

A Corpus-Based Analysis of Basic Verb Collocation Production by Japanese Learners of Chinese

Sakiko AKIYAMA

Kyushu University

Graduate School of Integrated Sciences for Global Society, Kyushu University, Japan

akiyama.sakiko.809@s.kyushu-u.ac.jp

Abstract

This study investigates the effects of L1-L2 congruency on the production of collocations involving three polysemous Chinese basic verbs, 开[kāi], 打[dǎ], and 放[fàng], among advanced Japanese learners of Chinese. Verb-noun co-occurrence data were extracted from the HSK Dynamic Composition Corpus and compared with native speaker reference data from the BCC Corpus. The quantitative analysis showed that learners did not uniformly favor congruent collocations. The analysis of learner errors further identified three processes underlying unnatural expressions: the projection of broad semantic categories associated with Japanese verbs, lexical selection based on semantic correspondences between Japanese and Chinese, and the overgeneralization of previously acquired L2 collocational schemas. These findings suggest that collocational errors do not result simply from direct L1 transfer but arise through interactions between L1 semantic knowledge and developing L2 lexical knowledge. The findings also have pedagogical implications, suggesting that Chinese collocations should be taught as integrated verb-noun units and that differences in the semantic ranges of verbs between Japanese and Chinese should be made explicit.

Keywords: collocation, corpus-based analysis, Chinese basic verbs, second language acquisition

1. Introduction

For second language learners, acquiring natural collocations remains challenging even at advanced proficiency levels (Nesselhauf 2003). Sinclair (1991: 114) argues that language use operates according to two principles: the *open-choice principle*, in which words are selected individually and combined freely, and the *idiom principle*, in which multiple words are processed together as prefabricated units. When learners have not yet fully acquired L2 collocations, they may rely on the vocabulary and expressive patterns of their first language when producing L2 output, resulting in expressions that are natural in the L1 but unnatural in the L2.

This tendency is particularly pronounced for basic verbs that have multiple meanings and uses and can co-occur with a relatively wide range of nouns. For such verbs, knowing the meaning of a word alone is often insufficient for producing appropriate collocations, making them particularly susceptible to L1 influence. It is therefore important to clarify the extent to which learners rely on L1-L2 correspondences when such correspondences exist and what kinds of misuse may result from this reliance in order to better understand the acquisition of L2 collocations more broadly.

This study accordingly examines how L1-L2 congruency affects the production of basic verb collocations in Chinese by native Japanese speakers learning Chinese. Specifically, it focuses on the basic verbs 开[kāi], 打[dǎ], and 放[fàng], which exhibit a “moderate restriction of collocability,” in order to investigate how L1-L2 congruency influences learner production.

2. Previous Studies

A substantial body of empirical research has examined L1 influence on second language processing from a psycholinguistic perspective. In particular, research in psycholinguistics and second language acquisition (SLA) has consistently shown that during L2 processing, the

equivalent L1 translation is automatically and unconsciously activated in learners’ minds, and that this activation can affect the processing and production of L2 collocations.

Studies by Chou and Morishima (2025), Chen (2024), and Ma and Hong (2026) have shown that when L1 and L2 expressions are congruent, L2 processing becomes faster and more accurate, a phenomenon known as the *Congruency Effect*. In particular, Ma and Hong (2026) demonstrated that stronger L1 activation is associated with a more pronounced congruency effect, identifying automatic L1 activation as a key factor underlying this effect. Yamashita and Jiang (2010) further showed that even learners in L2-rich environments are prone to errors in collocations where L1 and L2 expressions are not congruent. Jiang (2022) additionally points out that L2 learners find it difficult to fully grasp the strength of collocational bonding in the way native speakers do and that they rely heavily on L1-L2 semantic and expressive congruency as a cue in processing L2.

Taken together, these studies suggest that L1-L2 congruency plays a substantial role in the processing and acquisition of L2 collocations. However, most previous research has focused on learners of English as an L2. Although Ma and Hong (2026) examined Vietnamese learners of Chinese, the effect of L1-L2 congruency on collocation production among Japanese-speaking learners of Chinese remains insufficiently understood. It is therefore necessary to examine how L1 influence manifests itself in the production of Chinese collocations by native speakers of Japanese.

3. Purpose of the Study

This study accordingly focuses on the Chinese basic verbs 开, 打, and 放, which exhibit a moderate restriction of collocability and allow a relatively high degree of freedom in the nouns with which they co-occur. By comparing production data from advanced Japanese speaking learners of Chinese in the HSK Dynamic Composition Corpus with

native-speaker baseline data from BCC, this study aims to clarify how L1-L2 congruency affects the production of basic verb collocations. Specifically, the following two research questions are addressed:

RQ1: Do learners overproduce L1-L2 congruent collocations that correspond directly to expressions in their L1?

RQ2: What types of collocation errors are associated with the overapplication of L1 knowledge?

Based on these findings, the study further aims to draw implications for collocation-based instruction in the acquisition of basic verbs and for teaching approaches that take L1-L2 correspondences into account.

4. Methodology

4.1 Overview of the Corpora

This study uses the HSK Dynamic Composition Corpus as the source of production data from Japanese-speaking learners of Chinese and the BLCU Chinese Corpus (BCC) as the source of native speaker baseline data. The HSK Dynamic Composition Corpus is a learner corpus consisting of essay responses written by foreign examinees for the HSK examination between 1992 and 2005. From this corpus, 363 instances of nouns co-occurring with *kāi*, *dǎ*, and *fàng* were extracted from essays written by advanced Japanese-speaking learners.

BCC, on the other hand, is a large-scale corpus of contemporary Chinese produced by native speakers. It is used in this study as a baseline for identifying the co-occurrence patterns of the target verbs and nouns among native speakers. By comparing the two corpora, this study examines the characteristics of learners' collocation production and patterns associated with L1-L2 congruency.

4.2 Data Extraction and Processing

The extraction and processing of verb + noun combinations in this study combined automatic processing using natural language processing tools with manual verification by the researcher.

First, raw data extracted from the corpora using multiple search conditions were cleaned. Next, the Python3 natural language processing library spaCy was used to perform word segmentation and part-of-speech and dependency parsing. Nouns in a direct-object relationship with the target verb were extracted within a six-word window on either side of the verb. This method also captured instances such as 打了一个电话 (*dǎ le yí gè diànhuà*), in which an aspect marker or quantifier phrase intervenes between the verb and the noun.

Data that did not constitute suitable verb + noun combinations for analysis, including proper nouns, pronouns, and demonstratives, were then excluded, followed by a final manual verification by the researcher.

Using the resulting dataset, the frequency of each collocation in the BCC and HSK corpora was tabulated. A chi-square (χ^2) was then conducted after adjusting for

differences in corpus size to compare collocational production tendencies between the two corpora.

4.3 Criteria for Determining L1-L2 Congruency / Incongruency

In this study, L1-L2 congruency for each extracted verb + noun pair was determined based on the semantic correspondence between Japanese and Chinese. Congruency was assessed using the *Chinese-Japanese Dictionary* (3rd edition). A collocation was classified as "congruent" when its meaning could be explained by a Japanese expression corresponding to the first-listed sense of the Chinese verb. Collocations whose meanings could not be explained by the corresponding Japanese reference expression were classified as "incongruent."

The Japanese expressions used as the reference for each target verb were as follows:

开 [*kāi*] = *akeru*, *hiraku* ("to open")

打 [*dǎ*] = *utsu* (打つ) and *tataku* (たたく) ("to hit/strike")

放 [*fàng*] = *hanasu* (放す) ("to release/let go")

For example, 开门 [*kāi mén*], 打人 [*dǎ rén*], and 放人 [*fàng rén*] can be explained using the corresponding Japanese expressions *doa o akeru* ("to open a door"), *hito o tataku* ("to hit a person"), and *hito o hanasu* ("to release/let go of a person"), respectively, and were therefore classified as congruent. Similarly, 开宴会 [*kāi yànhuì*] ("to hold a banquet") was classified as congruent because its meaning can be expressed as *enkai o hiraku* (宴会を開く, "to hold a banquet") using the reference Japanese expression *hiraku*. In contrast, 开电视 [*kāi diànshì*] ("to turn on a television"), 打电话 [*dǎ diànhuà*] ("to make a phone call"), and 放假 [*fàng jià*] ("to have a holiday") were classified as incongruent because their meanings cannot be explained using the corresponding reference Japanese expressions for the respective verbs.

When multiple Japanese translations were possible, priority was given to an expression that satisfied the above criterion rather than to the most natural-sounding translation. For example, although 打孩子 [*dǎ háizi*] could also be translated as *kodomo o naguru*, it can equally be expressed as *kodomo o utsu*, which corresponds to the reference expression 打 (打つ) ("to hit/strike") used in this study. It was therefore classified as congruent.

5. Results

An analysis of the effect of L1-L2 congruency on learners' collocation production revealed distinct patterns across the three verbs in the distribution of collocations between learners and native speakers. For *kāi* and *dǎ*, statistically significant differences in collocational distributions were observed between learners and native speakers (both $p < .001$). The effect was particularly pronounced for *dǎ*, which showed the largest effect size ($\phi = .309$). Although the effect size for *kāi* was relatively small ($\phi = .086$), the difference in distribution was nevertheless statistically significant. In contrast, no statistically significant difference was observed for *fàng* ($p = .113$, $\phi = .064$). These findings indicate that the degree of correspondence

between learner and native-speaker collocational distributions varied across the three verbs.

V	Data	C	I	χ^2 value	CR	p	ϕ
开	HSK	65	22	21.48	74.7 %	< .001	.086
	BCC	1400	1429		49.5 %		
打	HSK	32	165	166.53	16.2 %	< .001	.309
	BCC	992	551		64.3 %		
放	HSK	1	78	2.51	1.3%	0.113	.064
	BCC	1	540		0.2%		

Table 1 Distribution of collocation production by learners and native speakers, by verb

Note. C = Congruent; I = Incongruent; CR= Congruency Rate

Verb	Noun	HSK Freq.	C / I
开	公司	8	C
	口	6	C
	晚会	4	C
	车	4	I
	商店	3	C
	电视机	3	I
	宴会	2	C
	摩托车	2	I
	服装店	2	C
	桃花	2	C

Table 2 Ten most frequent nouns co-occurring with *kāi*

For the verb *kāi*, congruent collocations accounted for 74.7% of learner production, compared with 49.5% of native speaker production. In the learner corpus, collocations such as 公司 [gōngsī], 口 [kǒu], 晚会 [wǎnhuì], and 商店 [shāngdiàn], which have clear semantic correspondences with the Japanese verbs *akeru* and *hiraku*, occupied prominent positions among the most frequently produced combinations. This distribution suggests that learners may preferentially produce collocations that have a clear semantic correspondence with Japanese, potentially reflecting the influence of their existing L1 knowledge that associates *akeru/hiraku* with *kāi*.

By contrast, native-speaker production included a wider range of incongruent collocations, such as 空调 [kōngtiào], 车 [chē], 玩笑 [wánxiào], and 灯 [dēng], alongside congruent collocations such as 门 [mén]. This distribution indicates that native speakers use *kāi* across a broader range of semantic contexts rather than restricting it primarily to the notion of “opening.” Depending on the noun with which it combines, *kāi* can express meanings such as “to turn on,” “to drive or operate,” and “to start.” Thus, the distributional pattern for *kāi* suggests that learners’ collocation production is more closely associated with L1-L2 semantic correspondence, whereas native speakers exhibit a broader range of collocational and semantic uses of the verb.

Verb	Noun	HSK Freq.	C / I
打	水	28	I
	电话	23	I
	网球	19	I
	棒球	16	I
	工	13	I
	孩子	9	C
	招呼	9	I
	排球	9	I
	乒乓球	5	I
	交道	5	I

Table 3 Ten most frequent nouns co-occurring with *dǎ*

For *dǎ*, the difference between learners and native speakers is more pronounced. Among native speakers, congruent expressions based on the core meaning of “to hit/strike,” such as 打人 [dǎ rén] (“to hit a person”) and 打玻璃 [dǎ bōli] (“to hit glass”), accounted for 64.3% of production, whereas incongruent expressions accounted for 83.8% of learner production. The native-speaker data also included a wide range of collocations referring to everyday activities, such as 打哈欠 [dǎ hāqian] (“to yawn”), 打游戏 [dǎ yóuxì] (“to play video games”), and 打交道 [dǎ jiāodao] (“to interact with someone”).

By contrast, although incongruent expressions predominated in learner production, they were concentrated in a relatively limited set of fixed collocations. This pattern suggests that, rather than productively extending the core meaning of *dǎ* to generate a diverse range of collocations, learners may rely primarily on prefabricated lexical chunks acquired during the learning process. Thus, for *dǎ*, the difference between learners and native speakers appears to lie not primarily in L1-L2 congruency itself, but in the extent to which learners are able to productively extend the verb across a diverse range of collocational contexts.

Verb	Noun	HSK Freq.	C / I
放	假	16	I
	暑假	9	I
	鞭炮	9	I
	音乐	6	I
	寒假	4	I
	机器	3	I
	烟	3	I
	烟火	3	I
	声音	2	I
	爆竹	2	I

Table 4 Ten most frequent nouns co-occurring with *fàng*

For *fàng*, no statistically significant difference was found between learners and native speakers ($p = .113$). Both groups produced incongruent expressions in over 98% of instances, and the surface-level distributions were highly similar. However, examination of the specific co-occurring nouns reveals qualitative differences in actual usage between the two groups. Learner production was concentrated on collocations involving 假 [jià] (“vacation/holiday”), such as 放假 [fàng jià] (“to have a vacation”), 放暑假 [fàng shūjià] (“to have a summer vacation”), and 放寒假 [fàng hánjià] (“to have a winter vacation”).

Native speakers, by contrast, favored expressions representing culturally and everyday Chinese activities, such as 鞭炮 [*biānpào*] (“firecrackers”), 风筝 [*fēngzheng*] (“kite”), 鸽子 [*gēzi*] (“to stand someone up; literally, ‘to release a pigeon’”), which ranked among the most frequent. Thus, for *fàng*, although no statistically significant distributional difference was found, learners appear to rely on a limited set of fixed collocations, whereas native speakers use a broader range of collocations that extend the core spatial meaning of “to release” to a variety of objects and contexts.

In summary, learners’ collocation production exhibits verb-specific characteristics. *kāi* shows a bias toward congruent collocations with a clear semantic correspondence to the L1, whereas *dǎ* shows a reliance on specific fixed collocations. Although *fàng* shows a distribution statistically similar to that of native speakers, differences remain in the internal composition and diversity of the collocations produced. These findings indicate that learner collocation production cannot be fully explained by a simple account of L1 transfer, according to which congruent expressions are produced more readily and incongruent expressions less readily. Rather, learners may rely on formulaic sequences for collocations they have already acquired, while constructing new combinations based on the meanings of individual words and existing linguistic knowledge when producing combinations that are not yet fully established or that require novel production. Learner collocation production therefore appears to involve both reliance on formulaic sequences and the ability to construct new combinations.

6. Discussion

Building on the statistical findings presented in the preceding chapter, this chapter provides a qualitative analysis addressing RQ2: what types of misuse arise from the overapplication of Japanese (L1) knowledge in Chinese production?

The unnatural expressions produced by learners for *kāi*, *dǎ*, and *fàng* reflect the overapplication of L1 semantic networks to L2 production. To substantiate this account, the discussion first establishes a theoretical framework based on language processing principles and then details the specific patterns and cognitive processes underlying these errors.

6.1 The Open-Choice Principle and Lexical Selection Based on L1 Knowledge

Human language processing can be understood in terms of two principles: the *open-choice principle*, whereby words with appropriate dictionary meanings are freely combined like pieces of a puzzle as long as grammatical rules are satisfied, and the *idiom principle*, whereby multiple words are processed together as a single prefabricated phrase or chunk. In native-speaker language use, the idiom principle has been argued to play a central role.

In L2 collocation production, however, these two principles may operate differently. As the preceding analysis suggests, learners appear to reproduce previously acquired collocations as formulaic sequences, while constructing combinations that are not yet fully established or that

require novel production on the basis of individual word meanings and grammatical knowledge. In other words, learner collocation production appears to involve two distinct processes: the reproduction of previously acquired collocations as formulaic sequences and the open-choice-based construction of unfamiliar combinations.

In particular, when Japanese learners construct unfamiliar or not-yet-established combinations through open choice, Japanese (L1) meanings and usage patterns, the semantic and embodied images associated with L1 expressions, and the formal similarity of Sino-Japanese homographs may influence Chinese (L2) lexical selection. Consequently, when learners have not yet fully internalized conventionalized Chinese word combinations, they may be more likely to rely on open-choice processing in lexical selection. At the same time, L1 knowledge is not the only source of information available to learners. Previously acquired L2 collocations and lexical schemas may also serve as cues for constructing new combinations. Thus, during open-choice processing, semantic and imagistic associations derived from L1 meanings, usage patterns, and homographs may be overapplied in some cases, whereas previously acquired L2 collocations and lexical schemas may be overgeneralized to new lexical items in others.

From this perspective, the errors observed above may not arise solely from negative L1 transfer but may result from multiple pathways, including (a) the overapplication of L1 meanings, usage patterns, and semantic imagery and (b) the overgeneralization of previously acquired L2 collocations and lexical schemas. Within these two patterns, we identify three distinct cognitive processes underlying learners’ unnatural productions: (1) the projection of broad L1 semantic categories and (2) the overgeneralization of L2-internal collocational schemas, both associated with pattern (a), and (3) lexical selection mediated by L1-L2 semantic correspondence, associated with pattern (b).

(a) Overapplication of L1 Semantic Knowledge

This pattern involves two related processes: the projection of the broad semantic category of a general-purpose Japanese verb, which can combine with a wide range of objects, onto the selection of a basic Chinese verb, and the extension of a previously acquired Chinese collocational schema to a new, semantically similar lexical item. In Japanese, a single verb may cover a broad semantic domain, whereas Chinese often requires different verbs depending on the object or type of action. Similarly, once learners have acquired a particular verb-noun combination in Chinese, they may extend the corresponding combinatory pattern to another semantically similar noun. Such errors may therefore arise through two pathways: the overprojection of L1 semantic categories and the overgeneralization of previously acquired L2-internal schemas.

For example, in the case of *kāi* corresponding to “to drive” (*untēn suru*, 運転する), the Japanese verb *untēn suru* can be used with a wide range of vehicles, including cars and motorcycles. Drawing on this broad semantic category of “to drive a vehicle,” a learner may select the Chinese basic

verb *kāi*, producing 开摩托车 [*kāi mótuōchē*] (“to ride a motorcycle”), as in the following example.

(1) 我爸爸很喜欢*开摩托车。

Wǒ bàba hěn xǐhuan kāi mótuōchē

(My father really enjoys riding motorcycles.)

In Chinese, however, the verb 骑 [*qí*], as in 骑摩托车 [*qí mótuōchē*] (“to ride a motorcycle”), is generally used to describe the action of riding and operating a two-wheeled vehicle. This example suggests that the broad semantic domain of the Japanese verb *untēn suru* may have influenced the selection of *kāi*, while the learner may also have extended the already-acquired collocational schema 开车 [*kāi chē*] (“to drive a car”) to a semantically similar object, “vehicle + driving.” Rather than attributing this error solely to direct negative transfer from the L1, it is therefore necessary to consider that both the L1 semantic category and an already-acquired L2 schema may have jointly served as cues for the selection of *kāi*.

A similar tendency can be observed in the case of *dǎ* corresponding to *suru* (する, “to do or play”). In Japanese, the general-purpose verb *suru* can be used with a wide range of sports and activities, as in *jūdō o suru* (柔道をする, “to do jūdō”), *barēbōru o suru* (バレーボールをする, “to play volleyball”), and *bujutsu o suru* (武術をする, “to practice martial arts”). The broad semantic category associated with *suru* may therefore have influenced the selection of the Chinese verb *dǎ*. At the same time, the previously acquired collocation 打排球 [*dǎ páiqiú*] (“to play volleyball”) may have provided a more general pattern for expressing participation in sports. The learner may then have extended this pattern to the new sport name 柔道 [*róudào*] (“jūdō”).

(2) 我哥喜欢*打柔道

wǒ gē xǐhuan *dǎ róudào

(My older brother likes doing jūdō.)

In Chinese, however, the verb 练 [*liàn*] is used for the practice or training of jūdō and martial arts, as in 练柔道 [*liàn róudào*] (“to practice jūdō”) and 练武术 [*liàn wǔshù*] (“to practice martial arts”), while *dǎ páiqiú* is a natural expression. This suggests that learners did not necessarily translate the Japanese *suru* directly into *dǎ*, but rather overgeneralized the L2-internal combinatory schema “*dǎ* + sport,” formed from the already-acquired *dǎ páiqiú*, to a different sport name, jūdō. That is, in addition to the broad semantic category carried by the Japanese *suru* potentially serving as a background cue, a combinatory pattern formed from an already-acquired L2 collocation may also have been extended to a new lexical item, contributing to the formation of this error.

Furthermore, in the case of *fàng* corresponding to “to broadcast/stream”, the Japanese verb *nagasu* (流す, “to broadcast or stream”) can also be used to mean “to broadcast” or “to play” content such as television programs or videos. Drawing on this meaning, learners may select the Chinese verb *fàng* as the corresponding verb, producing expressions such as the following.

(3) 在日本的电视新闻上,经常*放了关于这种问题的内容。

Zài Riběn de diànshì xīnwén shàng, jīngcháng fàng le guānyú zhè zhǒng wèntí de nèiróng.

(Japanese TV news often reports on issues like this.)

In Chinese, however, the verbs 播 [*bō*] or 播放 [*bōfàng*] is typically used to describe “broadcasting” content on television. The learner may therefore have mapped the meaning and usage of *nagasu* onto the Chinese verb 放 without sufficiently considering the conventionalized verb choice in Chinese. This example appears to reflect a more direct influence of L1 meaning and usage than the case of *dǎ róudào*, where the influence of an already acquired L2 collocational pattern appears to be more prominent.

These three examples share the feature that learners applied an existing semantic category or lexical schema when constructing a new collocation. However, the source of this knowledge was not the same in all cases. In *kāi mótuōchē*, both the broad semantic category associated with the Japanese verb *untēn suru* (運転する, “to drive”) and the previously acquired L2 collocation *kāi chē* (“to drive a car”) may have contributed to the selection of *kāi*. In *dǎ róudào*, the overgeneralization of an already acquired L2 collocation such as *dǎ páiqiú* appears to have played a particularly important role. In 放内容 [*fàng nèiróng*], by contrast, the correspondence between the meaning and usage of the Japanese verb *nagasu* and the Chinese verb *fàng* appears to have served as a more direct cue for lexical selection.

These findings suggest that learner collocation errors cannot be fully explained by negative L1 transfer alone. The overprojection of L1 semantic categories and the overgeneralization of previously acquired L2 schemas may both contribute to errors in the construction of unfamiliar collocations. When learners produce combinations that have not yet been fully established in their L2, they may draw not only on the meanings of individual words but also on patterns derived from previously acquired collocations. If the conventional constraints governing verb-noun combinations in Chinese have not yet been sufficiently internalized, these existing semantic categories and lexical schemas may be extended to inappropriate combinations, resulting in collocation errors.

(b) Interference from L1 Semantic Correspondence

This pattern involves an excessive reliance on semantic correspondences between Japanese and Chinese. Learners may select a Chinese verb that appears to correspond to the meaning or usage of a Japanese verb without sufficiently considering the collocational constraints specific to Chinese. In such cases, the perceived correspondence between the Japanese and Chinese verbs may serve as a cue for lexical selection. This correspondence does not necessarily involve a simple one-to-one translation relationship. Rather, it may provide a starting point that interacts with the learner’s existing L2 lexical knowledge and previously acquired collocational schemas.

For example, in the case of *dǎ* (“to sound or honk”), corresponding to *narasu* (鳴らす, “to sound or honk”) in Japanese, the following expression was produced.

- (4) 我觉得在中国开车的*打喇叭打得有点过头了, 真令人讨厌。

Wǒ juéde zài Zhōngguó, kāichē de *dǎ lǎba dǎde yǒudiǎnr guòtóu le, zhēn lìng rén tǎoyàn.

(I think drivers in China honk their horns a little too much. It’s really annoying.)

It is difficult to conclude that the learner selected *dǎ* simply by associating the physical action of “hitting” with the Japanese verb *narasu*. The Japanese expression for honking a horn is *kurakushon o narasu* (クラクションを鳴らす, “to honk a horn”), and the verb *narasu* does not conventionally describe the physical action of striking a horn. It is therefore unlikely that the learner directly mapped the physical action of “hitting” onto the Japanese verb.

Instead, the error may reflect the semantic proximity between the meaning of “producing a sound” associated with *narasu* and one of the uses of *dǎ* derived from its core meaning of “to hit or strike.” The learner’s previously acquired knowledge of *dǎ* may also have contributed to this selection. In Chinese, *dǎ* can be used when striking an object produces a sound, as in 打鼓 [*dǎ gǔ*] (“to beat a drum”). Building on this existing L2 knowledge, the learner may have selected *dǎ* as a verb associated with the semantic domain of producing a sound and constructed the new combination 打喇叭 [*dǎ lǎba*].

In Chinese, however, honking a horn is expressed as 按喇叭 [*àn lǎba*] (“to honk a horn”), based on the action of pressing the horn. This suggests that the Japanese verb *narasu* may have served as a semantic cue for selecting the Chinese verb *dǎ*, while the learner’s previously acquired knowledge of *dǎ* may have supported the extension of its use to a new noun. The learner may therefore not have sufficiently internalized the conventional Chinese collocation *àn lǎba*.

A similar tendency can be observed in the case of *fāng* (“to spread or apply”), corresponding to *maku* (まく, “to spread or apply”) in Japanese. The Japanese verb *maku* can be used with a range of objects, including pesticides and fertilizer, as in *nōyaku o maku* (農薬をまく, “to apply pesticide”) and *hiryō o maku* (肥料をまく, “to apply fertilizer”). The semantic image of spreading or dispersing an object associated with *maku* may therefore have influenced the selection of *fāng*.

- (5) 每年都要*放农药、化肥。

Měinián dōu yào *fàng nóngyào, huàfēi.

(Every year, we have to apply pesticides and chemical fertilizers.)

In Chinese, however, different verbs are selected depending on the object: 喷农药 [*pēn nóngyào*] (“to spray pesticide”) for spraying pesticide, and 施化肥 [*shī huàfēi*] (“to apply fertilizer”) for applying fertilizer. This example

suggests that mapping the semantic domain expressed by the single Japanese verb *maku* onto the Chinese verb *fāng* resulted in an insufficient grasp of the lexical differentiation in Chinese, whereby different verbs are selected depending on the object.

These two examples suggest that when a semantic correspondence exists between L1 and L2, this correspondence can serve as an important cue in L2 lexical selection among Japanese learners of Chinese. However, such correspondence alone does not determine the appropriate collocation. When learners select an L2 verb on the basis of an L1 semantic category but have not yet fully acquired the combinatory relationships between that verb and particular nouns, they may extend an existing L2 usage to a new noun or construct a new combination on the basis of an L1 semantic association.

The errors observed here should therefore not be understood simply as cases of word-for-word negative transfer in which a Japanese verb is directly substituted with a Chinese verb. Rather, they appear to arise during the construction of new collocations, where L1 semantic correspondence serves as a cue for lexical selection and interacts with previously acquired L2 lexical knowledge and collocational schemas. This perspective suggests that L1 influence may operate not only through direct transfer of individual lexical meanings but also through the interaction between L1 semantic knowledge and developing L2 lexical representations.

6.2 What Types of Collocation Errors Result from Overapplication of L1 Knowledge?

Section 6.1 identified two patterns in the production of Chinese collocations by native Japanese speakers: (a) Overapplication of L1 Semantic Knowledge, and (b) Interference from L1 Semantic Correspondence. However, not all of the errors observed in these patterns can be attributed directly to negative transfer from L1 to L2. As the preceding analysis has shown, learners sometimes extend previously acquired Chinese collocations to new lexical items. Learner errors therefore appear to involve an interaction between L1 semantic knowledge and previously acquired L2-internal schemas.

The basic verbs examined in this study are high-frequency words typically introduced at the beginner level, and their core meanings are presumably acquired relatively early in the learning process. The occurrence of collocational errors involving these verbs nevertheless suggests that knowledge of a verb’s meaning does not necessarily entail knowledge of its collocational restrictions, particularly the nouns with which it conventionally combines.

This point can also be understood in terms of Jiang’s (2000) model of L2 lexical development. According to Jiang (2000), until L2 words become fully integrated with semantic and syntactic information that is independent of the L1, the lemma information associated with the corresponding L1 translation continues to participate in L2 word processing. From this perspective, when native Japanese speakers produce Chinese basic verbs, the corresponding Japanese verb may continue to serve as a cue for lexical selection alongside the meaning of the L2 verb itself.

However, L1 influence does not necessarily result in erroneous expressions through direct transfer alone. For example, in *dǎ lǎba*, the Japanese verb *narasu* may have served as a semantic cue associated with producing a sound, while the learner may also have extended the previously acquired use of *dǎ* to the new noun *lǎba*. Similarly, in *dǎ róudào* and *kāi mótuōchē*, both L1 semantic categories and previously acquired L2 schemas, such as *dǎ páiqiú* and *kāi chē*, may have contributed to the selection of the verbs. In *fàng nóngyào*, by contrast, the mapping of the semantic category associated with the Japanese verb *maku* onto *fàng* appears to have exerted a more direct influence on lexical selection.

These findings suggest that the errors observed in this study are not simply instances of negative transfer from L1 to L2. Rather, they appear to arise during the construction of new collocations through the interaction of existing L2 lexical knowledge and schemas with L1 semantic knowledge. In particular, when L2-specific collocational knowledge has not yet been sufficiently established, learners may draw on L1 semantic categories as well as previously acquired L2 schemas when selecting lexical combinations. These sources of knowledge may consequently be extended beyond their conventional ranges, resulting in deviations from the collocational constraints governing verb-noun combinations in Chinese.

7. Conclusion

This study examined the collocation production of the Chinese basic verbs *kāi*, *dǎ*, and *fàng* among advanced Japanese speaking learners of Chinese, using learner data from the HSK Dynamic Composition Corpus and native-speaker reference data from BCC.

For RQ1, learners did not uniformly favor congruent collocations. Congruent expressions were more frequent with *kāi*, whereas incongruent expressions predominated with *dǎ*. This suggests that L1-L2 congruency does not affect all verbs in the same way and that highly frequent formulaic expressions may become strongly entrenched even when they are incongruent with the L1.

For RQ2, learner errors involved three distinct cognitive processes operating across the two descriptive patterns: the projection of broad L1 semantic categories and the overgeneralization of previously acquired L2 collocational schemas (under the pattern of L1 semantic knowledge overapplication), and the selection of verbs based on L1-L2 semantic correspondence (under the pattern of L1 semantic correspondence interference).

Pedagogically, verb-noun combinations should be taught as integrated lexical units, differences in semantic range between Japanese and Chinese should be made explicit, and L2 collocational constraints should be taught to support more flexible production.

The study is limited by the relatively small number of learner tokens, the focus on only three verbs, and the lack of longitudinal data. Future research should examine larger datasets, a wider range of basic verbs, and collocational development across proficiency levels.

8. References

- Chen, Y. (2024). The congruency effect in L2 collocational processing: The underlying mechanism and moderating factors. *Studies in Second Language Acquisition*, 46(1), 75–95.
- Chou, I.-S., & Morishima, Y. (2025). Daiichigengo to dainigengo ni okeru korokēshon shori e no hindo to icchisei no eikyō no gaiyō [An overview of the effect of frequency and congruency on collocational processing in the first and second language]. *Kokusai Kirisutokyō Daigaku Gakuhō 1-A: Kyōiku Kenkyū*, (67), 71–76.
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21(1), 47–77.
- Jiang, Y. (2022). Effects of L1-L2 congruency, collocation type, and restriction on processing L2 collocations. *Frontiers in Psychology*, 13, 947725.
- Ma, L., & Hong, W. (2026). How do congruency and L2 proficiency impact the processing of Chinese L2 collocations? Evidence from different L1 activation states. *International Journal of Applied Linguistics*, 36(1), 367–382.
- Nesselhauf, S. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223–242.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, 44(4), 647–668.

9. Language Resource References

- Shōgakukan. 2016. *Chū-Nichi jiten* [Chinese-Japanese Dictionary], 3rd edition.
- BCC Corpus (<https://bcc.blcu.edu.cn/>)
- HSK Dynamic Composition Corpus (<https://hsk.blcu.edu.cn/>)