong et al. 2013; Segalowitz and Freed 2004).

Productive vocabulary is closely related to task features such as topics and text types. L2 learners' vocabulary inventory is built up through many communicative tasks, depending on their target language contact experiences. It is important to consider task factor when we investigate the

relationship between lexical access and L2 language performance. Most previous studies related L2 learners' general vocabulary knowledge to their speaking performance in completing a small number of speaking tasks. The results might be affected by task selection. Increasing task number will improve the reliability of the experiments. However, the more tasks are included, the more overwhelmed participants will feel, which may affect the research results. Moreover, the type of word selection that best represents L2 learners' general vocabulary size is still debatable. Therefore, differently from previous studies examining general vocabulary size, in this study, we seek to narrow the investigation scope within specific tasks in order to look more closely into how lexical access interacts with L2 speaking performance within these tasks.

In this study, a task-specific vocabulary test was created based on native speakers' productive vocabulary in completing the same speaking tasks as those completed by L2 learners later. This test was used to explore L2 learners' lexical access. Two variables were used to measure lexical access: Vocabulary size (the number of words a learner knows) and lexical retrieval speed (how fast a learner recognizes and processes the words he/she knows). The purpose of this research is to investigate how lexical access relates to second language speaking performance.

2. L2 Speech Production

Levelt's model of speech production (1989, 1999) presents the process of first language (L1) speech production. The processing starts with conceptualizing the message. A pre-verbal plan consisting of concepts and language cues is generated at the conceptionalizer stage. Such messages are then processed at the formulator stage through lexico-grammatical encoding, morpho-phonological encoding, and phonetic encoding. The mental lexicon, including individual words and formulaic sequences, is activated to match with conceptual specifications and the language cues. At the articulator stage, words and sentences are executed as meaningful speech. During the whole process, speakers continuously monitor both their internal speech and overt speech output, modifying their speech as needed.

Based on his theory, Kormos (2006) proposed the Model of Bilingual Speech Production (see Figure 1). Two characteristics of L2 speech production are highlighted in this model:

(1) *L1 influence*. As proposed by Kormos (2006), L1 and L2 concepts, lexemes (word forms), lemmas (syntactic and morphological features), syllable programs, and grammatical rules are stored together. In speech production, L1 and L2 lexical items and rules are activated together and compete for selection. When L1 knowledge and encoding procedures are transferred during L2 processing, the efficiency of L2 speech production may be negatively affected.

(2) *Serial processing*. Compared to L1 automatized speech production, language processing in L2 requires more attention and is less automatized. L2 learners hold a great amount of L2 declarative rules in their long-term memory. Before declarative rules turn into procedural knowledge, L2 learners at a lower level of proficiency have to use limited attention to process at lexical, syntactic, and phonological levels as well as in monitoring. This causes L2 speech to be less fluent and more open to L1 influence (Kormos 2006; Vieira 2017).

In addition, Kormos (2006) emphasized that the L2 speech production system is lexically driven. The mental lexicon, including concepts, lemmas, and lexemes, affects conceptualization, lexico-grammatical encoding, and morpho-phonological encoding of phrases. Lexical access in the second language is an attention-demanding cognitive process. On the one hand, L2 learners face the challenge of retrieving the correct L2 words from the competition of L1 and related L2 items. On the other hand, time constraints in real communication require learners to process language efficiently, which increases the pressure of the cognitive process. As a result, the efficiency of lexical access influences the quality of L2 speech.

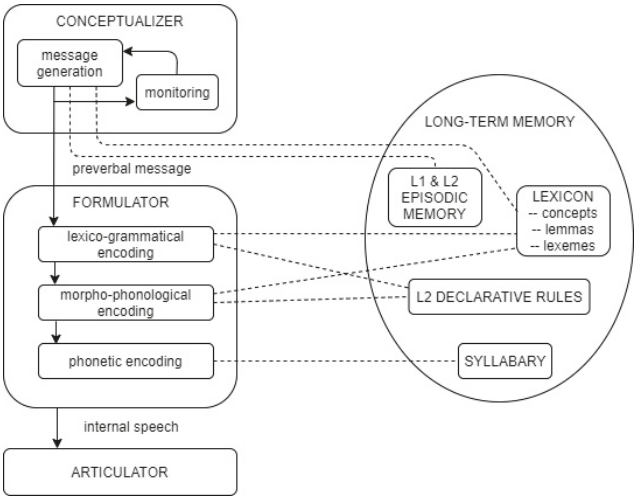


Figure 1. The Model of Bilingual Speech Production (adapted from Kormos 2006, p. 168).

3. Vocabulary Knowledge and L2 Speech

Lexical access is multifaceted. In literature, three categories have been distinguished in measuring vocabulary knowledge in literature (Anderson and Freebody 1981; Daller et al. 2007; Meara 1996; Milton and Fitzpatrick 2013):

1. *Vocabulary breadth*: The number of words a learner knows regardless of the form they are known in or how well they are known. Vocabulary breadth is also referred to as vocabulary size. Three forms of measurement have been found in literature: Yes/No format task (Uchihara and Clenton 2018), writing L2 forms corresponding to L1 meaning (Koizumi and In'nami 2013), and filling in the missing words in a sentence where the first letters are given (De Jong et al. 2013). Word selection in most studies is based on word frequency bank lists.
2. *Vocabulary depth*: How well or how completely words are known. Vocabulary depth is a rich concept that consists of various aspects. According to Nation's (2001, p. 27) description of "what is involved in knowing a word", vocabulary knowledge includes form (spoken, written, word parts), meaning (form and meaning, concepts and referents, associations), and use (grammatical functions, collocations, constraints on use). Read (2004) proposed that vocabulary involves word form and meaning, as well as associational knowledge, collocation knowledge, inflectional knowledge, and derivational knowledge. Meara and Wolter (2004) extended the vocabulary depth by including knowing the network words. Measuring vocabulary depth is less manageable because it is difficult to find a concept that holds together the variety of elements (Milton 2010).
3. *Vocabulary fluency*: The automaticity with which the words a person knows can be recognized and processed. It is also referred to as processing speed or lexical retrieval speed. Reaction time (RT) is recorded to measure vocabulary fluency in a vocabulary test (De Jong et al. 2013; Koizumi and In'nami 2013).

Previous studies have found a close relationship between L2 learners' vocabulary knowledge and their speech production. Some studies suggest that learners with larger receptive vocabulary sizes are more proficient in speaking (De Jong et al. 2013; Hilton 2008; Koizumi and In'nami 2013; Uchihara and Clenton 2018). Koizumi and In'nami (2013) found that vocabulary size explained up to 60% of the variance in speaking proficiency. However, Uchihara and Clenton (2018) found less predictive power in vocabulary size for L2 speaking, which was only 29%. The former study used an automated scoring system, whereas the latter used human ratings. Though Uchihara and Clenton

(2018) found a significant correlation between receptive vocabulary size and spoken lexical use based on human ratings, they also noticed that receptive vocabulary size and lexical sophistication measures were not significantly correlated. Their findings indicate that learners with larger vocabulary sizes do not necessarily produce more advanced words. De Jong et al. (2013) and Hilton (2008) focused on speaking fluency only. Both studies integrated lexical knowledge into a battery of tests that tested L2 learners' linguistic skills and discussed their relationship with L2 speaking fluency. The results in De Jong et al.'s (2013) study suggested that all measures of utterance fluency (e.g., speech rate, number of silent and filled pauses, repetitions, and repairs) were affected by linguistic knowledge. Hilton (2008) had similar findings in terms of temporal measures of L2 speaking fluency. He argued that the lack of lexical knowledge appeared to be the primary cause of the most serious disfluencies, and that it was the greatest impediment to L2 speaking fluency.

Among all studies, De Jong et al. (2013) and Koizumi and In'nami (2013) investigated more than one category of vocabulary knowledge. As De Jong et al. (2013) suggested, lexical retrieval speed was strongly correlated with all measures of L2 speaking fluency (e.g., number of silent pauses, filled pauses, repetitions, duration of silent pauses, speech rate). Koizumi and In'nami (2013) focused on three categories of vocabulary knowledge (e.g., size, depth, fluency). They found that vocabulary size and vocabulary depth substantially predicted L2 proficiency, while the correlation between vocabulary fluency and L2 speaking was much weaker. Koizumi and In'nami's (2013) findings were based on automatic ratings of L2 learners' speaking performance instead of human ratings or objective measures. The reliability of the automated scoring system may have had an effect on the findings.

Another area of vocabulary knowledge research measures productive vocabulary knowledge under the distinction of receptive vocabulary knowledge (passive knowledge, recognition) and productive knowledge (active knowledge, use). Only one study discussed how productive vocabulary knowledge affected L2 speaking performance (Uchihara and Saito 2019). In their study, the Lex30 test was used to investigate the productive mental lexicon. This test used a word association format, presenting learners with a list of 30 stimulus words and instructing them to write down the first four words in the target language they thought of when they read each word in the list (Meara and Fitzpatrick 2000). The data demonstrated that productive vocabulary knowledge could predict how fluently L2 learners can speak, but it was not significantly correlated with comprehensibility or accentedness. The results suggested that developed L2 lexicons led to less difficulty in retrieving L2 words, which therefore helped learners produce fluent speech.

4. Research Questions

To our knowledge, only a limited number of studies linked L2 learners' vocabulary knowledge with their speaking fluency, accuracy, and complexity, respectively (De Jong et al. 2013; Koizumi and In'nami 2013). No such study has been done in the research of Chinese as a second language (L2 Chinese), although some studies have discussed the relationship between vocabulary size and L2 Chinese reading (Wu 2016, 2017). In view of the important role that lexical access plays in second language speech production (Kormos 2006), this study aims at exploring the dynamics between L2 learners' lexical access, measured by vocabulary size and lexical retrieval speed, and three dimensions of their speaking performance: Fluency, accuracy, and complexity.

The following research questions are addressed in the present study:

1. How does lexical access affect the fluency in L2 learners' speech in four speaking tasks?
2. How does lexical access affect the accuracy in L2 learners' speech in four speaking tasks?
3. How does lexical access affect the complexity in L2 learners' speech in four speaking tasks?

5. Methodology

5.1. Participants

Fifteen English native speakers participated in the experiment. It was a small homogeneous group. According to the background questionnaire survey, these participants had very similar language-learning backgrounds. They have attended the same Chinese classes at the same university. All of the participants were enrolled in a third-year Chinese course at a U.S. university when the experiment was conducted. According to the instructor of the course, these participants' Chinese proficiency levels ranked at ACTFL intermediate-high to advanced-low level (ACTFL 2012). Their ages ranged between 20 and 25. Five of them were female. All of the participants replied to our call for participation on a voluntary basis. They gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of #15405.

5.2. Materials

A task-specific, native-referenced vocabulary test was created to investigate how lexical access interacts with L2 speaking performance within these tasks. Before the experiment, we invited six Chinese native speakers from the same university (ages 19–22, four females and two males) to complete four speaking tasks, which were also used to test fifteen L2 learners. By doing this, we were able to set up the reference on the basis of native speakers' productive vocabulary.

Four speaking tasks represented four text types, carrying four different communicative functions: Instructive, descriptive, explanatory, and argumentative. In the first task, both Chinese native speakers and L2 learners were asked to introduce the city where the university was located. In the second task, they described their first day at the university. In the third task, they were presented a data chart and were invited to explain the income gap between males and females of different age groups. In the last task, they talked about their opinions on a given topic, specifically "What kind of professors are good college professors?" These tasks were not culturally specific. The native speakers and the participants shared similar experiences at the same university. We assumed that most of the vocabulary output by these two groups should be within a limited range when they completed the same tasks.

Based on the vocabulary that the Chinese native speakers used in the four tasks, we compiled a list of vocabulary (198 items) that was most commonly used (being used at least three times by different speakers). All words were listed in a random order, controlling the effect derived from word frequency and task order. The list was then translated into English by the researcher for the vocabulary test.

5.3. Procedure

There were two parts in the experiment. The first part was a vocabulary test. In this part, participants were instructed to translate the words on the vocabulary list orally from English into Chinese as fast as possible. They were instructed to respond "I don't know" if they did not know the answer. They were not given pre-task planning time. The whole process was timed in order to record lexical retrieval speed. If participants were able to say the target words or their synonyms, the answers were rated as correct. Otherwise, the answers were rated as incorrect. Two L2 Chinese teachers rated the vocabulary test independently. There was no disagreement between the two ratings.

Participants took a five-minute break after completing the first part and then continued to finish the second part. The second part was a speaking test consisting of four monologue tasks. Participants completed four speaking tasks, which were the same as the ones completed by the six Chinese native speakers. For each task, participants had one minute to prepare and ten minutes to speak. Participants' speech was recorded through a recording software "Audacity" with the setting of stereo 44,100 Hz. The experiment was conducted in the researcher's office individually, administered by the researcher. The total time commitment for each participant in this experiment was about 1.5 to 2 h.

5.4. Measures and Statistical Procedures

L2 learners' lexical access is represented by both vocabulary size and lexical retrieval speed. Vocabulary size was measured based on the accuracy rate in the vocabulary test. Lexical retrieval speed was measured by calculating the average response time for each word in the vocabulary test. As for speaking performance, all participants' speech samples were first transcribed by a Chinese native speaker. Afterwards, they were encoded manually by the researcher as described in detail below. Then, thirteen variables of the following three categories were measured for statistical analysis:

Fluency: We measured three facets of speaking fluency (Tavakoli and Skehan 2005): Speed fluency (speech rate, mean length of runs); breakdown fluency (mean length of silent pauses, number of silent pauses, number of filled pauses), and repair fluency (number of disfluencies). A script programmed in PRAAT (Boersma and Weenink 2009) was used to detect silent pauses. Minimum silence duration was set to 350 milliseconds. We were therefore able to measure speech rate, mean length of runs, mean length of silent pauses, and the number of silent pauses. Filled pauses, such as *en* (嗯 "um"), *ranhou* (然后 "and then"), *jiushi* (就是 "that is"), and *nage* (那个 "that"), as well as disfluencies, such as repetitions, restarts, or repairs, were extracted manually from the transcripts of the speech samples. The number of filled pauses and the number of disfluencies were then calculated.

Accuracy: All speech samples were manually divided into different AS-units (main clauses and any attached subordinate clauses or sub-clausal units). They were also manually encoded by tagging lexical errors and syntactic errors. Lexical accuracy and syntactic accuracy were then calculated.

Complexity: Syntactic complexity was measured with two methods. The first method of measurement was to calculate the number of clauses, with the second method measuring the average sentence length. Lexical complexity was measured from three aspects: Lexical diversity, word frequency, and word difficulty. Guiraud's Index (Guiraud 1960) was used to measure lexical diversity. Word frequency was measured based on the SUBTLEX-CH corpus (Cai and Brysbaert 2010), whereas word difficulty was measured based on a reference word list designed for the standardized Chinese proficiency test Hanyu Shuiping Kaoshi (HSK): The HSK word difficulty ranking (Hanban 2012).

Table 1 lists the calculation methods used to measure L2 speaking performance in this study.

Table 1. Measures of speaking performance and their calculation methods.

		Variables	Calculation Methods
Fluency	Speed fluency	Speech rate	The total number of syllables divided by total time.
		Mean length of runs	The average number of syllables produced in utterances between pauses of 0.25 s and above.
	Breakdown fluency	Mean length of silent pauses	The total length of pauses above 0.2 s divided by the total number of pauses above 0.2 s.
		Number of silent pauses	The total number of pauses over 0.2 s divided by the total amount of time spent expressed in seconds and is multiplied by 60.
		Number of filled pauses	The total number of filled pauses divided by the total amount of time spent expressed in seconds and is multiplied by 60.
	Repair fluency	Number of disfluencies	The total number of disfluencies, such as repetitions, restarts, and repairs, divided by the total amount of time expressed in seconds and multiplied by 60.
Accuracy		Syntactic accuracy	The total number of correct AS units divided by the total number of AS units.
		Lexical accuracy	The total number of correct words divided by the total number of all words.

Table 1. *Cont.*

	Variables	Calculation Methods
Complexity	Number of clauses	The total number of clauses divided by the total number of AS units.
	Average sentence length	The total number of words divided by the total number of AS units.
	Lexical diversity	The number of types/the square root of the number of tokens.
	Word frequency	The average number of word frequency ranking of all words based on the SUBLEXT-CN word frequency ranking.
	Word difficulty	The average number of word difficulty rankings of all words based on the HSK word difficulty ranking.

6. Results

6.1. Vocabulary Test

In the experiment, it took participants an average of 516.5 s to complete the vocabulary test. The average reaction time for each word was 2.6 s. All participants correctly translated at least half of the words in the vocabulary list, with the accuracy rate ranging from 58% to 96%. The average accuracy rate was 81%, which indicates that these learners are familiar with most of the words on the list. However, none of them could translate all of the words correctly. Table 2 shows each participant's accuracy and average reaction time for each word in the vocabulary test. This accuracy represents the learners' task-specific receptive vocabulary size. The reaction time shows how fast lexical retrieval was. The stronger the link between the conceptual messages and the L2 lexical items (e.g., the concept of causal relation "because" and the Chinese words "因为 *yinwei*"), the more words were translated correctly, and the faster the reaction time was.

Table 2. Accuracy and average reaction time in the vocabulary test.

Students (N = 15)	Accuracy	Reaction Time per Word (s)
1	0.68	3.3
2	0.58	4.3
3	0.88	2.84
4	0.85	1.68
5	0.74	2.72
6	0.84	2.96
7	0.89	1.95
8	0.91	2.13
9	0.73	2.34
10	0.78	2.43
11	0.67	3.36
12	0.95	2.03
13	0.75	2.24
14	0.95	2.08
15	0.96	2.78
Mean (SD)	0.81 (0.12)	2.61 (0.68)

6.2. Vocabulary Size and L2 Speaking Performance

Table 3 shows L2 learners' speaking performance in terms of fluency, accuracy, and complexity in four speaking tasks. To determine the degree of the relationship between vocabulary size and all measures of learners' speaking performance, Pearson's correlations were calculated. The Appendix A

presents the correlations among vocabulary size, lexical retrieval speed, and all of the measures of speaking performance. A significant correlation was found between task-specific vocabulary size and all fluency measures except for the number of filled pauses (r = 0.012, p = 0.926). In particular, L2 learners' vocabulary size was strongly correlated with speed fluency (speech rate, r = 0.375, p = 0.003; mean length of runs, r = 0.354, p = 0.005), with breakdown fluency (mean length of silent pauses, r = −0.256, p = 0.048; number of silent pauses, r = −0.35, p = 0.006), and with repair fluency (number of disfluencies, r = −0.285, p = 0.027).

Table 3. Participants' speaking performance in four speaking tasks (N = 15).

	Task 1		Task 2		Task 3		Task 4		Average of All Tasks	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1. Speech rate	2.49	0.47	2.69	0.64	2.14	0.49	2.62	0.49	2.48	0.52
2. Mean length of runs	0.79	0.62	0.82	0.72	0.67	0.20	0.64	0.19	0.73	0.43
3. Mean length of silent pauses	106.44	54.35	98.44	37.45	98.90	39.16	102.70	40.11	101.62	42.77
4. Number of silent pauses	0.54	0.21	0.47	0.16	0.61	0.18	0.56	0.17	0.55	0.18
5. Number of filled pauses	9.97	5.52	8.74	4.45	10.29	4.20	8.85	5.47	9.46	4.91
6. Number of disfluencies	5.75	3.07	5.25	2.97	4.11	2.58	5.51	2.18	5.15	2.70
7. Syntactic accuracy	0.81	0.15	0.87	0.09	0.79	0.11	0.87	0.13	0.84	0.12
8. Lexical accuracy	0.988	0.006	0.992	0.008	0.990	0.009	0.990	0.009	0.990	0.008
9. Number of clauses	1.43	0.27	1.40	0.22	1.49	0.25	1.64	0.44	1.49	0.30
10. Sentence length	23.25	5.22	22.99	4.34	30.40	5.07	30.26	6.96	26.73	5.40
11. Lexical diversity	6.78	0.85	6.95	0.70	6.27	1.18	6.55	1.05	6.64	0.95
12. Word frequency	1924.2	908.3	1985.1	739.6	3028.3	1313.5	1930.0	632.7	2216.9	898.5
13. Word difficulty	2.45	0.24	2.38	0.21	2.48	0.21	2.39	0.17	2.42	0.20

Learners' vocabulary size was also significantly correlated with learner's speaking accuracy on both the lexical level (r = 0.377, p = 0.027) and the syntactic level (r = 0.313, p = 0.015).

In regard to speaking complexity, no significant correlation was found between learners' vocabulary size and syntactic complexity with two different measures (number of clauses, r = −0.23, p = 0.077; average sentence length, r = −0.098; p = 0.458). Regarding lexical complexity, mixed results were found based on different measures. The results showed that learners' vocabulary size was closely related to the lexical diversity measure (r = 0.413, p = 0.001) and to word difficulty measure (r = 0.397, p = 0.002), but not to the word frequency measure (r = 0.145, p = 0.271).

By using HSK word difficulty rankings to compare words produced by native speakers (NSs) and L2 learners (NNSs) in four speaking tasks, it was found that the word distributions of different groups had different patterns. Figure 2 showed the word distributions of two groups. Most of the words that L2 learners used were level 1 to level 3 words (35.06%, 22.1%, 19.36%), which were the simpler words. They also used some level 4 words (13.92%) and a few words from level 5, level 6, or above (7.03%, 2.53%), which were more advanced words. Differently from L2 learners, native speakers used mostly advanced words, especially words ranked level 3 and above (22.22%, 26.26%, 24.75%, 18.68%). Level 1 and Level 2 words were not frequently used by native speakers (1.52%, 6.57%), as opposed to L2 learners. When compared to native speakers, L2 learners generally used less advanced words. However, the more words they acquired, the more advanced and more diverse words they were likely to use. L2 learners' limited vocabulary size affected their word choice, which, as a result, might affect the quality of their speech.

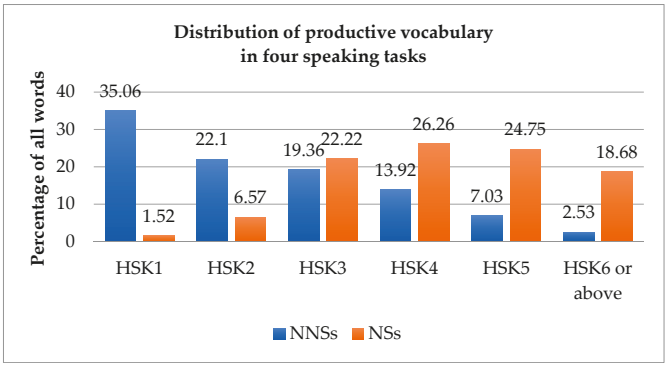


Figure 2. Distribution of words produced by second language (L2) learners (NNSs) and native speakers (NSs) based on the Hanyu Shuiping Kaoshi (HSK) difficulty ranking.

6.3. Speed of Lexical Access and L2 Speaking Performance

Pearson's correlations demonstrated a high correlation between task-specific lexical retrieval speed, which was measured by the average reaction time for each word in the vocabulary test, and speech rate ($r = -0.379$, $p = 0.003$). No significant correlation was found between lexical retrieval speed and other fluency measures, including mean length of runs ($r = 0.076$, $p = 0.565$), breakdown fluency (mean length of silent pauses, $r = 0.023$, $p = 0.862$; number of silent pauses, $r = 0.147$, $p = 0.262$; number of filled pauses, $r = -0.127$, $p = 0.332$), or with repair fluency (number of disfluencies, $r = 0.189$, $p = 0.148$).

Learners' lexical retrieval speed was significantly correlated with their speaking accuracy and complexity only on the lexical level, not on the syntactic level. Significant correlation was found between lexical retrieval speed and lexical accuracy ($r = -0.308$; $p = 0.017$), lexical diversity ($r = -0.365$; $p = 0.004$), and word difficulty ($r = -0.424$; $p = 0.001$), but not with word frequency measure ($r = 0.177$; $p = 0.176$). In terms of syntactic measures, no significant correlation was found between lexical retrieval speed and syntactic accuracy ($r = -0.192$; $p = 0.142$), syntactic complexity ($r = 0.176$; $p = 0.178$), or average sentence length ($r = 0.022$; $p = 0.867$).

6.4. Lexical Access and L2 Speaking Performance

A multiple linear regression was conducted to examine the role of L2 learners' lexical access in explaining their speaking fluency, accuracy, and complexity. Two variables of lexical access, vocabulary size and lexical retrieval speed, were entered into the regression equations to determine their contributions to the variance of thirteen measures of speaking performance. The results suggested that vocabulary size and lexical retrieval speed were significant predictors of speech fluency, explaining 16.7% of the variance of speech rate and 18.5% of the variance of the mean length of runs. Vocabulary size and lexical retrieval speed also had a significant predictive effect on breakdown fluency, accounting for 11.4% of the variance of the mean length of silent pauses and 14.2% of the variance of the number of silent pauses. Results showed that lexical access had no significant predictive effect on the number of filled pauses as well as the number of disfluencies (see Table 4).

In terms of accuracy, vocabulary size and lexical retrieval speed were found to be significant predictors of lexical accuracy, accounting for 14.6% of the variance. However, no significant predictive effect was found on syntactic accuracy (see Table 5).

As for complexity, vocabulary size and lexical retrieval speed were found to be significant predictors of lexical complexity, explaining 18.1% of the variance of lexical diversity and 19.9% of vocabulary difficulty. No significant predictive effect was found on vocabulary frequency. The results also suggested that lexical access had no significant predictive effect on syntactic complexity (see Table 6).

Table 4. Prediction of L2 speaking fluency by lexical access.

Measures of L2 Speaking Fluency	Lexical Access: Vocabulary Size; Lexical Retrieval Speed	*b*	SE *b*	*ß*	ΔR^2	F (2, 57)	*p*
1. Speech rate	Size	1.04	0.83	0.21			
	Speed	0	0	−0.23			
	Size * Speed				0.167	5.71	0.006 **
2. Mean length of runs	Size	2.56	0.72	0.59			
	Speed	0	0	0.34			
	Size * Speed				0.185	6.42	0.003 **
3. Mean length of silent pauses	Size	−176.43	65.19	−0.47			
	Speed	−0.1	0.06	−0.31			
	Size * Speed				0.114	3.68	0.031 *
4. Number of silent pauses	Size	−0.79	0.28	−0.49			
	Speed	0	0	−0.19			
	Size * Speed				0.142	4.7	0.013 *
5. Number of filled pauses	Size	−6.53	7.85	−0.15			
	Speed	−0.01	0.01	−0.23			
	Size * Speed				0.028	0.82	0.444
6. Number of disfluencies	Size	−7.21	4.28	−0.3			
	Speed	0	0	−0.02			
	Size * Speed				0.082	2.53	0.088

** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level.

Table 5. Prediction of L2 speaking performance by lexical access.

Measures of L2 Speaking Accuracy	Lexical Access: Vocabulary Size; Lexical Retrieval Speed	*b*	SE *b*	*ß*	ΔR^2	F (2, 57)	*p*
7. Syntactic accuracy	Size	0.39	0.2	0.35			
	Speed	0	0	0.06			
	Size * Speed				0.099	3.15	0.051
8. Lexical accuracy	Size	0.02	0.01	0.32			
	Speed	0	0	−0.08			
	Size * Speed				0.146	4.88	0.011 *

** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level.

Table 6. Prediction of L2 speaking performance by lexical access.

Measures of L2 Speaking Accuracy	Lexical Access: Vocabulary Size; Lexical Retrieval Speed	*b*	SE *b*	*ß*	ΔR^2	F (2, 57)	*p*
9. Number of clauses	Size	−0.58	0.5	−0.21			
	Speed	0	0	0.03			
	Size * Speed				0.053	1.61	0.209
10. Sentence length	Size	−12.7	10.44	−0.22			
	Speed	−0.01	0.01	−0.18			
	Size * Speed				0.026	0.76	0.475
11. Lexical diversity	Size	2.64	1.44	0.31			
	Speed	0	0	−0.15			
	Size * Speed				0.181	6.31	0.003 **
12. Vocabulary frequency	Size	−361.51	1653.43	−0.04			
	Speed	1.16	1.43	0.15			
	Size * Speed				0.032	0.95	0.395
13. Word difficulty	Size	0.35	0.3	0.19			
	Speed	0	0	−0.29			
	Size * Speed				0.199	7.08	0.002 **

** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level.

7. Discussion

The results in this study revealed that task-specific lexical access was closely related to all three dimensions of L2 speech: Fluency, accuracy, and complexity, though it is mainly on the lexical level. Vocabulary size and lexical retrieval speed were found to be significant predictors of speech rate, silent pauses, lexical accuracy, and lexical complexity. The explanatory power ranged from 11.4% to 19.9%. Among the three above outlined categories of speaking performance, fluency was most easily affected by vocabulary size and lexical retrieval speed.

In response to the first research question, "How does lexical access affect the fluency in L2 learners' speech in four speaking tasks?", the results showed that both vocabulary size and lexical retrieval speed were found to be highly correlated with speech rate ($p < 0.01$). L2 learners' speech rate was most easily affected by lexical access. A significant correlation was found between vocabulary size and most of the fluency measures, whereas lexical retrieval speed was not significantly correlated with other measures of speaking fluency. Those who knew more L2 words and had faster processing speed produced more fluent speech. This finding is aligned with previous studies (De Jong et al. 2013; Hilton 2008; Koizumi and In'nami 2013; Uchihara and Clenton 2018). This can be explained by the theory of automatization in Kormos's model (2006). L2 lexical access is an attention-demanding task. This is because in the L2 lexico-semantic system, the link between conceptual messages and L2 lexical items is weaker than that of L1. In addition, L1 and L2 words, as well as related L2 words, are activated and compete for selection. Moreover, the syntactic rules of lexical forms are stored as declarative knowledge, which requires attention control to be executed. Under the impact from the above factors, the L2 speech production process is serial, rather than automatized, which causes L2 speech to be less fluent.

Answering the second research question, "How does lexical access affect the accuracy in L2 learners' speech in four speaking tasks?", the results demonstrated that both vocabulary size and lexical retrieval speed were significantly correlated with lexical accuracy ($p < 0.05$). Syntactic accuracy was significantly correlated with vocabulary size ($p < 0.05$), but not with lexical retrieval speed. Those who knew more L2 words and had faster processing speed spoke more accurately. It should be noticed that, although in this study knowing more words resulted in more accurate sentences, language processing on the syntactic level is very complicated because knowing L2 words not only means matching the concepts and the L2 lexical forms, but also means being aware of these words' grammatical rules. Since L2 rules are stored as declarative knowledge in learners' long-term memory, learners often encounter challenges in selecting the correct words among synonyms and using them correctly in sentences. It is necessary for future studies to investigate the relationship between vocabulary depth and L2 speaking performance.

In terms of the third question, "How does lexical access affect the complexity in L2 learners' speech in four speaking tasks?", it was found that vocabulary size was highly correlated with lexical diversity and word difficulty ($p < 0.01$). Those who acquired more L2 words and had faster processing speed produced more diverse and more advanced words as they spoke. This finding differs from that of Uchihara and Clenton's study (2018). In their study, the receptive vocabulary size and lexical sophistication measures were not significantly correlated. The reason may be that the word selection methods used in the vocabulary test were different in both studies. In their study, researchers used 100 real and 100 imaginary words to create the word list. Participants were asked to judge which words were real and which ones were imaginary. Judging real words reflected L2 learners' word recognition ability, whereas translating words from L1 to L2 reflected the effectiveness of mapping concepts and L2 words.

Furthermore, two different patterns of word distribution in L1 and L2 were found in completing the same tasks. L2 learners tended to use more lower-level words: The more advanced the words, the less they were used. Word distribution was presented as a curve from high to low. On the contrary, most of the words used by native speakers were more advanced words (HSK level 3 and above).

The word distribution curve went from low to high and turned into a flat line, showing a balanced distribution from level 3 to higher-level words.

According to the Model of Bilingual Speech Production (Kormos 2006), conceptualization, lexico-grammatical encoding, and morpho-phonological encoding are all directly affected by lexical access. L2 speech production is a lexical-driven process. Retrieving vocabulary from long-term memory in the process of second speech production has been claimed to be less automatic, serial processing instead of automatic, parallel processing. The evidence in this study supports this model. Learners' performance in the vocabulary test revealed their ability to retrieve L2 vocabulary, which was shown to be significantly correlated with the quality of their speaking performance. The more efficient lexical retrieval was, the faster the lexico-grammatical encoding and morpho-phonological encoding were, and the more fluent speech was produced by L2 learners. When L2 Learners encountered difficulty in retrieving the corresponding L2 words or it took longer to successfully retrieve them, their speech was less fluent, less accurate, and simpler. The evidence in this study provides details of how L2 learners' fluency, accuracy, and complexity are affected by their capability or incapability of retrieving vocabulary in their L2 speech. The importance of vocabulary learning in second language acquisition is echoed in this study.

8. Limitations and Directions for Future Research

In this study, we adopted a task-specific, native-referenced approach in vocabulary test design, which is rarely seen in current related literature. Rather than testing L2 learners' general vocabulary size, we focused on learners' vocabulary size within specific tasks. We believe that by using this more focused approach, the dynamics between L2 learners' lexical access and their speaking performance can be presented more clearly. However, there are also limitations in doing so. One limitation is that the number of vocabularies for investigation is limited. Future studies can include more tasks with different topics and text types. It is also necessary to consider individual differences of productive vocabulary in completing the same tasks, especially the "vocabulary gap" between native speakers and non-native speakers in relation to native-referenced vocabulary test design. To improve the reliability of the experiment for future work, a larger number of participants at different proficiency levels can be included to control the effect from individual differences. With a larger scale of investigation and duplicated studies, we would be able to further the discussion of how lexical access affects L2 speech. It would also be useful to compare the word distribution of native speakers and non-native speakers by using a computer-coding approach, in order to explore more deeply the link between receptive vocabulary knowledge and productive vocabulary knowledge.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Pearson's correlations among vocabulary size, lexical retrieval speed, and all of the measures of speaking performance.

	Accuracy of Vocabulary Test	Speed of Lexical Access	Speech Rate	Mean Length of Runs	Mean Length of Silent Pauses	Number of Silent Pauses	Number of Filled Pauses	Number of Disfluencies	Syntactic Accuracy	Lexical Accuracy	Number of Clauses	Sentence Length	Lexical Diversity	Lexical Richness	Lexical Complexity
Accuracy of vocabulary test															
Speed of lexical access	−0.703 **														
Speech rate	0.375 **	−0.379 **													
Mean length of runs	0.354 **	−0.076	0.255 *												
Mean length of silent pauses	0.256 *	0.023	−0.236	−0.732 **											
Number of silent pauses	−0.350 **	0.147	−0.568 **	−0.526 **	0.461 **										
Number of filled pauses	0.012	−0.127	−0.068	−0.404 **	0.589 **	0.12									
Number of disfluencies	−0.285 *	0.189	0.098	−0.097	0.131	0.008	−0.212								
Syntactic accuracy	0.313 *	−0.192	0.123	0.008	−0.062	−0.302 *	0.257 *	−0.297 *							
Lexical accuracy	0.377 **	−0.308 *	0.248	0.085	−0.081	−0.166	0.222	−0.306 *	0.245						
Number of Clauses	−0.23	0.176	0.071	−0.06	−0.024	0.105	−0.239	0.304 *	−0.482 **	−0.377 **					
Sentence length	−0.098	−0.022	−0.128	−0.165	−0.063	0.21	−0.126	0.21	−0.284 *	−0.155	0.68 5 **				
Lexical diversity	0.413 **	−0.365 **	0.228	0.181	−0.139	−0.071	−0.008	−0.407 **	0.358 **	0.039	−0.200	−0.290 *			
Word frequency	−0.145	0.177	−0.047	0.06	−0.292 *	−0.034	−0.192	−0.04	−0.243	−0.133	0.131	0.235	−0.222		
Word difficulty	0.397 **	−0.424 **	0.325 *	0.440 **	−0.327 *	−0.268 *	−0.222	−0.137	0.005	−0.115	0.2	0.19	0.316 *	−0.024	

** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level.

References

ACTFL. 2012. *ACTFL Proficiency Guidelines—Speaking*, 3rd ed. Alexandria: American Council on the Teaching of Foreign Languages. Available online: https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking (accessed on 29 January 2019).

Albrechtsen, Dorte, Kirsten Haastrup, and Birgit Henriksen. 2008. *Vocabulary and Writing in a First and Second Language: Processes and Development*. New York: Palgrave Macmillan.

Anderson, Richard C., and Peter Freebody. 1981. Vocabulary knowledge. In *Comprehension and Teaching: Research Reviews*. Edited by John T. Guthrie. Newark: International Reading Association, pp. 77–117.

Boersma, Paul, and David Weenink. 2009. *Praat: Doing Phonetics by Computer* [computer program]. Version 5.1.17. Available online: http://www.praat.org (accessed on 27 January 2019).

Cai, Qing, and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE* 5: e10729. [CrossRef] [PubMed]

Daller, Helmut, James Milton, and Jeanine Treffers-Daller, eds. 2007. *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press.

De Jong, Nivja H., Margarita P. Steinel, Arjen F. Florijn, Rob Schoonen, and Jan H. Hulstijn. 2012. Facets of speaking proficiency. *Studies in Second Language Acquisition* 34: 5–34. [CrossRef]

De Jong, Nivja H., Margarita P. Steinel, Arjen Florijn, Rob Schoonen, and Jan H. Hulstijn. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics* 34: 893–916. [CrossRef]

Engber, Cheryl A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4: 139–55. [CrossRef]

Grant, Leslie, and April Ginther. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9: 123–45. [CrossRef]

Guiraud, Pierre. 1960. *Problèmes et méthodes de la statistique linguistique*. Paris: Presses Universitaires de France.

Hanban. 2012. HSK (Chinese as a second language proficiency test: Hanyu Shuiping Kaoshi) Vocabulary. [新汉语水平考试(HSK)词汇(2012年修订版)]. Available online: http://www.chinesetest.cn/userfiles/file/HSK/HSK-2012.xls (accessed on 27 January 2019).

Hilton, Heather. 2008. The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal* 36: 153–66. [CrossRef]

Jeon, Eun Hee, and Junko Yamashita. 2014. L2 reading comprehension and its correlates: A meta-analysis. *Language Learning* 64: 160–212. [CrossRef]

Koizumi, Rie, and Yo In'nami. 2013. Vocabulary knowledge and speaking proficiency among Second Language Learners from Novice to Intermediate Levels. *Journal of Language Teaching & Research* 4: 900–13.

Kormos, Judit. 2006. *Speech Production and Second Language Acquisition*. Abingdon: Routledge.

Laufer, Batia. 1992. How much lexis is necessary for reading comprehension? In *Vocabulary and Applied Linguistics*. Edited by Pierre J. L. Arnaud and Henri Béjoint. London: Macmillan International Higher Education, pp. 126–32.

Levelt, Willem J. M. 1989. *Speaking: From Intention to Articulation*. Cambridge: MIT Press.

Levelt, Willem J. M. 1999. Producing spoken language. In *The Neurocognition of Language*. Edited by Colin M. Brown and Peter Hagoort. Oxford: Oxford University Press, pp. 83–122.

Meara, Paul. 1996. The Vocabulary Knowledge Framework. Available online: http://www.swan.ac.uk/cals/calsres/vlibrary/pm96d.htm (accessed on 28 January 2019).

Meara, Paul, and Tess Fitzpatrick. 2000. Lex30: An improved method of assessing productive vocabulary in an L2. *System* 18: 19–30. [CrossRef]

Meara, Paul, and Brent Wolter. 2004. V_Links: Beyond vocabulary depth. *Angles on the English Speaking World* 4: 85–96.

Milton, James. 2010. The development of vocabulary breadth across the CEFR levels. In *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*. Edited by Bartning Inge, Maisa Martin and Ineke Vedder. Eurosla Monograph Series, 1; pp. 211–32. Available online: http://eurosla.org/monographs/EM01/EM01home.html (accessed on 28 January 2019).

Milton, James. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In *L2 Vocabulary Acquisition, Knowledge and Use*. Edited by Bardel Camilla, Christina Lindqvist and Batia Laufer. Amsterdam: EUROSLA-the European Second Language Association, pp. 57–78.

Milton, James, and Tess Fitzpatrick, eds. 2013. *Dimensions of Vocabulary Knowledge*. London: Macmillan International Higher Education.

Nation, Paul. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Qian, David. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review* 56: 282–308. [CrossRef]

Read, John. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined? In *Vocabulary in a Second Language: Selection, Acquisition and Testing*. Edited by Bogaards Paul and Batia Laufer. Amsterdam: John Benjamins Publishing, vol. 10, pp. 209–27.

Schoonen, Rob, Amos van Gelderen, Kees de Glopper, Jan Hulstijn, Annegien Simis, Patrick Snellings, and Marie Stevenson. 2003. First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning* 53: 165–202. [CrossRef]

Segalowitz, Norman, and Barbara F. Freed. 2004. Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition* 26: 173–99. [CrossRef]

Stæhr, Lars Stenius. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal* 36: 139–52. [CrossRef]

Stæhr, Lars Stenius. 2009. Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition* 31: 577–607. [CrossRef]

Tavakoli, Parveneh, and Peter Skehan. 2005. Strategic planning, task structure and performance testing. In *Planning and Task Performance in a Second Language*. Edited by Ellis Rod. Amsterdam: John Benjamins Publishing, pp. 239–73.

Uchihara, Takumi, and Jon Clenton. 2018. Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*, 1–17. [CrossRef]

Uchihara, Takumi, and Kazuya Saito. 2019. Exploring the relationship between productive vocabulary knowledge and second language oral ability. *The Language Learning Journal* 47: 64–75. [CrossRef]

Vieira, Gicele Vergine. 2017. Lexical access in L2 speech production: A controlled serial search task. *Ilha do Desterro* 70: 245–64. [CrossRef]

Wu, Sina. 2016. Influence of vocabulary, syntactic and metacognitive awareness of Japanese students on their Chinese reading comprehension [词汇、句法和元认知策略对日本学生汉语阅读理解的影响]. *Language Teaching and Linguistic Studies* [语言教学与研究] 178: 59–66.

Wu, Sina. 2017. Contribution of vocabulary knowledge, morphological awareness and lexical inference ability to reading comprehension: Evidence from SEM [语素意识、语法知识与汉语二语阅读理解——来自结构方程模型的证据]. *Chinese Teaching in the World* [世界汉语教学] 31: 420–32.

languages

MDPI

Article

Syntactic and Discourse Features in Chinese Heritage Grammars: A Case of Acquiring Features in the Chinese Sentence-Final Particle *ba*

Shanshan Yan

School of Chinese as a Second Language, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing 100871, China; shanshanyan@pku.edu.cn

Received: 24 February 2020; Accepted: 5 June 2020; Published: 22 June 2020

Abstract: This study investigates how syntactic and discourse features of Chinese sentence-final particles (the question particle *ba* and the suggestion particle *ba*) are reconfigured in Chinese heritage grammars. It has been argued that features of the Chinese particles *ba* are present in English but are configured differently. An acceptability judgment task, a discourse completion task, and a translation task were adopted in this study. In total, 35 Chinese heritage speakers and 18 Chinese native speakers took part in this study. The results show that none of the heritage speaker groups had any problem in configuring the discourse feature of the suggestion particle *ba* and the syntactic features of the question particle *ba*. However, none of them could successfully reconfigure the discourse feature of the question particle *ba*. It seems that the effects of dominant language transfer, reduced Chinese input, and limited processing resources play roles in the reconfiguration of discourse features in heritage grammars. As compared to previous L2 studies regarding the same phenomenon, heritage speakers with more and early Chinese input seem to have advantages over L2 learners in terms of syntactic features. L2 learners are found to be slightly better than heritage speakers in terms of reconfiguring some discourse properties.

Keywords: sentence-final particle; Chinese heritage speakers; features

1. Introduction

Heritage speakers are those simultaneous or sequential bilinguals whose weaker language corresponds to the minority language of their society and whose stronger language is the dominant language of that society (Polinsky 2018b)[1]. These speakers' acquisition of their heritage language is considered an important issue in the field of language acquisition, as it bridges L1 and L2 acquisition. Many studies involving heritage speakers concern their ultimate achievements in their heritage language (Montrul 2009), the effects of language transfer from the dominant language (Montrul 2010; Montrul and Ionin 2012), and heritage language attrition (Montrul 2002, 2008, 2011a; Pallier 2007; Polinsky 2011), as well as their differences from L1 learners and L2 learners (Montrul 2010; Montrul and Foote 2014; Polinsky and Kagan 2007; Rothman 2007). Recently, a significant body of research has explored the underlying characteristics of Spanish heritage speakers (Cuza 2012, 2016; Valenzuela et al. 2015), Russian heritage speakers (Polinsky 2011, 2016, 2018a; Sekerina and Sauermann 2015),

[1] In the literature, it is difficult to obtain a clear-cut definition of heritage language speakers. Polinsky (2018b) has reviewed several major arguments and pointed out that her definition tries to tie together different dimensions such as early bilingualism, simultaneous and sequential acquisition, and the unbalanced relationship between the two languages. By quoting Polinsky's definition, we hope to present a more general picture of the definition of heritage speakers and their complex nature. For more definitions and discussions, please see Valdés (2000), Fishman (2001), Rothman (2009), and Kupisch and Rothman (2018), among others.

and Korean heritage speakers (Chung 2013; Kim 2007; Kim et al. 2009). However, there have been few investigations of Chinese heritage speakers' acquisition of syntactic and discourse features. This study tries to fill this gap by providing evidence related to Chinese heritage speakers in terms of their acquisition of Chinese language properties.

Among the research on heritage grammars, the feature properties of linguistic phenomena (i.e., syntactic, semantic, discourse features) and the influence of the dominant language have all been extensively investigated, particularly under the theoretical framework of the Interface Hypothesis (Montrul 2012; Montrul and Ionin 2012; among others). The Interface Hypothesis (Sorace 2011; Sorace and Filiaci 2006; Sorace and Serratrice 2009; Tsimpli and Sorace 2006) argues that structures that lie at the core of a syntax system (narrow syntax) or at internal interfaces (such as syntax-semantics) are more impervious to cross-linguistic influences than features that lie at syntax interfaces with other modules (external interfaces), such as the syntax–discourse interface. Montrul and Polinsky (2011) further argue that heritage speakers are an "important testing ground" for the Interface Hypothesis, directing researchers' attention from the L2 field to heritage language acquisition. On the basis of findings from empirical studies concerning the linguistic knowledge of heritage speakers and L2 speakers, Montrul (2012) summarizes that syntax (and phonology) are the most resilient area(s) of grammar for heritage speakers, where they have the most advantages over L2 learners. However, aspects at the syntax-discourse interface (semantics and inflectional morphology) seem to be very vulnerable for heritage speakers. By investigating the performance of verb placement (the V2 phenomenon) and nominal agreement in the NPs (gender, number, and definiteness) of both Swedish heritage speakers and L2 learners of Swedish, Håkansson (1995) found that the Swedish heritage speakers followed the V2 rule whereas L2 learners were highly inaccurate in their application. In contrast, in the syntax-discourse domain Keating et al. (2011) showed that although both Spanish L2 learners and heritage speakers were significantly different from Spanish native speakers in preferring the relevant antecedent for overt and null pronouns, there was no difference between them in their choices, which suggests that the heritage speakers did not outperform the L2 learners in terms of syntax and discourse related areas. Regarding transfer effects from dominant languages at interfaces, Montrul (2010) argues that the effects may be found in both Spanish heritage speakers and L2 learners' grammars, and more so at syntax-semantic/pragmatic interfaces than in the core syntax. Montrul and Ionin (2012) further confirmed these findings, reporting transfer effects from the dominant language in the semantic field during the interpretation of definite plural articles in Spanish by L2 learners of Spanish and English-dominant Spanish heritage speakers. Lee (2016) also found that both L2 learners and heritage speakers of Korean exhibited transfer effects from their dominant language, but this effect was stronger in the L2 group. Other recent studies concerning properties at interfaces include Kaltsa et al. (2015), Laleko and Polinsky (2016), Leal et al. (2018) and others.

However, the Interface Hypothesis is not without some debates. Hopp (2011), Pires and Rothman (2011), and O'Grady (2011) propose that computational or processing complexity are better factors than interfaces. Duffield (2011), Montrul (2011b), Pérez-Leroux (2011), and White (2011) all criticize the problematic distinction between internal and external interfaces, pointing out that linguistic structures may involve multiple interfaces and learners' difficulties may reside at several of them, and there may also be conflict at the same interface from different linguistic phenomena. Sorace (2012, p. 213) proposes that researchers should "instead allow for a range of interface conditions, graded according to their computational complexity and their dependence on extra-linguistic factors", and this may lead to a better understanding of learner grammar facts.

In a further discussion of acquisition difficulties in bilingualism and heritage grammars, de Prada Pérez (2010) argues that complexity, as an important alternative explaining factor to interfaces, resides in variability, and she proposes a Vulnerability Hypothesis (de Prada Pérez 2019) that establishes a cross-linguistic permeability hierarchy along a variability continuum spanning from categorical distributions to highly variable context. In her proposal, a linguistic phenomenon is classified by its relative frequency as being a categorical distribution or a variable distribution. Her study on the use of

null and overt pronominal subjects by Spanish-Catalan bilinguals (and Spanish- and Catalan-dominant monolinguals) shows that for categorical distributions, where a specific form is used (near) exclusively, are more invulnerable, whereas variable distributions, where more than one form can be used, are vulnerable to cross-linguistic influences. She further argues that learners with higher proficiency may be affected by cross-linguistic influences in highly variable phenomena, but lower proficiency learners also exhibit this influence in less variable phenomena. However, as Hoot (2017) has pointed out, complexity may vary from structure to structure, and it is hard to generalize its power across linguistic constructions and in heritage language acquisition. As for the variability proposal, it depends on ratings of the frequency/distribution of structures, which relate to social contacts (mostly relying on monolinguals' ratings) and which may be hard to decisively define. As will be discussed below, in the literature the underlying operating mechanism in heritage grammars is influenced by various conditions, as compared to monolingual baselines.

Polinsky and Scontras (2020) discuss the deviation of heritage grammars from the relevant baseline. They propose that this deviation is mainly triggered by the input of the heritage language from which the heritage grammar is acquired (both quantity and quality), as well as the economy of online resources when operating in the heritage language. They conclude that the unique outcomes among heritage speakers (i.e., differences from both the baseline and L2 speakers) are affected by reduced input quality as well as reduced quantity in these heritage speakers' daily lives. In a summary of empirical studies concerning heritage grammars, Polinsky and Scontras (2020) argue that in terms of quantity, the greater their exposure to the heritage language over a longer period of time, the more effectively the heritage grammars are acquired, and thus these speakers will obtain a more balanced bilingual ability. Furthermore, the recency of exposure to the heritage language, both cumulative exposure and current exposure in daily life, are predictive of grammatical outcomes. In terms of quality, community size has an effect on the proficiency of heritage grammars (Gathercole and Thomas 2007; Gollan et al. 2015). When it comes to the online processing of heritage languages, due to the limited nature of processing resources as well as the additional cost of operating in a less proficient language, heritage speakers tend to apply economical, superficial, and direct strategies, which leads to their preferences for avoidance of ambiguity, resistance to irregularity, and shrinking of structures.

To summarize, it seems that both the recent complexity/variable proposals and the Interface Hypothesis indicate that heritage speakers and bilinguals have difficulty with certain properties (mostly discourse/pragmatics), and this seems ultimately to be an inherent part of bilingualism itself (Hoot 2017). In addition, the nature of the linguistic phenomenon being acquired, the influence of dominant languages and factors leading to divergence from monolinguals are also important in heritage language acquisition. The Chinese sentence-final particles (SFPs) *ba*, which will be discussed in the next section and which bear mainly discourse as well as other syntactic features, will provide good testing grounds for these theoretical debates.

2. Features of SFPs *ba*

Features, as the undividable ultimate building blocks of language (Chomsky 2005; Jackendoff 2002) have long been the focus of much linguistic research in both the theoretical and applied linguistic fields. For Jackendoff (2002), words in the lexicon are fully specified with lexical, syntactic, and even discourse pragmatic features already, and no syntactic procedures are needed to enable the specification of features. It is these features that explain linguistic variations. The analysis of the linguistic phenomena in question (i.e., the SFPs *ba*) will be based on this feature perspective.

2.1. Discourse Features of the Suggestion SFP ba

The suggestion SFP *ba* in Chinese mainly types a suggestion sentence, as shown in (1). In (1b), with the assistance of the suggestion SFP *ba*, it sounds soft to signify a suggestion. In contrast, in (1a) without the particle, the tone of the sentence is very harsh, as in giving an order. Therefore, we propose

that there is a discourse [suggestion] feature attached to the suggestion SFP *ba*, bearing softening and suggestive functions in Chinese.

(1) a. Kuài diǎnr zǒu!
hurry little go
'Hurry up!'
b. Kuài diǎnr zǒu ba!
hurry little go BA
'Hurry up, please!' (Chao 1968, p. 807)

However, there is no suggestive SFP in English. The [suggestion] feature is realized by several suggestive structures—for example, 'Let's (Let us) . . . ', 'We/You/They/He'd (We/You/They/He had) better . . . '. Sentences without such structural elements are merely general statements; compare (2a–c) and (3a–c).

(2) a. We are going to the park this afternoon.
b. We are going inside as it is going to rain outside.
c. We will finish this work before starting something new.
(3) a. Let's go to the park this afternoon.
b. We'd better go inside as it is going to rain outside.
c. Shall we finish this work before starting something new?

2.2. Syntactic and Discourse Features of the Question SFP *ba*

The question SFP *ba* in Chinese has both syntactic ([Q] and [-wh]) and discourse ([confirmation seeking]) features. In syntax, the addition of the question SFP *ba* to a declarative sentence converts the sentence into a question, as shown in (4). It is merely a declarative sentence in (4a), but in (4b) the addition of the particle at the end of the sentence means that it becomes a question. Thus, there is a syntactic [Q] feature attached to this particle. Furthermore, interrogative wh-phrases are not allowed to be present in an SFP *ba* question, i.e., a syntactic [-wh] feature is attached to this particle; see (5) for an example.

In discourse, the use of the question SFP *ba* in a sentence presents the speaker's attitude of certainty in the conversation (Hu 1987; Lu 1985; Qu 2006). In other words, the use of the question SFP *ba* provides an assumption and asks for a confirmation (Xu 2003). In (4b), the speaker asks the SFP *ba* question with some certainty and expects an answer to confirm his/her assumptions. Therefore, a discourse [confirmation seeking] feature is attached to this particle.

(4) a. Tā huì kāi chē.
3SG can drive car
'He(She) can drive a car.'
b. Tā huì kāi chē ba?
3SG can drive car BA
'He(She) can drive a car, can't he(she)/right?'

(5) *Nǐ wèn le shéi ba?
2SG ask PERF who BA
Intended: 'Whom did you ask?'

In English, tag questions "express a need for confirmation of the statement expressed in the declarative" (Huddleston and Pullum 2005, p. 164). We propose that the functions (typing a question, asking for confirmation) and restrictions (disallowing interrogative wh-phrases) of the question SFP *ba* are manifested by the tag questions in English, which is in accordance with Tang (2015a, 2015b, 2016a, 2016b). The tag question in English (e.g., *The dog is yours, is it/isn't it/right?*) mainly consists of a declarative clause (i.e., *The dog is yours*), which expresses the view of the speaker, and a tag clause (i.e., *is it/isn't it/right?*), which indicates that the speaker wishes his/her view to be confirmed. The tag part of the question can be viewed as a reduced form of the yes-no question, whose content can be recovered

from the preceding declarative clause (Huddleston and Pullum 2005). There are two categories of tag questions in English: the canonical tag question, and the invariant tag question. The former can be further divided into the non-disjunctive question (e.g., *The dog is yours, is it?*) and the disjunctive question (e.g., *The dog is yours, isn't it?*); the latter uses *right*, *yeah* and is very colloquial (e.g., *The dog is yours, right?*). In syntax, both types of tag question in English are conjunction structures (Kayne 2016; Sailor 2009), which have two individual CPs. Kayne (2016) argues that the declarative clause in the English tag question is labeled as the external conjunct, i.e., *John is a student*, and the yes-no tag as the internal conjunct, i.e., *isn't he a student*, as in (6b). He also proposes that these two conjuncts are connected by a disjunctive conjunction F. Sailor (2009) suggests that the surface derivation of the tag question is due to ellipsis of the repeated NP or VP, as in (6), where the tag question in (6a) is a result of the deletion of the NP *a student* in (6b). As for the derivation of invariant tag questions, i.e., those with 'right', 'yeah', etc., Kayne (2016) points out that there is a silent element, *ISN'T THAT*, as in (7b) in the internal conjunct. In this sense, the internal conjunct of the invariant tag questions is also a fully-fledged CP. Therefore, we propose that the [Q], [-wh] and [confirmation seeking] features, as with the Chinese question SFP *ba*, are attached to the internal conjunct of the tag question in English, which is a fully-fledged CP (the tag part), specifically the C of the CP.

(6) a. John is a student, isn't he?
b. [$_{CP}$John is a student] F [$_{CP}$isn't he ~~a student~~?]

external conjunct internal conjunct

(7) a. We are on the list, right?
b. [$_{CP}$We are on the list] F [$_{CP}$ ISN'T THAT right?] (irrelevant syntactic details omitted)

In fact, as pointed out by Tang (2016b), Chinese does have tag questions that have a parallel underlying syntactic structure as tag questions in English. There are also canonical and invariant forms of tag questions in Chinese (Li and Thompson 1981; Tang 1988, 2015a, 2015b, 2016a, 2016b; see examples 8, 9). The surface derivation of both forms of Chinese tag questions is also due to ellipsis of the repeated element (which is a CP in Chinese, as in 8, but an NP or VP in English, as in 6), as well as the existence of a silent null element (which is an empty category *e* in Chinese, as in 9, but a silent element *ISN'T THAT* in English, as in 7).

(8) a. Nǐ yào táotàidiào zhè pǐ mǎ, shì ma?
2SG will eliminate this CL horse COP MA
'You will eliminate this horse, will you?' (Tang 1988, p. 271)
b. [$_{CP}$Nǐ yào táotàidiào zhè pǐ mǎ] F [$_{CP}$ shì ~~Nǐ yào táotàidiào zhè pǐ mǎ~~ ma?]
(Tang 2016b, p. 32)

(9) a. Nǐmen shì jiǔ diǎnzhōng kāimén de, duìbuduì?
2PL COP nine o'clock open door DE yes-NEG-yes
'Yours (/your store) opens at nine o'clock, right?' (Li and Thompson 1981, p. 546)
b. [$_{CP}$ Nǐmen shì jiǔ diǎnzhōng kāimén de]$_i$ F [$_{CP}$ e_i duìbuduì?] (Tang 2016b, p. 31)

What, then, is the role of the question SFP *ba*? In line with Yan and Yuan (2020), we agree that it is an 'additional' means to realize the function of English tag questions in Chinese.

Furthermore, there is an additional fact that we cannot ignore, which is that Chinese has an SFP *ma* that has same syntactic features (both [Q] and [-wh]; see examples 10, 11) as the question SFP *ba*. Unlike the question SFP *ba*, there is an [information seeking] discourse feature in the SFP *ma*. That is, in an SFP *ma* yes-no question (e.g., 10b), genuine information is being sought from the listener; thus either a positive answer (like 'Yes, he can') or a negative one ('No, he can't') would be suitable, but no particular answer is necessarily expected. This is in contrast to the question SFP *ba* question, as in (4b).

(10) a.	Tā	huì	kāi	chē.	
	3SG	can	drive	car	
	'He(She) can drive a car.'				
b.	Tā	huì	kāi	chē	ma?
	3SG	can	drive	car	MA
	'Can he(she) drive a car?'				

(11)	*Nǐ	wèn	le	shéi	ma?
	2SG	ask	PERF	who	MA
	Intended: 'Whom did you ask?'				

3. Previous Research

On the basis of the linguistic comparison between Chinese and English in terms of the features of the SFPs *ba* in the above section, it seems that the realization of relevant features is asymmetric across the two languages. In other words, there is structural ambiguity between Chinese and English. For the suggestion SFP *ba*, there seems to exist a one-to-many mapping between Chinese and English, whereas for the question SFP *ba* there is a many-to-one mapping (also see Yan and Yuan 2020). The structural ambiguity, which has been summarized by Müller (1998), states that in language acquisition, if a phenomenon has two possible structures in one language but only one in another language, only the similar structure shared by the two languages will be kept. That is, the structure that is not shared will be abandoned in learners' grammars.

Scontras et al. (2017) showed that in a situation where English allows both surface and inverse interpretations of doubly quantified sentences (e.g., *A shark attacked every pirate*, which can be interpreted in two different ways: a. *There was a single shark that attacked multiple pirates*, or b. *For each pirate, there was a (different) shark that attacked him.*), English-dominant heritage speakers of Mandarin lack scope ambiguities in their Mandarin heritage grammar, which allows only interpretation (a) for the above English sentence. This demonstrates that heritage speakers prefer a one-to-one surface mapping from structure to interpretation. Ronai (2018) followed up on Scontras et al.'s work and found that both Hungarian-dominant heritage speakers of English and English-dominant heritage speakers of Hungarian disallow inverse interpretations of doubly quantified sentences (Hungarian is like Mandarin, disallowing inverse interpretation of scope ambiguity). This further demonstrates that heritage speakers tend to prefer a simplified structure, and they struggle with one-to-many mappings between form and meaning in their heritage grammars. Polinsky (2018b) also concluded that heritage speakers may have morphology problems, as they overextend regular tense morphology and they prefer to reduce multiple allomorphs into one. However, Anderssen et al. (2018) found a different result; in their study, the acquisition of possessive DPs by English-dominant heritage speakers of Norwegian was investigated. In Norwegian, both Possessive-N and N-Possessive are allowed; however, only Possessive-N is allowed in English. Their results show that Norwegian heritage speakers overuse N-Possessive structures and this effect is proficiency related; it seems that learners "inhibit" similar structures between the two languages, thus contradicting to structural ambiguity arguments (i.e., Müller 1998; among others).

In terms of the Chinese SFPs, Wen (2014) studied the pragmatic development of 48 learners of Chinese in producing requests through a discourse completion task. One of her research findings showed that Chinese learners behaved differently from Chinese native speakers, particularly in the use of pragmatically functional particles and phrases such as the suggestion SFP *ba* and the tag-question appealer *xíngma* ('OK') for permission. Wen also found that learners have problems with opaque and similar non-literal expressions such as *xíngma* ('OK') and *hǎobùhǎo* ('OK'). By comparing the performance of requests in the grammars of Chinese heritage learners and L2 learners, Wen (2019) found that the SFP *ba* was rarely produced by heritage and L2 learners, although it was frequently used by Chinese native speakers. Furthermore, she also found that the heritage speakers slightly under-produced tag-questions in comparison with L2 learners, and therefore she argued that the

heritage speakers adopted a more conservative strategy. This is in line with Dekeyser (2005), in that the optionality of a form and meaning connection leads to a lack of transparency, which increases the acquisition difficulties.

Using a similar experimental design (see Section 4 for details) but a different learner group, Yan and Yuan (2020) investigated features of the question and suggestion SFPs *ba* and their behaviors in the grammars of English-speaking L2 learners of Chinese. This study concluded that L2 learners have no problem with the features of the suggestion SFP *ba*, but have substantive problems with both the syntactic [Q] and the discourse [confirmation seeking] features of the question particle *ba*, and only at advanced levels could L2 learners successfully reconfigure the discourse [confirmation seeking] feature. These results suggest that the input and the L1–L2 (i.e., English-Chinese) structural difference in realizing the features of the particles significantly affected the successful acquisition of the features. In particular, the 'one-to-many' form-meaning mapping imposes great difficulties for L2 learners.

In this study, we look further at the re-configuration of the features of the SFPs *ba* in heritage grammars. With relatively more and earlier input than L2 speakers of Chinese, but also with influence from the dominant language (English), we expect to discover whether the syntactic and discourse feature properties carried by the Chinese SFPs *ba* will exhibit any differences for heritage speakers.

4. This Study

In particular, we are interested in the following questions:

(1) How are the syntactic and discourse features represented in Chinese heritage speakers' grammars?

(2) What are the possible hidden mechanisms that influence their successful acquisition?

4.1. Participants

In total, 35 Chinese heritage speakers and 18 Chinese native speakers participated in this study. Table 1 gives detailed information about the participants. All the heritage speakers were born and raised in an English-speaking country, except for six who reported immigrant status (before the age of 5). They all reported English as their dominant (native) language. Almost all the heritage speakers had frequent exposure to Chinese, including both Mandarin and dialects, since they were young (32 reported 'always', 3 'seldom')—i.e., in their family there was at least one parent frequently speaking Chinese to them. Among these heritage speakers, 71.4% (25 out of 35) had exposure to Mandarin. The remaining 28.6% had exposure to Chinese dialects, which were mainly the Cantonese and Fujian dialects[2]. In this study, the amount of naturalistic exposure and formal learning of Chinese by heritage speakers were combined and seen as early input for these speakers[3]. Data collection took place in four universities in Beijing and Shanghai in China while the heritage speakers were taking a Chinese language course.

---

[2] Three heritage speakers reported that they had exposure to the Wenzhou dialect, Taiwanese, and Hakka respectively, but at the same time they also reported exposure to Mandarin.

[3] As a reviewer has pointed out, it is better to minimize the amount of learning (especially classroom instruction) that heritage speakers receive, in order to be more confident that any effect that we are seeing is really the outcome of naturalistic heritage language acquisition rather than classroom "re-acquisition". However, it was almost impossible for us to find ideal data from HSs who have no experience of formal learning of Chinese. Among our Chinese HSs, almost all of them reported early exposure to and learning of Chinese since a very young age, either by attending formal classes or weekend classes or via informal learning at home (this was probably due to a cultural phenomenon whereby the parents of our HSs expect more connections to the Chinese language and culture). Moreover, the inclusion of the time spent learning Chinese allows us to make comparisons with L2 learners of Chinese, who have also been learning Chinese but from a later age. Therefore, the findings can to some extent inform us about whether early Chinese input has any effects on the acquisition of relevant features.

Table 1. Information about participants in each group.

Groups	Number of Participants	Average Age	Average Months of Studying Chinese[4]	Average Months in China	Mean Score in Cloze Test (SD)
HS High-int	16	24	199	13.5	25.3 (3.5)
HS Advanced	19	22	187	23.2	33.2 (2.4)
Native	18	23	N/A	N/A	38.8 (1.1)

Notes: 'HS High-int' hereafter stands for heritage speakers at high-intermediate proficiency levels, and 'HS Advanced' hereafter stands for heritage speakers at advanced proficiency levels.

All the Chinese native speakers were born and raised in northern China, and none of them had ever been to an English-speaking country. They were mainly post-graduate students majoring in Arts and Humanities at a university in Beijing. The heritage speakers were divided into two proficiency groups according to their performance in an established cloze test, which consisted of two short passages with 40 blanks (cf. Yuan 2014, 2015; Yuan and Dugarova 2012). The Chinese native speakers also participated in the cloze test, and they were treated as the baseline group. Results of a one-way ANOVA and post hoc Tukey tests indicate that all the groups were significantly different from each other (F (2, 50) = 123.537, $p < 0.001$).

4.2. Tasks and Procedures

4.2.1. Acceptability Judgment Task (AJT)

The Acceptability Judgment Task explicitly asks participants to judge sentences as acceptable or not. In particular, this task is used to test whether participants have acquired the syntactic [Q] and [-wh] features of the question SFP *ba*. Examples are shown in (12) and (13). Participants who accepted the sentence in (12a), and at the same time rejected that in (12b), counted as evidence that [Q] and [-wh] features were attached to the question *ba* in their Chinese grammars. Meanwhile, control sentences, as shown in (13), were also included to ensure that the acceptance or rejection of the experimental sentences was not for any lexical or structural reasons.

The major consideration in adopting the AJT in this study is because of its advantages in providing acceptance and especially rejections of the structures. Mackey and Gass (2005) have pointed out that the AJT explicitly provides evidence of what participants include and particularly exclude in their grammars. In this study, we are interested in syntactic features (i.e., [Q] and [-wh] features), which were tested by asking participants to judge grammatical and ungrammatical sentences. However, ungrammatical sentences never (or at least rarely) occur in real linguistic environments. AJT can thus provide us with at least some evidence of the learners' rejection of ungrammatical sentences.

There were four tokens for each of the four types. In total there were 16 test items, and they were randomized with 128 other sentences that were used either to test other aspects of L2 Chinese grammars or to serve as distractors. In the AJT, as well as in the tasks in the following sections, only basic everyday words were included in the tasks. Words that might be new or difficult for the participants were provided with their English translations. The instructions were given to the HSs in English and to the Chinese native speakers in Chinese.

[4] Here, the statistics are an overall measurement of their time spent learning Chinese. Future studies could further investigate the effects of different learning status (i.e., formal vs. informal) on the acquisition of Chinese.

(12) a. Experimental A:

Mǎlì	chángcháng	qù	chāoshì	mǎi	cài	ba?
Mary	often	go	supermarket	buy	vegetable	BA

'Mary often goes to the supermarket to buy vegetables, right?'

b. Experimental B:

Mǎlì	wèishénme	chángcháng	qù	chāoshì	mǎi	cài	ba?
Mary	why	often	go	supermarket	buy	vegetable	BA

Intended: 'Why does Mary often go to the supermarket to buy vegetables?'

(13) a. Control A:

Mǎlì	chángcháng	qù	chāoshì	mǎi	cài.
Mary	often	go	supermarket	buy	vegetable

'Mary often goes to the supermarket to buy vegetables.'

b. Control B:

Mǎlì	wèishénme	chángcháng	qù	chāoshì	mǎi	cài?
Mary	why	often	go	supermarket	buy	vegetable

'Why does Mary often go to the supermarket to buy vegetables?'

This task was designed with the E-Prime 2.0 software, and sentences were presented on a computer monitor one by one. Participants were tested individually. They were required to press a key to indicate their judgment of the degree of the acceptability of each sentences: key 'A' represented 'completely unacceptable', key 'D' 'probably unacceptable', key 'J' 'probably acceptable', and key 'L' 'completely acceptable'. A separate key 'T' was used to represent the option 'I don't know'. To make these keys and their representations clearer, small word tags were stuck on the relevant keys. Participants were also given six practice examples before they begin the main test.

4.2.2. Dialogue Completion Task (DCT)

The acquisition of the [confirmation seeking] feature of the question *ba* and the [suggestion] feature of the suggestion *ba* were tested in this task. The DCT allows us to see whether participants can distinguish different particles in a specific context which particularly favors only one particle. Many prior studies have provided evidence that the DCT is a valid instrument to investigate discourse or pragmatic features (Mackey and Gass 2005; Pinto and Raschio 2007; Rose 2009; Wen 2014, 2019). In terms of practical issues (i.e., space and time limitations on the tasks in the project), the dialogues in this task were designed to be miniature and as self-evident as possible. However, the data elicited from this task can still provide us with information about specific and authentic situations for social communication and pragmatic purposes.

Participants were presented with short dialogues which had a part missing. Their task was to choose the most appropriate item to complete the dialogue. In the instructions the participants were particularly asked to read the whole conversation and told that only the single most appropriate answer should be chosen. This was to exclude cases where participants may have found more than one item that was suitable to complete the conversation, thus ensuring that only the most appropriate item was selected. The list of items provided was *de* (an auxiliary word), *yòu* ('again'), SFP *ma*, SFP *ba*, SFP *ne*, SFP *le*, and SFP *a*. Examples are shown in (14) and (15). In (14), the question SFP *ba* was expected mainly due to the answer from Speaker B, which implies that Speaker A already knows that Speaker B can speak Chinese (as suggested by Speaker B's utterance *'nǐ zěnme zhīdào?'*). There were four tokens for each SFP, randomized with 30 other sentences, which either tested other aspects of Chinese grammars or served as distractors and fillers. This task was administered before the other two tasks to avoid any priming effects.

(14) (the question SFP *ba* is expected)[5]

Speaker A:	Nǐ	huì	shuō	Hànyǔ	_____?
	2SG	can	speak	Chinese	
	'You can speak Chinese, right?'				
Speaker B:	Wǒ	huì,	nǐ	zěnme	zhīdào?
	1SG	can	2SG	how	know
	'I can. How do you know?'				

(15) (the suggestion SFP *ba* is expected)

Speaker A:	Wàimiàn	fēng	tài	dà	le.	Guānshàng	huānghu _____.
	outside	wind	too	big	LE	close-PREP	window
	'It is too windy outside. (Please could you) close the window.'						
Speaker B:	Hǎo,	méi	wèntí.				
	good	NEG	problem				
	'OK, no problem.'						

4.2.3. Translation Task (TT)

The translation task asks participants to translate Chinese sentences into English ones. It was used to supplement the results of the other tasks; specifically, it allows us to see what potential English counterparts exist for the Chinese question SFP *ba* and suggestion SFP *ba* in heritage speakers' grammars, and further to see whether there is any effect from the dominant language. Since there are no SFPs in English, it is important to see how the features of the Chinese SFPs *ba* are manifested in English, which was the dominant language of the Chinese heritage speakers studied in this research. There was only one token for each SFP in this task. In addition, there were also 9 other sentence types, which tested other aspects of Chinese grammars[6]. This task was administered after the two tasks described above in order to avoid any priming effects. Examples are shown in (16) and (17).

(16)

Xiǎohóng	xǐhuān	lǚyóu	ba?	
Xiaohong	like	travel	BA	
'Xiaohong likes travelling, doesn't she/right?'				

(17)

Wǒmén	yìqǐ	qù	xué	Hànyǔ	ba.	
1PL	together	go	learn	Chinese	BA	
'Let's go to learn Chinese together!'						

5. Results

5.1. AJT: Syntactic Features of the Question SFP ba

Participants were first screened by their acceptance of control sentences (i.e., those in 13). Those who failed to accept at least three out of the four tokens in each control sentence type were excluded from further analysis in this task. The results showed that all participants successfully accepted those sentences. Table 2 shows the descriptive results for all groups.

---

[5] It may be argued that at a first glance at Speaker A's utterance in (14), the SFP *ma* may be a possible candidate. However, in the instruction for the task, participants were explicitly told that to complete the task they should read the whole conversation including the response from Speaker B and should only choose the single most appropriate answer. The utterance by Speaker B in (14) indicates that the question SFP *ba*, not the SFP *ma*, is required in Speaker A's utterance. However, as a reviewer has pointed out, we cannot totally exclude the possibility that participants' errors were from their sloppy reading, such as neglecting the second sentence. Future research adopting other tasks may reveal more evidence on this.

[6] The Chinese native speakers did not participate in this task as it was designed to see how learners of Chinese treat the SFPs in their native English.

Table 2. Mean scores in the AJT for sentences testing the [Q] and [-wh] features of the question SFP *ba*.

Group	N (after Screening)	Control A (SD) (e.g., (13a))	Experimental A (SD) (e.g., (12a))	* Experimental B (SD) (wh-Phrase + Question *ba*, e.g., (12b))	Control B (SD) (wh-Phrase, e.g., (13b))
HS High-int	16	3.91 (0.2)	2.69 (1.1)	1.72 (0.8)	3.92 (0.2)
HS Advanced	19	4.00 (0.0)	3.55 (0.7)	1.22 (0.4)	3.93 (0.2)
Native	18	4.00 (0.0)	3.89 (0.2)	1.32 (0.5)	3.96 (0.1)

Notes: * is standing for ungrammatical sentences.

Recall that participants were making judgments on an acceptability scale. In the data analysis, participants' judgments were transformed into numerical values, where 1 stands for 'completely unacceptable', 2 for 'probably unacceptable', 3 for 'probably acceptable', and 4 for 'completely acceptable'. Therefore, participants' judgments ranged from 1 to 4. For the [Q] feature (i.e., experimental A sentences) of the question SFP *ba*, both the HS advanced group and the Chinese native group were able to accept these sentences (mean scores > 3), but the HS high-intermediate group showed indeterminacy in accepting them. However, all groups rejected the ungrammatical sentences (i.e., those where a wh-phrase occurs in a question SFP *ba* sentence; mean scores approaching 1).

Linear mixed-effects models were performed to find whether there was any statistical difference among the variables (fixed factors and random factors). In this study, R (R Core Team 2014) and *lme4* (Bates et al. 2013) were applied to perform a linear mixed-effects analysis of the relationships between different fixed factors. The fixed factors included Condition (i.e., grammatical (G) vs. ungrammatical (UG) sentences), Group (Chinese native, HS advanced, HS high-int), Age, and Length (i.e., number of words included in each sentence). As random effects, we have intercepts for subjects and items, as well as by-subject and by-item random slopes included.

The results from the mixed-effects models show that both the HS advanced group and the HS high-intermediate group (all $t < |2|$, $p > 0.05$) were found not to be significantly different from the Chinese native group in accepting the grammatical experimental A sentences (i.e., testing the [Q] feature). However, the high-intermediate group was found to be significantly different from the Chinese native group in rejecting the ungrammatical experimental B sentences ($t > |2|$, $p < 0.05$; see Table 3). Therefore, these results suggest that all HS groups are able to reconfigure the [Q] feature of the question SFP *ba*, as well as the [-wh] feature, except the high-intermediate group.

Table 3. Summary of the fixed effects and random effects in the mixed-effects model (for the [-wh] feature).

Parameters	Fixed Effects			Random Effects			
				Subject Analysis		Item Analysis	
	Estimates	SE	t	Var	SD	Var	SD
(Intercept)	1.33	0.21	6.25 ***	0.29	0.54	0.03	0.16
Condition G	2.55	0.27	9.29 ***	0.31	0.55	-	-
Cage	−0.02	0.02	−1.45	-	-	-	-
HS Advanced	−0.13	0.19	−0.70	-	-	-	-
HS High-int	0.40	0.20	2.04 ***	-	-	-	-
Condition G: Cage	0.03	0.02	1.66	-	-	-	-
Condition G: HS Advanced	0.19	0.21	0.91	-	-	-	-
Condition G: HS High-int	−0.37	0.21	−1.74	-	-	-	-

Notes: reference level for condition=UG (ungrammatical), for group=Chinese native; Model formula: ratings ~ condition * cage * clength[7] + factor (group) + condition: factor (group) + (1 + condition|subject) + (1|item); here values of $t > |2|$ can be informally considered significant (Gelman and Hill 2007); *** $p < 0.001$.

5.2. DCT: Discourse Features of Suggestion SFP ba and Question SFP ba

In the analysis of data in this task, all participants' selections of items were converted into numerical values. Specifically, if they selected the expected SFP *ba* in relevant dialogues, their selections received a score of 1; otherwise, they received a score of 0. The 'I don't' know' option was treated as a missing value.

5.2.1. [Suggestion] Feature of Suggestion SFP *ba*

As shown in Table 4, the mean scores of the participants in all groups were close to the ceiling score of 4. Generalized linear mixed-effects models were conducted to see whether there was any statistical difference between groups. In the models, the fixed factors included Group (Chinese native, HS advanced, HS high-int), Age, and Length (i.e., the number of words included in sentences). As random effects, we included intercepts for subjects and items, as well as by-subject and by-item random slopes. The results show that there were no statistically significant differences between any pairs of groups ($z < |2|$, $p > 0.05$). This indicates that all the heritage groups were able to select the expected suggestion SFP *ba* in required dialogues, and they had no problem with the [suggestion] feature of the suggestion SFP *ba*.

Table 4. Mean scores of all groups in the DCT concerning the [suggestion] feature of the suggestion SFP *ba*.

Groups	N	Use of *ba* in Suggestions (Max. = 4, SD)
HS High-Int	16	3.94 (0.3)
HS Advanced	19	3.95 (0.2)
Native	18	4.00 (0.0)

5.2.2. [Confirmation Seeking] Feature of Question SFP *ba*

The participants were first screened by their acceptance (3 out of 4 tokens) of the experimental A sentences in (12a), which indicates their knowledge of the question SFP *ba* as used in questions (i.e., a [Q] feature of this SFP). The rationale for this is that being a question (having a [Q] feature) for a question SFP *ba* is the prerequisite for investigation of the [confirmation seeking] feature of the SFP. None of the participants had a problem with the experimental A sentences in (12a), and no one was excluded. Table 5 shows the results.

Table 5. Mean scores of all groups in the DCT concerning the [confirmation seeking] feature of the question SFP *ba*.

Groups	N (after Screening)	Use of *ba* in Confirmation Seeking (Max. = 4, SD)
HS High-Int	16	0.94 (1.3)
HS Advanced	19	1.95 (1.3)
Native	18	3.78 (0.4)

These results clearly show that Chinese native speakers selected the expected question SFP *ba* in required dialogues, as their mean scores were at 3.78 (close to the ceiling 4). However, heritage speakers at advanced levels were indeterminate (mean score at 1.95) in selecting the expected SFP *ba* in required dialogues, and those at high-intermediate levels tended not to select the question SFP *ba* (mean score was only 0.94).

[7] Here 'cage' stands for age (centered), 'c' is abbreviated for centered, 'clength' for length (centered), 'ratings' for the acceptability scores. In the models, we included the length (centered) as a factor. However, due to the collinearity issue, its effect is converged with condition and thus cannot be obtained in the models; it is not reported here.

A generalized linear mixed-effects analysis was applied to examine the statistical relationships between different fixed factors. The results are summarized in Table 6 and show that both the Chinese heritage high-intermediate and advanced speakers were significantly different from the Chinese native group (both $z > |2|$, $p < 0.001$). Therefore, it seems that all heritage speaker groups have difficulty in reconfiguring the [confirmation seeking] feature of the question SFP *ba*.

Table 6. Summary of the fixed effects and random effects in the generalized linear mixed-effects model (for the question SFP *ba*).

Parameters	Fixed Effects			Random Effects			
				Subject Analysis		Item Analysis	
	Estimates	SE	z	Var	SD	Var	SD
(Intercept)	119.06	19.84	6.00 ***	11,208	105.87	6711	81.92
HS Advanced	−64.11	13.87	−4.62 ***	-	-	-	-
HS High-int	−115.64	17.57	−6.58 ***	-	-	-	-

Notes: reference level for group=Chinese native; Model formula: cbind[8] (ratings, 1 − ratings) ~ factor (group) + (1|subject) + (1|item); here values $z > |2|$ can be informally considered significant; *** $p < 0.001$.

Since heritage speakers in two proficiency groups did not select the expected question SFP *ba* in the required dialogues, we were curious to see what items that they preferred rather than the expected question SFP *ba*. An error analysis was conducted, to give us more information about the behavior and nature of the question SFP *ba* in the heritage speakers' Chinese grammars. Specifically, we analyzed the cases where the participants consistently (3 out of 4 tokens) selected items other than the expected question SFP *ba*. The results are presented in Figure 1.

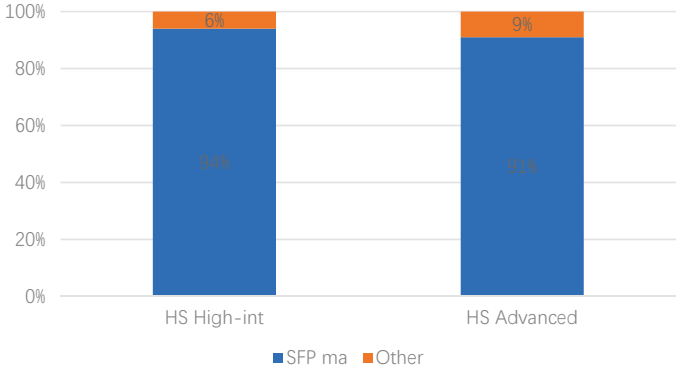


Figure 1. Percentage of *ma* and other errors consistently selected by heritage speakers at high-intermediate and advanced levels in confirmation seeking contexts in the DCT.

As shown in Figure 1, more than 90% of the heritage speakers at both high-intermediate and advanced levels selected the SFP *ma* in contexts favoring the question SFP *ba*. This suggests that all heritage speakers have difficulty making a clear distinction between the [confirmation seeking] feature attached to the question SFP *ba* and the [information seeking] feature of the SFP *ma*.

[8] Here 'cbind' represents the scores of participants.

5.3. TT: Feature Mapping between Chinese SFP ba and English Structures

Participants' translations were counted and converted into percentages, except for the Chinese native speakers as they did not take part in this task. The results are presented in Figures 2 and 3.

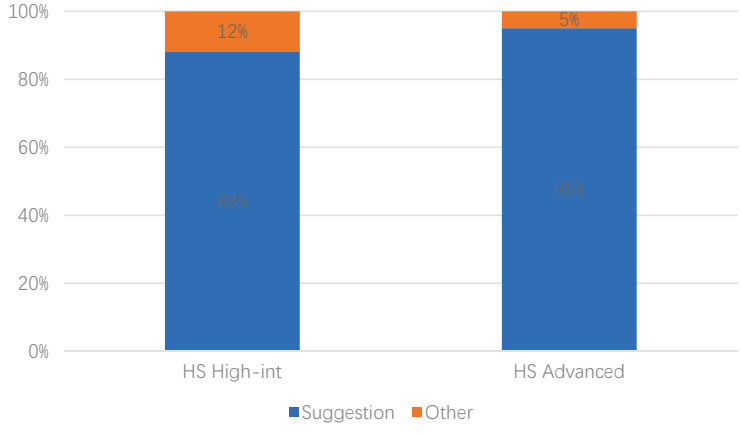


Figure 2. Percentages of heritage speakers who translate the suggestion SFP *ba* sentence into an English suggestion and heritage speakers who translate it into other types of sentences in English.

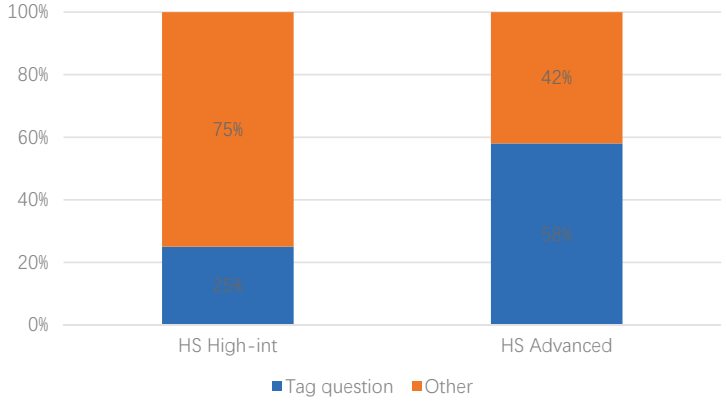


Figure 3. Percentages of heritage speakers who translate the question SFP *ba* sentence into an English tag question and heritage speakers who translate it into other types of sentences in English.

For the suggestion SFP *ba*, as shown in Figure 2, more than 87% of the heritage speakers translated the sentences containing the suggestion SFP *ba* into suggestion sentences in English (e.g., *Let's go and study Chinese together*), which indicates that in the heritage speakers' Chinese grammars, suggestion SFP *ba* sentences are treated as suggestions in their dominant language, i.e., English. This further suggests that the heritage speakers have established the mapping between Chinese suggestion SFP *ba* sentences and English suggestion sentences.

In terms of the question SFP *ba*, before analysis of the data, participants were first screened by their acceptance of experimental sentence A in (12a), similar to the rationale in the DCT (see Section 5.2.2). Since all participants met the screening criterion, no participant was excluded. Figure 3 presents the percentages of the heritage speakers' translation of the question SFP *ba* sentences into English structures. It shows that only 25% of the high-intermediate heritage speakers translated the question

SFP *ba* sentences into tag questions in English (e.g., *Xiǎohóng likes travelling, doesn't she/right/yeah?*). There was an increase at advanced levels (58%). However, more than 41% of the heritage speakers at both high-intermediate and advanced proficiency levels translated the question SFP *ba* sentences into other structures in English. A closer examination shows that all these translations were general yes-no questions in English (i.e., *Does Xiǎohóng like travelling?*). This indicates that heritage speakers have problems with interpreting Chinese question SFP *ba* sentences as tag questions in English (instead, they interpret them as yes-no questions, similar to the English equivalent of the SFP *ma*), even at advanced levels. This converges with the findings in DCT.

6. Discussion

6.1. Comparing Heritage Speakers and L2 Learners in the Acquisition of the Same Features: Addressing Question (1)

Representations of the features of the SFPs *ba* in Chinese heritage speakers' grammars are different from those in English-speaking L2 learners' Chinese grammars. Recall that Yan and Yuan (2020) investigated the same features of the SFPs *ba* in L2 grammars using a similar research design, finding that L2 learners have no problem with the discourse [suggestion] feature of the suggestion SFP *ba* and the syntactic [-wh] feature of the question SFP *ba*, but they do have difficulties with the syntactic [Q] feature and the discourse [confirmation seeking] feature of the question SFP *ba*. However, the findings in this study show that in the heritage speakers' Chinese grammars, the [Q] feature of the question SFP *ba* was successfully reconfigured, which suggests an advantage of heritage speakers over L2 learners in terms of acquiring this syntactic feature. Heritage speakers at high-intermediate levels are indeterminate in reconfiguring the [-wh] feature of the question SFP *ba*, which seems like a lower level of competence in comparison with L2 learners in terms of acquiring this syntactic feature. However, at advanced levels both heritage and L2 learners have no problem with this feature. We argue that this optionality in acquiring the [-wh] feature is probably a temporary confusion with wh-phrases in Chinese[9], which does not lead to permanent representational problems in heritage grammars. This confirms the argument of Montrul (2012) that heritage speakers have a better mastery of the syntactic features in comparison with L2 learners.

With respect to the [confirmation seeking] feature, though both heritage speakers and L2 learners have substantial problems in successfully reconfiguring it (as shown in the DCT and the TT), which is in line with Keating et al. (2011) and Montrul (2012), the English-speaking L2 learners outperformed the heritage speakers, since L2 advanced learners were found to be able to reconfigure this discourse feature successfully by Yan and Yuan (2020). As for the discourse [suggestion] feature of the suggestion SFP *ba*, similar to its behavior in the grammars of L2 learners, it is also successfully represented in the grammars of Chinese heritage speakers. Therefore, it seems that not all discourse features pose difficulties to heritage speakers and L2 learners.

Furthermore, the heritage speakers do not always have advantages over L2 learners regarding the acquisition of properties at the (syntax-)discourse level, i.e., in terms of the [confirmation seeking] discourse feature (Wen (2019) reports a similar finding). Montrul (2012) has summarized that there is a vulnerability in the syntax-discourse domain for heritage language grammars, but the findings of this study show that this may not be across the board as the heritage speakers seem to have no problem at all with the [suggestion] discourse feature. In terms of the Vulnerability Hypothesis (de Prada Pérez 2019), it seems that the discourse [confirmation seeking] feature, which may have a more variable

[9] In Chinese, some non-interrogative wh-phrases are able to appear in SFP *ba* questions, for example,
Tā zài chī shenme ba?
3SG PROG eat something BA
'He(She) is eating something, isn't /is he(she)?/right?'

distribution, i.e., more than one form can be used in the target Chinese, causes the most mapping and reconfiguration difficulties for heritage speakers. However, the discourse [suggestion] feature and the syntactic [Q] and [-wh] features, which may be categorical, seem to impose little difficulty for heritage speakers. Nevertheless, since this study did not intentionally test the Vulnerability Hypothesis since the frequency/distribution of the features of the SFPs *ba* are not calculated, more research is called for to further test this speculation. If this is on the right track, however, it seems that the Vulnerability Hypothesis may have more explanatory power than the Interface Hypothesis in heritage and bilingual acquisition.

6.2. Possible Factors Affecting the Acquisition of Features of the Question SFP ba: Addressing Question (2)

The role of dominant language transfer: as shown by Lee (2016), Montrul (2010), and Montrul and Ionin (2012), dominant language transfer has been found at both semantic and syntax-semantic/pragmatic interfaces, affecting the restructuring of heritage grammars. Difficulties in reconfiguring the [confirmation seeking] feature of the question SFP *ba* in Chinese heritage grammars may be also due to the same reason. Recall that this feature has different realizations in Chinese and English (see Section 2.2 for details), whereby it is attached to the question SFP *ba* in one single CP, but it is on the C of the internal CP in the English tag questions. Therefore, it is probable that this different syntactic structural realization prevents a smooth and successful reconfiguration of the [confirmation seeking] feature of the question SFP *ba* in heritage grammars. In other words, the dominant language, English, which has a different realization of the [confirmation seeking] feature, has led to difficulty in acquiring this feature among Chinese heritage speakers.

Even worse, as demonstrated in Section 2.2, there are canonical tag questions in Chinese (i.e., using *shì ma*, 'will you'; *duìbuduì*, 'right' as tags), which are syntactically the same as those used in English, and these might further 'enhance' heritage speakers' confidence that the question SFP *ba* sentences are different from the tag questions in English. This is in line with the findings of Yan and Yuan (2020), who have also found that English-speaking L2 learners of Chinese have substantive difficulties in acquiring the [confirmation seeking] feature of the question SFP *ba*. However, Yan and Yuan (2020) further shows that even though it is difficult, advanced English-speaking L2 learners do finally reconfigure this feature of the question SFP *ba*. Why, then, do heritage speakers, especially those with more and earlier Chinese input, have problems even at advanced levels? This may be explained as follows.

Input as a possible factor: Polinsky and Scontras (2020) have concluded that the unique representations of heritage speakers (i.e., their deviations from either/both heritage baseline and dominant language baseline) are probably due to the input that they have gained. In particular, the quantity (i.e., their exposure to and recency of the heritage language) and the quality (the community size) of their input all contributes to their final achievements in heritage language acquisition. In this study, though the majority of heritage speakers have early and frequent exposure to Chinese, this was up to the age of 7 (we only obtained information about participants' language backgrounds until that age in this research). That is, in the absence of evidence about their Chinese exposure after this age, we cannot exclude the possibility that it might be the lower exposure and lack of recency of Chinese (in contrast with their dominant language English) that has resulted in their problems with reconfiguring the [confirmation seeking] feature of the question SFP *ba*.

Effects of limited processing resources: it has been argued that limitations in processing resources, as well as the additional cost of handling a less proficient language, may be factors affecting the acquisition of heritage languages (Polinsky and Scontras 2020; Ronai 2018; Scontras et al. 2017; among others). Under such processing pressures, heritage speakers tend to adopt a simplified structure and a one-to-one surface mapping from structure to interpretation. Dekeyser (2005), Wen (2014), Wen (2019) and Yan and Yuan (2020) all show that the optionality of a form and meaning connection leads to a lack of transparency, and thus increases acquisition difficulties. Therefore, as demonstrated in this study, because of the similarity in structure, Chinese heritage speakers prefer the mapping of Chinese SFP *ba*

questions onto English yes-no questions (i.e., similar to the SFP *ma* question being the equivalent of the English yes-no question). They adopt this preference instead of treating them as tag questions in English, which requires the recognition of an additional mapping besides the canonical ones.

However, this study to some extent contradicts the findings of Anderssen et al. (2018), who found that heritage speakers inhibited similar structures between the target and their dominant languages. We propose that this may due to the fact that the form-meaning mappings in this study are more complex as compared to Anderssen et al. (2018), requiring recognition of the sequential order of N and possessives. In addition, heritage speakers' preferences for one-to-one form and meaning mappings may also explain the effects of successful reconfiguration in terms of different discourse features of the question SFP *ba* and the suggestion SFP *ba* in heritage grammars. Recall that the [suggestion] discourse feature of the suggestion SFP *ba*, which requires a one-to-many mapping between Chinese and English, was readily acquired by heritage speakers in this study, but the [confirmation seeking] feature of the question SFP *ba*, which requires a many-to-one mapping between Chinese and English, imposes substantive difficulties. We propose that this is probably a manifestation of the fact that the one-to-many form and meaning connection between the target heritage language (i.e., Chinese) and the dominant language (i.e., English) is easier than many-to-one form and meaning connections in heritage language acquisition. This lends further support to the findings in Yan and Yuan (2020) in terms of the form-meaning mappings in L2 Chinese.

6.3. Pedagogical Implications

The findings of this study suggest that the successful acquisition of features of the SFPs *ba* (especially discourse features of the question SFP *ba*) by heritage speakers requires disentangling similar particles and structures, distinguishing different feature realizations, and obtaining enough input. Kupisch and Rothman (2018) have argued that formal training between later childhood and early adulthood can facilitate the acquisition of heritage languages. Therefore, pedagogical measures may have influential effects on the successful acquisition of the features of the SFPs *ba*. However, current textbooks and formal teaching do not explicitly highlight the relationship between question SFP *ba* sentences and English tag questions. Awareness of the differences between the question SFP *ba* and its English counterparts, as well as its relationship with Chinese canonical tag questions, will enhance heritage speakers' ability to notice relevant evidence from the target input. Furthermore, differences between similar features of similar particles, i.e., the yes-no question particle *ma* and the question SFP *ba*, should also be highlighted and explicitly compared in formal teaching. In addition, since both suggestion and question SFPs *ba* are discourse particles, robust input, especially from spoken Chinese and from a variety of native speakers, should also be incorporated in formal teaching.

6.4. Limitations and Future Directions

Although this study has provided insightful findings concerning the acquisition of syntactic and discourse features by Chinese heritage speakers, there are a number of limitations. First, several methods have been adopted to investigate the relevant features attached to the suggestion and question SFPs *ba*, but it is necessary to add more methods from different dimensions to gain more evidence and further triangulate the data. Though AJT in this study provided us with understanding of relevant syntactic features, data from online processing tasks (such as reaction time studies or eye-tracking studies) would provide us with more knowledge about the implicit representations of syntactic features. Moreover, though the DCT is a valid method to obtain insights about the nature of discourse features, production tasks involving natural contexts and discourses could provide more in-depth data. Furthermore, besides the TT task, qualitative methods such as interviews should be incorporated in future studies to better explore the underlying mechanisms behind the performance of heritage speakers.

Second, although we tried to control for the language backgrounds of the heritage speakers in this research, it would be better to collect data from one unified Chinese language group (either Mandarin

or a single dialect). This would give more confidence to conclude that the findings are influenced by the factors under investigation, rather than being artifacts of different Chinese dialects. Third, as Kupisch and Rothman (2018) have highlighted, knowledge of heritage speakers in terms of their heritage language should be considered native, and thus future investigation of heritage speakers could compare their grammars with matched bilingual native groups to provide more insights.

7. Conclusions

This study investigated the representations of features of suggestion and question SFPs *ba* in the grammars of English-dominant Chinese heritage speakers. The results show that the features of the suggestion SFP *ba* and the syntactic features of the question SFP *ba* are properly represented, but the discourse feature of the question SFP *ba* poses a difficulty for Chinese heritage speakers. We argue that not all syntactic and discourse features are problematic in heritage language acquisition, and a transfer effect from the dominant language, the effect of input, as well as limitations in processing resources may all have contributed to the deviation in the heritage speakers' performance. Future studies are invited to particularly examine the quantity and quality of input that heritage speakers obtain in acquiring a heritage language, as well as their online instant processing of relevant feature properties.

Funding: This research received no external funding.

Acknowledgments: We are grateful to the guest editor Xiaohong Wen and three anonymous reviewers, for their valuable suggestions and comments on the manuscript. We would also like to thank all participants for their contributions to this research, which enables the study possible and interesting.

Conflicts of Interest: The author declares no conflict of interest.

References

Anderssen, Merete, Björn Lundquist, and Marit Westergaard. 2018. Crosslinguistic similarities and differences in bilingual acquisition and attrition: Possessives and double definiteness in Norwegian heritage language. *Bilingualism: Language and Cognition* 21: 748–64. [CrossRef]

Bates, Douglas M., Martin Mächler, Ben Bolker, and Steven Walker. 2013. lme4: Linear Mixed-Effects Models Using Eigen and S4. R Package Version 1.1-23, Published on 7 April 2020. Available online: https://cran.r-project.org/web/packages/lme4/index.html (accessed on 29 May 2020).

Chao, Yuen Ren. 1968. *A Grammar of Spoken Chinese*. Los Angeles: California University Press.

Chomsky, Noam. 2005. *The Minimalist Program*. Cambridge: MIT Press.

Chung, Eun Seon. 2013. Exploring the Degree of Native-Likeness in Bilingual Acquisition: Second and Heritage Language Acquisition of Korean Case-Ellipsis. Unpublished Doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

Cuza, Alejandro. 2012. Crosslinguistic influence at the syntax proper: Interrogative subject-verb inversion in heritage Spanish. *International Journal of Bilingualism* 17: 71–96. [CrossRef]

Cuza, Alejandro. 2016. The status of interrogative subject-verb inversion in Spanish-English bilingual children. *Lingua* 180: 124–38. [CrossRef]

de Prada Pérez, Ana. 2010. Subject position in Spanish in contact with Catalan: Language similarityvs. interface vulnerability. In *Proceedings of the 2009 Mind/Context Divide Workshop*. Edited by Michael Iverson, Ivan Ivanov, Tiffany Judy, Jason Rothman, Roumyana Slabakova and Marta Tryzna. Somerville: Cascadilla Proceedings Project, pp. 104–15.

de Prada Pérez, Ana. 2019. Theoretical implications of research on bilingual subject production: The Vulnerability Hypothesis. *International Journal of Bilingualism* 23: 670–94. [CrossRef]

Dekeyser, Robert M. 2005. What makes learning second language grammar difficult? A review of issues. *Language Learning* 55: 1–25. [CrossRef]

Duffield, Nigel. 2011. Loose ends? Commentary on Sorace. *Linguistic Approaches to Bilingualism* 1: 35–38. [CrossRef]

Fishman, Joshua A. 2001. 300-plus years of heritage language education in the United States. In *Heritage Languages in America: Preserving a National Resource*. Edited by Joy K. Peyton, Donald A. Ranard and Scott McGinnis. Washington, DC: Delta Systems and Center for Applied Linguistics, pp. 81–98.

Gathercole, Virginia C. Mueller, and Enlli Môn Thomas. 2007. Factors contributing to language transmission in bilingual families: The core study—Adult interviews. In *Language Transmission in Bilingual Families in Wales*. Edited by Virginia C. Mueller Gathercole. Cardiff: Welsh Language Board, pp. 59–181.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gollan, Tamar H., Jennie Starr, and Victor S. Ferreira. 2015. More than use it or lose it: The number-of-speakers effect on heritage language proficiency. *Psychonomic Bulletin and Review* 22: 147–55. [CrossRef]

Håkansson, Gisela. 1995. Syntax and morphology in language attrition: A study of five bilingual expatriate Swedes. *International Journal of Applied Linguistics* 5: 153–71. [CrossRef]

Hoot, Bradley. 2017. Narrow presentational focus in heritage Spanish and the syntax-discourse interface. *Linguistic Approaches to Bilingualism* 7: 63–95. [CrossRef]

Hopp, Holger. 2011. Extended patterns and computational complexity. *Linguistic Approaches to Bilingualism* 1: 43–47. [CrossRef]

Hu, Mingyang, ed. 1987. Modal particles and interjections in Peking Mandarin (Beijinghua de yuqi zhuci he tanci). In *Original Exploration on Peking Mandarin (Beijinghua chutan)*. Beijing: The Commercial Press, pp. 74–107.

Huddleston, Rodney, and Geoffrey K. Pullum. 2005. *A Student's Introduction to English Grammar*. Cambridge: Cambridge University Press.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Kaltsa, Maria, Ianthi M. Tsimpli, and Jason Rothman. 2015. Exploring the source of differences and similarities in L1 attrition and heritage speaker competence: Evidence from pronominal resolution. *Lingua* 164: 266–88. [CrossRef]

Kayne, Richard S. 2016. The silence of heads. *Studies in Chinese Linguistics* 37: 1–37. [CrossRef]

Keating, Gregory D., Bill VanPatten, and Jill Jegerski. 2011. Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Studies in Second Language Acquisition* 33: 193–222. [CrossRef]

Kim, Ji-Hye. 2007. Binding Interpretations in Adult Bilingualism: A Study of Language Transfer in L2 Learners and Heritage Speakers of Korean. Unpublished Doctoral dissertation, University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Kim, Ji-Hye, Silvina Montrul, and James Yoon. 2009. Binding interpretations of anaphors by Korean heritage speakers. *Language Acquisition* 16: 3–35. [CrossRef]

Kupisch, Tanja, and Jason Rothman. 2018. Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *International Journal of Bilingualism* 22: 564–82. [CrossRef]

Laleko, Oksana, and Maria Polinsky. 2016. Between syntax and discourse: Topic and case marking in heritage speakers and L2 learners of Japanese and Korean. *Linguistic Approaches to Bilingualism* 6: 396–439. [CrossRef]

Leal, Tania, Emilie Destruel, and Bradley Hoot. 2018. The realization of information focus in monolingual and bilingual native Spanish. *Linguistic Approaches to Bilingualism* 8: 217–51. [CrossRef]

Lee, Teresa. 2016. Dominant language transfer in the comprehension of L2 learners and heritage speakers. *International Journal of Applied Linguistics* 26: 190–210. [CrossRef]

Li, Charles N., and Sandra A. Thompson. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Lu, Jianming. 1985. About the question modal particles in modern Mandarin Chinese (Guanyu xiandai Hanyu li de yiwen yuqici). *Research and Exploration in Chinese Linguistics (Yufa Yanjiu he Tansuo)* 3: 233–46.

Mackey, Alison, and Susan M. Gass. 2005. *Second Language Research Methodology and Design*. London: Lawrence Erlbaum Associates Publishers.

Montrul, Silvina. 2002. Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and Cognition* 5: 39–68. [CrossRef]

Montrul, Silvina. 2008. *Incomplete Acquisition in Bilingualism: Re-Examining the Age Factor*. Amsterdam: John Benjamins.

Montrul, Silvina. 2009. Knowledge of tense-aspect and mood in Spanish heritage speakers. *International Journal of Bilingualism* 13: 239–69. [CrossRef]
Montrul, Silvina. 2010. Dominant language transfer in adult second language learners and heritage speakers. *Second Language Research* 26: 293–327. [CrossRef]
Montrul, Silvina. 2011a. First language retention and attrition in an adult Guatemalan adoptee. *Language Interact and Acquisition* 2: 276–311.
Montrul, Silvina. 2011b. Multiple interfaces and incomplete acquisition. *Lingua* 121: 591–604. [CrossRef]
Montrul, Silvina. 2012. Is the Heritage Language like a Second Language? *Eurosla Yearbook* 12: 1–29. [CrossRef]
Montrul, Silvina, and Rebecca Foote. 2014. Age of acquisition interactions in bilingual lexical access: A study of the weaker language of L2 learners and heritage speakers. *International Journal of Bilingualism* 18: 274–303. [CrossRef]
Montrul, Silvina, and Tania Ionin. 2012. Dominant language transfer in Spanish heritage speakers and L2 learners in the interpretation of definite articles. *The Modern Language Journal* 96: 70–94. [CrossRef]
Montrul, Silvina, and Maria Polinsky. 2011. Why not heritage speakers? *Linguistic Approaches to Bilingualism* 1: 58–62. [CrossRef]
Müller, Natascha. 1998. Transfer in bilingual first language acquisition. *Bilingualism: Language and Cognition* 1: 151–71. [CrossRef]
O'Grady, William. 2011. Interfaces and Processing. *Linguistic Approaches to Bilingualism* 1: 63–66. [CrossRef]
Pallier, Christophe. 2007. Critical periods in language acquisition and language attrition. In *Language Attrition Theoretical Perspectives*. Edited by Barbara Köpke, Monika S. Schmid, Merel Keijzer and Susan Dostert. Amsterdam: John Benjamins, pp. 99–120.
Pérez-Leroux, Ana. T. 2011. What I don't understand about interfaces. *Linguistic Approaches to Bilingualism* 1: 71–73. [CrossRef]
Pinto, Derrin, and Richard Raschio. 2007. A comparative study of requests in heritage speaker Spanish, L1 Spanish, and L1 English. *International Journal of Bilingualism* 11: 135–55. [CrossRef]
Pires, Acrisio, and Jason Rothman. 2011. An integrated perspective on comparative bilingual differences: Beyond the Interface problem? *Linguistic Approaches to Bilingualism* 1: 74–78. [CrossRef]
Polinsky, Maria. 2011. Reanalysis in adult heritage language: A case for attrition. *Studies in Second Language Acquisition* 3: 305–28. [CrossRef]
Polinsky, Maria. 2016. Structure vs. use in heritage language. *Linguistics Vanguard* 2: 20150036. [CrossRef]
Polinsky, Maria. 2018a. Bilingual children and adult heritage speakers: The range of comparison. *International Journal of Bilingualism* 22: 547–63. [CrossRef]
Polinsky, Maria. 2018b. *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.
Polinsky, Maria, and Olga Kagan. 2007. Heritage languages: In the "wild" and in the classroom. *Language and Linguistic Compass* 1: 368–95. [CrossRef]
Polinsky, Maria, and Gregory Scontras. 2020. Understanding heritage languages. *Bilingualism: Language and Cognition* 23: 4–20. [CrossRef]
Qu, Chengxi. 2006. *Mandarin Discourse Grammar (Hanyu Pianzhang Yufa)*. Beijing: Beijing Language and Culture University Press.
R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, Version 4.0.1, Published on 6 June 2020. Available online: http://www.R-project.org/ (accessed on 12 June 2020).
Ronai, Eszter. 2018. Quantifier scope in heritage bilinguals: A comparative experimental study. In *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*. Edited by S. Hucklebridge and M. Nelson. Amherst: GLSA, University of Massachusetts, vol. 3, pp. 29–38.
Rose, Kenneth R. 2009. Interlanguage pragmatic development in Hong Kong, phase 2. *Pragmatics* 41: 2345–64. [CrossRef]
Rothman, Jason. 2007. Heritage speaker competence differences, language change and input type: Inflected infinitives in heritage Brazilian Portuguese. *International Journal of Bilingualism* 11: 359–89. [CrossRef]
Rothman, Jason. 2009. Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *International Journal of Bilingualism* 13: 155–63. [CrossRef]
Sailor, Craig. 2009. Tagged for deletion: A typological approach to VP ellipsis in tag questions. Unpublished Master's thesis, University of California, Los Angeles, CA, USA.

Scontras, Gregory, Maria Polinsky, C.-Y. Edwin Tsai, and Kenneth Mai. 2017. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics* 2: 36. [CrossRef]

Sekerina, Irina A., and Antje Sauermann. 2015. Visual attention and quantifier spreading in heritage Russian bilinguals. *Second Language Research* 31: 75–104. [CrossRef]

Sorace, Antonella. 2011. Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism* 1: 1–33. [CrossRef]

Sorace, Antonella. 2012. Pinning down the concept of interface in bilingual development: A reply to peer commentaries. *Linguistic Approaches to Bilingualism* 2: 209–17. [CrossRef]

Sorace, Antonella, and Francesca Filiaci. 2006. Anaphora resolution in near-native speakers of Italian. *Second Language Research* 22: 339–68. [CrossRef]

Sorace, Antonella, and Ludovica Serratrice. 2009. Internal and external interfaces in bilingual language development: Beyond structural overlap. *International Journal of Bilingualism* 13: 195–210. [CrossRef]

Tang, Ting-chi. 1988. *Studies on Chinese Morphology and Syntax*. Taipei: Taiwan Student Book Co., Ltd.

Tang, Sze-Wing. 2015a. Some syntactic and phonological properties of sentence-final particles in Chinese. Paper presented at *SytaxLab*, Cambridge, UK, April 28.

Tang, Sze-Wing. 2015b. A generalized syntactic schema for utterance particles in Chinese. *Lingua Sinica* 1: 1–23. [CrossRef]

Tang, Sze-Wing. 2016a. Sentence-final particles as conjuncts under the cartographic approach (Zhitu lilun yuzhuci de lianhe jiegoushuo). *Bulletin of Linguistic Studies (Yuyan Yanjiu Jikan)* 16: 1–10.

Tang, Sze-Wing. 2016b. A syntactic analysis of tag questions in English and Chinese (Yingyu he Hanyu yiwen weiju de jufa fenxi). *Foreign Language Teaching and Research (Waiyu Jiaoxue yu Yanjiu)* 48: 29–35.

Tsimpli, I. Maria, and Antonella Sorace. 2006. Differentiating interfaces: L2 performance in syntax-semantics and syntax-discourse phenomena. In *Proceedings of the 30th Annual Boston University Conference on Language Development (BUCLD 30)*. Somerville: Cascadilla Press, pp. 653–64.

Valdés, Guadalupe. 2000. Introduction. In *Spanish for Native Speakers*. AATSP Professional Development Series Handbook for Teachers K-16. Edited by N. Anderson. New York: Harcourt College, vol. 1, pp. 1–32.

Valenzuela, Elena, Michael Iverson, Jason Rothman, Kristina Borg, Diego Pascual y Cabo, and Manuela Pinto. 2015. Eventive and stative passives and copula selection in Canadian and American heritage speaker Spanish. In *New Perspectives on the Study of Ser and E star*. Edited by Isabel Pérez-Jiménez, Manuel Leonetti and Silvia Gumiel-Molina. Amsterdam: John Benjamins Publishing Company, pp. 267–92.

Wen, Xiaohong. 2014. Pragmatic Development: An Exploratory Study of Requests by Learners of Chinese. In *Studies in Second Language Acquisition of Chinese: A Series of Empirical Studies*. Edited by Zhaohong Han. Clevedon: Multilingual Matters, pp. 30–57.

Wen, Xiaohong. 2019. Requests in Chinese by heritage and foreign language learners. In *Studies on Learning and Teaching Chinese as a Second Language*. Edited by Xiaohong Wen and Xin Jiang. London: Routledge, pp. 38–63.

White, Lydia. 2011. Second language acquisition at the interfaces. *Lingua* 121: 577–90. [CrossRef]

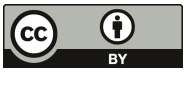
Xu, Jingning. 2003. Modality interpretation for the tone particle *ba* (Yuqi zhuci *ba* de qingtai jieshi). *Journal of Peking University (Philosophy and Social Sciences Section)* 40: 143–48.

Yan, Shanshan, and Boping Yuan. 2020. Asymmetric form-meaning mappings in L2 acquisition of Chinese sentence-final particles *ba*. Unpublished manuscript, last modified March 3. Microsoft Word file.

Yuan, Boping. 2014. "Wh-on-earth" in Chinese speakers' L2 English: Evidence of dormant features. *Second Language Research* 30: 515–49. [CrossRef]

Yuan, Boping. 2015. The effect of computational complexity on L1 transfer: Evidence from L2 Chinese attitude-bearing wh-questions. *Lingua* 167: 1–18. [CrossRef]

Yuan, Boping, and Esuna Dugarova. 2012. Wh-topicalization at the syntax-discourse interface in English speakers' L2 Chinese grammars. *Studies in Second Language Acquisition* 34: 533–60. [CrossRef]



MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Languages Editorial Office
E-mail: languages@mdpi.com
www.mdpi.com/journal/languages

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03943-271-4