The activation and selection of lexico-syntactic features in speech production: behavioural and electrophysiological evidence from L1 and L2 speakers

Wang, S.

CHAPTER 5

L2 noun phrase production of English-Chinese speakers: a web-based experiment

Abstract
The processing of classifiers in L2 noun phrase production of Mandarin Chinese learners without a corresponding lexico-syntactic feature in their L1 has been rarely studied. Here, we explored the modulation effect of classifiers in English-Chinese speakers using the picture-word interference paradigm in a web-based setting. Participants overtly named pictures using noun phrases while ignoring superimposed distractor words in a picture-naming task. We recorded naming responses and extracted naming latencies off-line from audio files. Despite the small sample size, we found descriptive trends for the effects of classifier congruency and semantic interference. We discussed challenges and recommendations for running online L2 production experiments with overt articulation.

5.1. Introduction

In Chapter 2 of this thesis, we investigated how Dutch learners of Mandarin Chinese process classifiers (CL) in Chinese noun phrase (NP) production. The study included a picture-naming task using the picture-word interference (PWI) paradigm (Bürki et al., 2020) to test Dutch learners of Chinese with intermediate proficiency levels. As an extension, a question arises: how do learners without gender or classifiers in their L1 process classifiers in L2 Chinese? The current chapter investigated this question specifically with English learners of Chinese.

The English language does not employ lexico-syntactic features such as gender or classifiers (Tokowicz & MacWhinney, 2005). Numerals can

combine with nouns when the unit of entity is clear, as in “two tables” and “three doors”, specifying the exact number of entities. In contrast, in Chinese, classifiers are always required to specify the unit of entities regardless of what the noun denotes. For example, the noun for “table” requires the classifier 张 “*zhang*” when specifying the number of tables, forming the NP 两 张 桌 子 “liang *zhang* zhuozi” [two CL tables]. This typical NP construction involves three components, i.e., numeral - classifier - noun (Cheng & Sybesma, 1999, 2012; Li & Thompson, 1981).

Generally, in Chinese, the unit of quantification must be explicitly specified through the use of classifiers, while in English, this is not required. The exception in English is when the entity unit cannot be determined, as in the case of mass nouns, i.e., water or paper, where a measure word is necessary to specify the unit, such as “two *bottles of* water” and “one *piece of* paper”. Therefore, the primary distinction between English measure words and Chinese classifiers in NPs lies in the reference to countable entities, such as tables and doors.

In this study, we only focused on classifiers used to specify countable entities, a feature not employed in English. By studying how L1 English speakers process classifiers in L2 Chinese, we aim to provide insights into how speakers without a corresponding syntactic feature in their L1 process such a feature in their L2. Chinese classifiers are determined according to the properties of entities, such as animacy, shape, functionality, and size (Shi, 1996; Tai, 1994; Tai & Chao, 1994; Tai & Wang, 1990). For instance, in the case of a table, the classifier 张 “*zhang*” is commonly used to specify objects with a flat surface (Tai & Chao, 1994). 条 “*tiao*” is another shape-based classifier that is often used to label entities with elongated shapes, such as ropes and snakes (Tai & Wang, 1990).

Although the selection of classifiers is, to some extent, based on the natural properties of the individuals, the application of classifiers still needs memory representation, as merely knowing the semantics of nouns is not sufficient for selecting their proper classifiers. For instance, both 桌子 “zhuozi” [table] and 门 “men” [door] have a flat surface, yet they pair with different classifiers, 张 “*zhang*” and 扇 “*shan*”, respectively. Moreover, the

same noun can be associated with different classifiers. For instance, 画 "hua" [painting] can be paired with the classifier 张 "*zhang*" to indicate a flat surface or with a more specific classifier 幅 "*fu*" in a more formal occasion (Zhang, 2007). Therefore, the combination of classifiers and nouns cannot be generalized by a single rule. Still, the choice of classifiers is predominantly influenced by the noun itself within the given context (Shao, 1993).

The acquisition and processing of a classifier system is a challenge for learners of Chinese, regardless of whether or not their L1 has gender. Dutch learners of Chinese have a two-gender system in their L1, but the system is different from the classifier system. Although both gender and classifiers are inherent properties of nouns and cannot be omitted in NP production, gender assignment in Dutch is less predictable from the properties of an entity and basically follows a one-to-one mapping between nouns and gender. L2 processing can be modulated by proficiency level and other individual variables such as the age of acquisition (AoA), the duration of learning, input quality, and vocabulary knowledge, etc. (Cornips & Hulk, 2008; Hao et al., 2021; Unsworth, 2008). Among Dutch learners of Chinese who reached intermediate proficiency levels and had good knowledge of classifiers, we observed evidence for their sensitivity to classifiers in Chinese NP production at behavioral and electrophysiological levels (Chapter 2). Behaviorally, the sensitivity was reflected in faster production of classifier congruent NPs compared to incongruent NPs in the PWI task. This replicated the classifier congruency effect observed in L1 Chinese NP production (Huang & Schiller, 2021).

Compared to Dutch L1 speakers, English L1 speakers may be relatively less familiar with assigning a lexico-syntactic feature (e.g., classifier) to a noun, and it may be more challenging for English L1 speakers to apply the correct classifier rules in this regard. This difficulty can be attributed to the greater structural distance between English and Chinese, as research suggests that L2 speakers face increased challenges in learning and processing when the structural distance between their L1 and L2 is greater (Myles, 1995; Sabourin, 2001; White et al., 2004). The model relevant to this study is the *Language Distance Hypothesis* (LDH) (Zawiszewski & Laka, 2020), which provides a theoretical framework for understanding L2 processing modulated

by similarities and differences between L1 and L2. The model predicts greater processing challenges in cases of morphologically different structures between the L1 and L2. In this chapter, we adopted the same tasks and materials employed in Chapter 2, testing Dutch-Chinese speakers to investigate English-Chinese language pairs. An innovative aspect of our approach was the transition from a lab-based design to a web-based setting for the L2 production experiment. Online settings offer several advantages compared to lab-based settings. For example, it allowed us to access a diverse pool of participants across countries and nationalities (e.g., Gallant & Libben, 2019; Peer et al., 2017) while also consuming fewer resources (e.g., Grootswagers, 2020). Web-based platforms have been considered an efficient alternative in the restrictions of the COVID-19 pandemic. As a result, online platforms have facilitated the replication and extension of behavioral experiments, including those that require precise measures of reaction times (Anwyl-Irvine, Dalmaijer, et al., 2020; Anwyl-Irvine, Massonnié, et al., 2020; De Leeuw, 2015; Gallant & Libben, 2019; Hilbig, 2016; Pinet et al., 2017).

However, online language production experiments involving overt articulation have been relatively scarce. This may be due to concerns regarding technical difficulties and data quality associated with capturing accurate voice onset latencies and synchronizing stimuli (Bridges et al., 2020). Nevertheless, recent evidence suggests that recording audio data online to investigate speech production is feasible. Fairs and Strijkers (2021) successfully replicated the word frequency effect in a picture-naming paradigm on the platform FindingFive (FindingFive Team, 2019). Moreover, both Stark et al. (2022) and Vogt et al. (2021) replicated lab-based effects of semantic interference in both overt naming and key-pressing response modalities with L1 speakers using the PWI paradigm. Vogt et al. (2021) implemented their experiment in SoSciSurvey (Leiner, 2019; Khan, 2020) and jsPsych (De Leeuw, 2015) platforms, each with 48 participants. Similarly, Stark et al. (2022) conducted their overt naming task (Experiment 1) using the same SoSciSurvey platform with 30 participants, providing evidence for the feasibility of time-sensitive overt naming experiments in web-based settings when employing a sufficient sample size.

The web-based experiment requires careful consideration of the technical setup due to potential variations in equipment across participants, which may compromise data quality (Anwyl-Irvine, Dalmaijer et al., 2021; Bridges et al., 2020). For example, Kim et al. (2020) found that the input and output components of the audio interfaces could lead to delays of around 5-15 ms. However, it is important to note that, as of now, none of the online experimental software can ensure perfect synchrony (Vogt et al., 2021). Nevertheless, high data precision for within-participant comparison can be obtained by a large sample size and sufficient trials (Bridges et al., 2020; Stark et al., 2022). Thus, we expected that with adequate statistical power (Vogt et al., 2021) the effects of interest should be detectable.

Conducting L2 overt production in web-based settings presents several challenges. Technical hurdles, such as equipment variations (e.g., microphone quality, internet speed, etc.), may impact the precision of time-locked vocal responses to picture stimuli displayed on a screen. Moreover, the absence of direct experimenter monitoring poses a challenge to online experiments. In the lab, experimenters can promptly address issues if participants do not comply with instructions or lose focus during the task. However, in online production tasks, experimenters can only check participants' responses after the experiment.

Another challenge is recruiting genuine L2 participants online. Unlike in the lab, where proficiency levels of L2 speakers can be assessed intuitively and relatively efficiently, online settings have to rely on participants' self-reported proficiency levels. As reported by Fairs and Strijkers (2021), around 20% of "bad-faith" L1 participants in their overt naming experiment merely ran the task for payment without actual participation, even when recruited from their lab mailing list. Given the greater difficulty of overt naming in L2 compared to L1, L2 participants may remain silent or produce non-sense responses when unsure how to respond. Moreover, the lack of a reliable mailing list with genuinely interested and qualified L2 participants is a concern. Therefore, encountering "bad-faith" and non-qualified participants must be anticipated in L2 online studies. Overcoming these challenges will be crucial for the reliability and validity of data collected in online overt L2 production experiments.

5.1.1. The current study

The current study extended Chapter 2 and built upon Huang and Schiller (2021) to investigate the classifier processing in L2 NP production of English learners of Chinese. We study this through a picture-naming task using the PWI paradigm. The experimental materials were kept identical to those used in Chapter 2.

In the PWI task, *classifier congruency* (congruent vs. incongruent) and *semantic interference* (related vs. unrelated) were manipulated between target pictures and distractor words. The task was to name pictures in Chinese using NPs, such as 两张桌子 "liang *zhang* zhuozi" [two CL tables], while ignoring distractor words superimposed on pictures.

The research questions were, first, can the semantic interference effect be found in online L2 PWI experiment? If naming is delayed by semantically related distractor words, this would indicate lexical competition in L2 NP production. Second, can the classifier congruency effect be replicated? If naming is facilitated by classifier-congruent distractor words, this would suggest competition at the lexico-syntactic level. Third, what recommendations can be provided regarding running L2 overt naming experiment online?

In order to increase the recruitment efficiency and try to prevent the "bad-faith" participants from taking part in the PWI task, we set proficiency tests as a separate phase prior to the main experimental (second) phase as a pre-screening. This is an effort to ensure that the second phase invites only participants who have a sufficient L2 proficiency level (intermediate level) and equipment functionality. Proficiency tests included a classifier knowledge test, a Mandarin Elicited-Imitation test (EI; Yan et al., 2020), and a Language History Questionnaire (Li et al., 2020). Tasks were identical to Chapter 2.

The experiment was programmed and run on the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine, Massonnié, et al., 2020), a tooled experiment design, host, and run platform that is fully compliant with the EU General Data Protection Regulation and with NIHR and BPS guidelines. Participants were recruited via the commercial platform Prolific

(www.prolific.co.uk; Palan & Schitter, 2018) and the participant pool from the Faculty of Social Science (FSW) at Leiden University.

5.2. Methods

5.2.1. Participants

In the proficiency test (first) phase, a total of 510 native English speakers, aged between 18 and 50 years, signed up and accessed the task via Prolific. They were from four countries (the United Kingdom, the United States, Canada, and Australia). They were provided with the experiment information in Prolific before signing up. When participants opened the experiment link in Prolific but did not complete it, they were added to a randomized list but were not considered as paid participants. The large majority of the participants (i.e., 494 or 96.67%) did not complete the first phase and was removed from the data analysis. Among them, 92.94% (475) dropped out and returned to Prolific without submissions, and 2.55% (13) timed out. 1.18% (6) had to be manually rejected due to issues with the returned link.

There were 16 participants in the first phase who were assigned as completed participants because they successfully submitted responses and were invited to the second phase, in which 9 of them actually took part. The final sample consisted of 5 participants (1 female; M = 32.80 years of age, SD = 9.26) who met all inclusion criteria. See Section 3.1 for details on data exclusion.

The basic criteria established using the Prolific defaulted pre-screen function to allow only participants who meet the requirements to find this study were as follows: the L1 is English; being "fluent" in Chinese and English (see discussion in Section 5.4); no reported language-related disorders such as dyslexia; no impairments or diseases such as hearing difficulties, neurodiversity, and anxiety.

Participants read and signed informed consent forms before the experiment in compliance with the Ethics Code for linguistic research in the Faculty of Humanities at Leiden University. Participants received monetary compensation distributed via Prolific. The data from participants who did not

provide consent, dropped out, or timed out (i.e., exceeded the maximum duration of 1 hour and a half for each phase) were automatically excluded by Gorilla and were not included in the analysis.

Participants' linguistic profile was assessed through an adapted version of the Language History Questionnaire LHQ (Li et al., 2020) in the first phase. In the LHQ, all five participants in the final sample indicated that English was their L1. Three participants reported Mandarin Chinese as their L2 with M_{AOA} = 16.33 (SD = 5.31). Two participants reported Mandarin Chinese as their L3 with M_{AOA} = 2.00 (SD = 1.00). One participant reported Cantonese as L2. Moreover, three participants provided additional information that they were raised by Cantonese-speaking parents or had one parent who spoke Cantonese at home. Their parents did not speak Mandarin Chinese. On a scale from zero to seven (seven being maximally proficient), participants reported a speaking proficiency of M = 5.40 (SD = 1.02), listening proficiency of M = 5.20 (SD = 1.17), reading proficiency of M = 4.40 (SD = 1.86), and writing proficiency of M = 4.00 (SD = 1.90).

5.2.2. Tasks and materials

The experiment was split into two phases. The first phase was pre-screening, including an EI task (Yan et al., 2020), a classifier knowledge task, and a language background questionnaire LHQ (Li et al., 2020). The second phase was the main experiment, where participants performed a picture-naming task and a LexTale-Ch vocabulary test (Chan & Chang, 2018). Tasks and materials were almost identical to those in Chapter 2. The biggest difference was they were transmitted to the online platform.

Elicited imitation (EI) task

To evaluate participants' L2 listening and speaking proficiency and assess the quality of their microphone recordings, we employed an adapted version of the EI task (Yan et al., 2020). In this task, participants listened to a sentence in Mandarin and then repeated it as accurately as possible. The EI task is based on the theoretical rationale that accurate repetition requires proper comprehension of the sentence's meaning. This task has been recognized as a valid and practical assessment of language proficiency (Bowles, 2011; Ellis, 2005; Yan et al., 2020).

The adapted version consisted of 36 stimulus sentences, which is half the original material in Yan et al. (2020). Each sentence contains key vocabulary and/or Chinese grammar corresponding to the three proficiency levels (beginner, intermediate, and advanced). The stimulus audios provided by the original authors (Yan et al., 2020) were recorded by a male native Mandarin speaker, maintaining a speech rate similar to the typical language input encountered by learners in Mandarin classrooms. These audios were uploaded to Gorilla for administration on participants' devices.

Classifier knowledge test

In order to assess participants' knowledge of classifiers, we adapted the classifier knowledge test. The test consisted of twenty multiple-choice questions, where each question presented an NP with a blank space for the classifier. For example, 一__书 "yi (ben) shu" [one (CL) book]. Four options of classifiers were provided for each blank, and participants were required to select the correct one.

LexTale-Ch test

The LexTale-Ch test (Chan & Chang, 2018) was employed to assess participants' Chinese vocabulary. To use it on Gorilla, we adapted the original web-based test. The test consists of 90 characters, including 60 real items and 30 nonce items. Participants were required to identify the real characters from the given options. The materials were obtained from the original authors. Each individual character was uploaded to Gorilla as a separate picture.

Picture-naming task

The picture-naming task employed a 2 × 2 factorial within-subjects design using the PWI paradigm. Two factors were manipulated: classifier congruency (congruent vs. incongruent) and semantic relatedness (related vs. unrelated), leading to four experimental conditions: classifier congruent and semantically related (C+S+), classifier incongruent and semantically related (C-S+), classifier congruent and semantically unrelated (C+S-), and classifier incongruent and semantically unrelated (C-S-).

The target nouns, classifiers, as well as distractor nouns used in the task were identical to Chapter 2. All Chinese stimuli were selected from

"Integrated Chinese" (Liu & Yao, 2009), a series of Mandarin textbooks used by China Studies at Leiden University. The selection of target classifiers and nouns was based on three criteria. First, items were selected from textbooks targeted at beginning and intermediate learners. This was an effort to ensure that participants with intermediate proficiency levels could understand the materials. Second, each noun represented clear concepts and was easily depicted. Third, we specifically considered nouns that could share a classifier despite belonging to different semantic categories. For example, both 鱼 "yu" [fish] and 河 "he" [river] share the classifier 条 "*tiao*", despite belonging to the categories of animals and natural landscapes, respectively. A total of nine target nouns representing concrete categories and three classifiers were selected. Each classifier was assigned three nouns. Nine black-and-white drawings from Severens' picture database (Severens et al., 2005) or similar line drawings were selected as picture stimuli. In order to control for quantifiers in NPs, the pictures were presented either individually (numeral 一 "yi" [one]) or in pairs (numeral 两 "liang" [two]). Each picture was accompanied by four distractor words. Each combination of target pictures and distractor words was assigned to one of four experimental conditions, resulting in a total of 18 stimuli per condition (nine pictures x two number conditions) and 72 stimuli in total. These stimuli were presented in a pseudo-random order generated by the Windows program MIX (Van Casteren & Davis, 2006).

5.2.3. Design and procedure

Participants first read the study information on Prolific, which included language and device requirements, as well as the general procedure. They were informed that the first phase of the experiment would be conducted first, and if they completed and passed it, they would be invited to participate in the second phase within two days. Participants accessed the experiment through the Gorilla link.

They were instructed to check their browser sound settings and microphone through two testing nodes before accessing the consent form. This step was an effort to ensure that the auto-play feature on participants' browsers and the functionality of their microphones were enabled and suitable for use in the EI task and the picture-naming task, both of which

task designs require listening and speaking. If their devices did not work properly, they had the option to quit the experiment and return to Prolific with minimal loss of time.

In the sound testing node, participants were instructed to close applications using sound functions (e.g., communication software, games, or video players) and adjust the volume to an appropriate level. Then, a test music piece was automatically played using the *Web Audio Zone*. Participants clicked the "Yes" button to indicate they could hear the music. If participants could not hear the music and clicked the "No" button, the screen would be moved to detailed instructions with explanations and screenshots on setting auto-play on the browser.

In the subsequent microphone testing node, participants were instructed to accept the microphone permission notification and click the "Start Recording" button on the screen to initialize their microphone and speak a sentence within 3,000 ms. The voice was recorded by the *Audio Recording Zone* in the *Test Mode*. The recorded audio would be played after clicking the "Stop Recording" button. If they could hear their audio clearly, they clicked the "Yes" button to proceed, indicating that the microphone was functioning properly. If they encountered issues, they were advised to adjust the microphone settings and given the option to attempt again until the issues were solved.

Elicited imitation task

The first phase of the experiment started with the EI task. Participants were instructed to turn on the full-screen mode to reduce distractors from the environment and practice by listening to a sentence first and then repeating the sentence after hearing a ringtone and seeing a microphone picture. They were informed that imitation of the sound was unnecessary. The voice in practice was recorded using the *Audio Recording Zone* in the *Test Mode*, and the recorded audio was replayed for participants for self-assessment. Participants were allowed to re-record and were able to proceed regardless of their answers.

After one practice trial, the main task started. Following a fixation cross, presented for 250 ms, the stimulus audio was presented auditorily. After a 2 s pause and a prompt ringtone, the audio recording was started,

accompanied by a microphone picture as a visual prompt. Participants could click "Next" on the screen after recording to proceed, and the recording lasted for a maximum of 10 s with a 3 s countdown. The voice was recorded using the *Audio Recording Zone* in the *Record Mode*, without a replay option. The subsequent trial started automatically. There were three blocks and two breaks, and each block had 12 trials. The order of trials was randomized. The whole task lasted around 10 mins on average.

Classifier knowledge test

After the EI task, participants proceeded to the classifier knowledge test. The test was presented as a one-page questionnaire using the *Questionnaire Builder* feature. Participants first chose between traditional and simplified versions of Chinese characters based on their preference. Then, they were instructed to select the correct classifier from four options built using the Radio Buttons Widget for each question. The test took around 5 mins.

After completion, the first phase finished with the language background questionnaire and a debriefing page.

Picture-naming task

The second phase started with the picture-naming task. Participants went through the same device-checking procedure as in the first phase, turned on the full-screen mode, and choose between traditional and simplified versions of Chinese characters before starting the task.

Participants were instructed to speak a sentence, i.e., "Take me to the task". The response was mandatory and saved as a reference trial for rapidly checking data quality before approving their submissions on Prolific.

Participants were first familiarized with pictures by presenting each picture on the screen with Chinese names underneath. Each picture was displayed for 3,000 ms. After that, participants were instructed to name each picture in a Mandarin NP in the form of a quantifier, a classifier, and a noun (e.g., 一条河 "yi *tiao* he" [one CL river]) as quickly as possible in the practice session. Each picture was superimposed by a string of "XX", and lasted for 3,000 ms. The audio recording was done using the *Audio Recording Zone* in the *Test Mode*, which was initialized on screen start-up,

and automatically replayed after 3,000 ms. Participants were able to hear what they responded and were allowed to make another attempt by clicking the re-recording button until their voice was clear. After completion of each trial, feedback was given in the form of a suggested NP and lasted for 3,000 ms. The feedback was provided as an effort to keep the practice session consistent with the experiment in Chapter 2, where an experimenter provided oral feedback in the lab. The next practice trial started automatically.

The experimental session started with a detailed instruction. Participants were then shown pictures superimposed by Chinese distractor words when ready to start. They were instructed to name each picture using a Chinese NP as fast and clearly as possible. Each trial began with a fixation cross "+" displayed for 400 ms, followed by a 300 ms pause, and the target picture superimposed by a distractor word appeared for 3,000 ms. The audio recording was done using the *Audio Recording Zone* in the *Record Mode*, which was initialized on the picture screen startup. The next trial started automatically. No replay was possible and no feedback was provided.

LexTale-Ch test

Following the picture-naming task, participants proceeded to the LexTale-Ch test. The test started with an instruction on the procedure with example screenshots. Participants were instructed to click on each item if they recognized it as a real Chinese character, even if they may not know its precise meaning. If they thought an item was not a real character, they were asked not to click on it. They were asked to perform the task based on their first impression without any external aid (e.g., consulting a dictionary).

Each item was presented as a button using the *Response Button (Image) Zone*. When pressed, the button transformed into a "✓" symbol, indicating selection. Each item button was configured to allow only one press and remain in the pressed state until the end of the screen. Participants were presented with nine items on each screen and proceeded to the next set of items in a self-paced manner by clicking on the "Next" button. The task lasted around 3 mins on average.

5.3. Results

5.3.1. Behavioral data exclusion

The data were retrieved and downloaded from Gorilla. Among the nine participants who actually participated in the second phase, one participant was excluded due to time outs of their responses. Responses were included in the analysis if they met the following criteria: first, the audios were clear, without obvious distractive environment noise; second, the responses were complete, containing three items in each NP (i.e., quantifier, classifier, and noun); third, the responses were correct. The inclusion threshold for the picture-naming task was set at 60%. As a result, three more participants were excluded for exceeding the threshold. In total, only five datasets were included in the analysis. For the picture-naming task, 20.28% of all data points (i.e., 73 trials) were removed because of (a) participants responding incorrectly or exceeding the time limit, resulting in incomplete responses (19.72%); (b) delayed reaction and outliers (i.e., naming latencies exceeding 3 SDs around a participant's mean reaction time; 0.56%).

5.3.2. Behavioral data analysis

The recorded audio files from the EI test were evaluated by a native Chinese speaker using a rating scale provided by Yan et al. (2020), ranging from 0 to 4, with the following criteria: 0-minimal response, 1-inadequate response, 2-half repetition, 3-minor deviation, 4-exact repetition. The maximum score per participant was 144. Proportionally calculated, the highest correct rate was 100%. The experimenter manually evaluated the scores from the classifier knowledge test and the LexTale test. The highest score was 100.

For the audio files recorded from the picture-naming task, we used Praat (Boersma & Weenink, 2021) to extract naming latencies. Although the sample size was small, we did statistical analyses using the *lme4* package (Bates et al., 2020) in RStudio (Version 2023.03.0.) to examine the potential effects of classifier congruency and semantic relatedness on naming latency. For positively skewed naming latency data, we performed generalized linear mixed effect models (GLMMs) using the *glmer()* function in combination with a gamma distribution and the identity link function to model correct

trials. We generated the maximal model at the beginning based on two main manipulations: *classifier congruency* and *semantic relatedness*. We used treatment coding as our contrast. The default reference levels were *congruent* and *related*. *LexTale-Ch score*, *EI score*, *self-rating score*, *AoA*, *year of use*, and *the order of acquisition* were added to the model as co-variables. We added *participant* and *item* (i.e., the target) as random effects. We followed a top-down selection procedure to simplify the model structure by removing non-significant factors (Barr, 2013; Bates et al., 2014). In the case of non-convergence or singular fit, correlations between random slopes and interaction between fixed effects were discarded. The *anova()* function was used to perform model comparisons. We checked the model fit by plotting the model residuals against the predicted values.

5.3.3. Behavioral data results

Picture-naming task

Table 5.1 reported averaged naming latencies and accuracy for each condition. For naming latencies, we found, first, that participants were faster in the semantically unrelated condition compared to the related condition. Second, participants were faster in the classifier congruent condition compared to the incongruent condition. These findings indicated trends of the semantic interference effect and the classifier congruency effect. Moreover, for naming accuracy, we found that participants were slightly more accurate in the semantically related condition compared to the unrelated condition and more accurate in the classifier congruent condition compared to the incongruent condition.

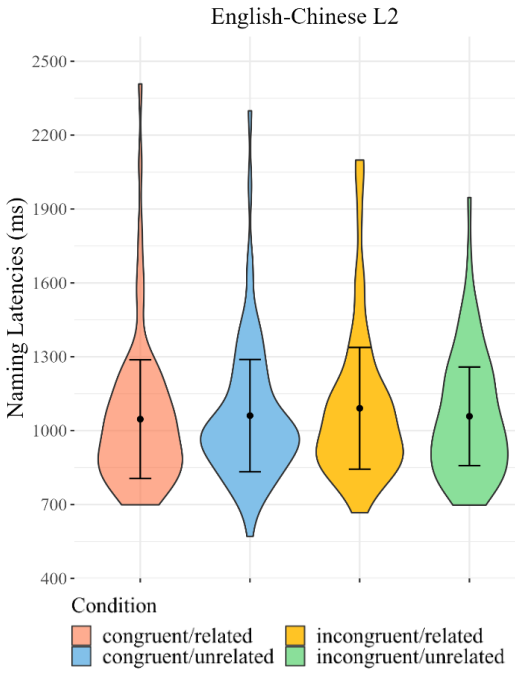
Table 5.1: *Mean naming latencies (only correct responses included) and mean naming accuracy (%) for each condition (n = 5).*

Condition	Latencies (ms)		Accuracy (%)	
	Mean	SD	Mean	SD
congruent-related (C+S+)	1047	301	84.4	14.4
incongruent-related (C-S+)	1091	309	77.8	18.8
congruent-unrelated (C+S-)	1061	285	81.1	15.0
incongruent-unrelated (C-S-)	1059	250	77.8	18.8
Semantic relatedness				
related (S+)	1068	305	81.1	15.8
unrelated (S-)	1059	268	79.4	16.3
estimated difference	**9**		1.7	
Classifier congruency				
congruent (C+)	1054	293	82.8	14.2
incongruent (C-)	1077	280	77.8	17.1
estimated difference	**23**		5.0	

In the best-fitting model for naming latencies, co-variables *LexTale-Ch score*, *EI score*, *self-rating score*, *AoA*, *year of use*, and *the order of acquisition* resulted in non-convergence or singular fit and were dropped from the model fitting procedure. The interaction of *classifier congruency* and *semantic relatedness* did not significantly improve the model fit with $\chi 2(1) = 0.111$, $p = 0.739$, and was thus removed. The best-fitting model was: naming latencies ~ semantic relatedness (related vs. unrelated) + classifier congruency (congruent vs. incongruent) + (1| participant) + (1| item). Statistically, *semantic relatedness* did not significantly affect naming latencies with $\beta = -0.003$, $SE = 0.027$, *95% CI* [-0.056, 0.049], $t = -0.120$, $p = 0.905$. Moreover, *classifier congruency* did not significantly influence naming latencies with $\beta = 0.027$, $SE = 0.027$, *95% CI* [-0.025, 0.080], $t = 1.029$, $p = 0.304$.

In summary, neither *semantic relatedness* nor *classifier congruency* showed a statistically significant effect, primarily due to the limited sample size and insufficient statistical power. Model parameters are reported in Appendix A. See Figure 5.1 to visualize the naming latencies.

Figure 5.1: *Mean picture naming latencies with 95% confidence intervals across the four conditions (n = 5).*



EI test, Classifier knowledge test, and LexTale-Ch

Among the five participants included in the final sample, the average score of the EI test was *M* = 83.47 (*SD* = 16.33). The maximal score was 144 in this adapted version. According to the descriptive statistics of test scores provided by the original authors (Yan et al., 2020), an accuracy rate of around 70% was scaled as intermediate level, and around 90% was scaled as advanced. Therefore, the listening and production of our participants were rated as upper-intermediate levels.

The average score of the LexTale-Ch test was M = 64.50 (SD = 8.41). According to the scoring method by Lemhöfer and Broersma (2012), scores

of 80 to 100 are categorized as “advanced”, scores of 60 to 79 are categorized as “upper intermediate”, and scores of 59 and below are categorized as “lower intermediate”. Our participants’ vocabulary size can be classified as “upper intermediate” level.

The average score on the classifier knowledge test was M = 83.00 (SD = 15.03), with a maximal score of 100. The scores indicate that the five participants have a good knowledge of classifiers.

5.4. Discussion

In this study, we set up an online experiment to explore classifier processing in L2 overt production. We employed a largely identical experimental design to Chapter 2, Experiment 1, to test how classifier congruency and semantic relatedness influenced L2 NP production in English learners of Chinese. In addition to the picture-naming task, we administered a range of proficiency tests to provide a comprehensive evaluation of L2 skills.

We adapted the original lab-based setting to a web-based setting using the Gorilla Experiment Builder to allow the administration in participants’ devices. The transition to the web environment resulted in some modifications of the experimental procedure. First, this study employed a longitudinal design consisting of two phases, i.e., a proficiency test phase and a main experiment phase. The proficiency test phase assessed speaking and comprehension skills, as well as classifier knowledge. It aimed to ensure that participants invited to the picture-naming task had a certain proficiency level. Second, participants were required to self-assess their devices for microphone and browser sound when accessing the study link in each phase. The testing was taken to ensure the consistency of device settings. The microphone checking was performed again at the beginning of the picture-naming task, and reference audio was recorded for later assessment. Third, in the absence of an experimenter, participants were instructed to self-assess their verbal responses by replaying the recorded audio during the practice session of both the EI test and the picture-naming task.

The final sample size was limited, resulting in insufficient statistical power to detect significant effects in the analysis. Nevertheless, we found a

descriptive trend for the classifier congruency effect in the picture-naming task, with faster naming when a picture was superimposed by a congruent distractor word compared to an incongruent word. The estimated difference in naming latencies between the congruent and incongruent conditions was approximately 22 ms. This trend appeared to replicate the classifier congruency effect observed in L1 and L2 NP production in lab settings (Huang & Schiller, 2021; Wang & Schiller, submitted). Moreover, we found a trend for the semantic interference effect, with faster naming in the semantically unrelated condition compared to the related condition. These findings indicated that the naming latencies of our L2 participants were affected by both classifier congruency and semantic relatedness in NP production. Based on these trends one may speculate that both classifier congruency and semantic interference effects may become statistically significant with sufficient sample size and statistical power.

However, it is worth noting that among the final sample of L1 English speakers (n = 5), all of them had been learning Mandarin Chinese for several years, and three of them had at least one parent from Guangdong province, China, who spoke Cantonese. This indicates that they were likely exposed to Cantonese, in addition to English, at home from an early age, and such a linguistic environment could have potentially facilitated their Mandarin learning compared to English-speaking families without exposure to Cantonese at all. The other two participants did not provide reports on their parents' language background; however, they had an average learning duration of 33.5 years (*SD* = 13.5), which suggests they had significant experience in learning Mandarin. Therefore, it remains an open question whether English learners of Mandarin Chinese with intermediate proficiency levels, who were not exposed to Cantonese from an early age, would demonstrate sensitivity to classifier congruency in NP production.

Regarding the recruitment of L2 speakers in the web-based setting, one issue we encountered is the high exclusion rate because the majority of participants dropped out or made too many errors (with correct rates below 60%) in the proficiency tests. Surprisingly, a large number of participants who registered for the first phase had limited Chinese language knowledge or appeared to be unable to speak Chinese. This finding was unexpected, considering that the language requirement had been clearly emphasized in

the title, study information, and requirement sections of our study description on the Prolific platform. Moreover, the language requirement had been specified in the Prolific built-in pre-screening criteria to ensure that only individuals meeting the criteria would be invited.

We presumed that, first, many participants did not thoroughly read the study description before signing up, leading to the language requirement being overlooked. Given that this study focused on L2 speech production, the language requirement stood as the most critical criterion. Unfortunately, the lack of attention to this requirement resulted in a high drop-out rate, and the recorded data became unusable. Second, the Prolific built-in pre-screening criteria proved insufficient for a study in L2 linguistics that demands specific proficiency levels. This limitation was reflected in the inability to set a precise L2 proficiency level as a criterion, relying instead on participants' self-reported, rough, and subjective language proficiency during the registration process on Prolific. Participants who self-reported as "fluent" in the Chinese language appeared to have exaggerated their actual proficiency levels. This was evident in their performance in the EI test, where some participants were silent, produced unintelligible repetitions, or repeated only a single word, despite reporting fluency in Chinese. Up to the date of data collection in the current study, there was no effective method available on Prolific to directly target language learners with specific proficiency levels (e.g., intermediate).

Moreover, we presumed that the single-anonymized mode also contributed to the high error rate in the L2 study. In this mode, the experimenter was unaware of the participants' identities and had no direct contact with them. Consequently, the absence of experimenter monitoring required participants to take a more independent and self-directed approach to the tasks. This may have resulted in a lack of guidance, leading to a higher likelihood of errors and drop-outs, especially among participants with lower target language proficiency levels. To ensure that participants clearly understood what they needed to do, we made careful efforts to improve the clarity of instructions by providing detailed explanations, highlighting key points, inserting example screenshots, and using visual prompts such as pictures and arrows. Despite these efforts, it was unavoidable that some participants remained uncertain about the task, as reflected in instances of

hesitation, questions, or discussions with others captured and recorded in the audio. In addition, we also encountered "bad-faith" participants who left the task running for payment, resulting in audio recordings capturing only their environmental sounds, such as TV programs, birds, or family dinners.

In order to make recruitment and data collection more efficient, we adjusted the recruitment strategy to recruit L2 speakers through the participants' pool of the FSW at Leiden University instead of Prolific, as these participants might be more familiar with speech production experiments similar to the current study. We also attempted to reach out to Chinese language institutions in English-speaking countries such as the United Kingdom, the United States, and Australia, as they potentially had a larger pool of participants who met the L2 requirement. However, despite these efforts, the sample size could not be expanded in the current study. Nevertheless, online experiments provide more flexibility and a broader scope for recruitment possibilities.

In sum, conducting L2 overt production experiments in web-based settings poses challenges in terms of recruiting L2 learners with specific proficiency levels (e.g., intermediate and higher levels) and consistent language backgrounds, as well as collecting data without the experimenter operation. The current experiment attempted to address these challenges and improve the effectiveness by dividing the proficiency tests, which were originally conducted concurrently with the picture-naming task in the lab in Chapter 2, into a separate phase as a participant pre-screening procedure.

From a technical perspective, ensuring the reliability of the timing of audio recordings is crucial. One issue that requires particular attention is the synchronization of the actual recordings with the onset of the stimulus. We observed that there may be a delay between Gorilla initiating the request for microphone access to start recording and the actual data captured by the microphone. This issue could cause the loss of the beginning of recorded audio. However, the length of delay was unable to be measured due to it varying between devices. To ensure sufficient time for the microphone to start up, we followed Gorilla's recommendation to add a 1,000 ms delay in stimulus onset using a provided script. The script was applied to both practice and experimental sessions in the picture-naming task. As a result,

each stimulus screen had a total duration of 4,000 ms, with a 1,000 ms delay before the stimulus appeared and a 3,000 ms duration for the stimulus itself. The request for microphone activation was initiated at the start of the screen and lasted until the end of the screen, resulting in each recorded audio having a duration of around 4,000 ms. During data analysis, the naming latencies extracted from the audio files were adjusted by subtracting this 1,000 ms delay prior to the onset of voice.

However, the timing issue was still reflected in varying audio lengths across participants, which has also been reported by Vogt et al. (2021) and Stark et al. (2022). Vogt et al. (2021) conducted a study to test the online PWI task (Experiment 1) using the experimental platform SoSciSurvey (Leiner, 2019) and the experiment library jsPsych (De Leeuw, 2015). They found that only a small number of audio files matched the expected length precisely. The variation in audio lengths is assumed to be caused by technical factors such as device variability and fluctuation in internet connection. The variation can occur either at the beginning or the end of the recording, and it is not possible to precisely measure the variance.

The technical aspects related to the recording process are outside the scope of this study (Vogt et al., 2021). We have taken measures to improve the synchronization of recordings. Moreover, our audio data showed consistent audio length within participants. This suggests that the audio recordings in Gorilla were relatively less susceptible to variability within participants' devices. Furthermore, Stark et al. (2022) reported that length variation did not affect the effect in their study involving thirty L1 speakers. This suggests that with adequate sample size and statistical power, errors induced by such variation may be minimized.

Based on the above discussion, three main recommendations for conducting L2 PWI tasks online were presented as follows: First, it is crucial to carefully pre-screen participants based on their L2 proficiency levels and adherence to the task. Adding a separate proficiency test phase before the main experiment was effective in pre-screening participants' proficiency levels. However, the longitudinal design also increased the possibilities of participant attrition, where some participants who completed and passed the first phase did not return for the second phase.

Second, clear instructions help participants understand the task requirements more accurately. In online settings, instructions should be much more precise and easily readable, providing detailed information about the visual or auditory stimuli participants will encounter and the expected corresponding responses.

Third, for the language production experiment that focuses on the timing of voice onset, maintaining the reliability of audio data is essential. Special attention should be given to the functionality of the microphone on participants' devices and the accurate timing of the recording onset relative to the stimulus onset. These factors play a crucial role in ensuring the reliability and validity of the recorded audio data. Variation caused by technical and internet factors exists across participants' devices and cannot be completely eliminated, but can be minimized by employing a reliable experiment builder, such as Gorilla and jsPsych (for discussion, see Anwyl-Irvine et al., 2021), and a large sample size.

The lab environment offers better control over the L2 overt production experiment, providing several advantages. First, lab equipment ensures the consistency of stimulus presentation and audio recordings, and technical issues can be promptly addressed by the experimenter. Second, the involvement of the experimenter is valuable, as they can provide oral feedback, detailed explanations, and timely answers to questions, ensuring active engagement from every participant. Moreover, the lab setting reduces the probability of casual or irresponsible participation. Participants are directly contacted by the experimenter, scheduled for an appointment, and required to physically travel to a designated experiment location, compared to simply accessing the experiment via clicking on a web link in online settings.

Nevertheless, the advantages of an online setting are significant in terms of flexibility, especially for certain types of participant groups where it is not easy to recruit the target sample size within a limited period at the same site where the lab is located. Online settings are not restricted by geography, allowing data collection from anywhere in the world without needing travel. In this study, our final sample, although limited, included participants from four countries (i.e., the United States, the United Kingdom, Canada, and

Australia). This approach saved significant time, manpower, and financial resources compared to the alternative of traveling to each of these countries to collect data in a lab setting.

Taken together, the findings of this online L2 overt production experiment suggest that classifier congruency and semantic relatedness appear to impact naming latencies in L2 NP production for L1 English speakers, particularly those at the upper-intermediate proficiency level and who were exposed to Cantonese from an early age. Due to the challenges of running the L2 overt experiment online, we had a small final sample size and were unable to gain a comprehensive understanding of classifier processing in this specific group. Nevertheless, our attempt provides valuable insights and recommendations for future web-based empirical studies in L2 production.

5.5. Conclusion

This study explored the classifier congruency effect and the semantic interference effect in L2 speech production of English-Mandarin Chinese speakers using the PWI paradigm. In the picture-naming task, target names were overtly articulated in the form of NPs consisting of quantifier, classifier, and noun. The trends for classifier congruency and semantic interference effects in naming latencies were observed, indicating that our L2 participants appeared to be sensitive to classifiers and semantics.

Our study was a valuable endeavor to implement a reaction-time-sensitive, overt L2 production experiment in a web-based setting. It added evidence for the feasibility of capturing millisecond differences in experimental tasks using the Gorilla Experiment Builder, and it supported the flexibility of collecting data from participants across diverse geographical locations. We have provided recommendations for studying L2 production online, highlighting the recruitment and technical considerations. We hope these suggestions can help future researchers effectively conduct web-based experiments in L2 production.

Declaration of competing interest

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported.

Contribution statement

Shaoyu Wang: Conceptualization, Methodology, Programming, Data Acquisition, Formal Analysis, Data Curation, Writing-Original Draft, Writing-Review and Editing, Funding acquisition. **Niels O. Schiller**: Conceptualization, Methodology, Writing-Review and Editing, Supervision, Funding acquisition.

Appendix

5.A Model parameters: naming latencies

Table 5.A.1: *Specification of the model of best fit for naming latencies (n = 5), including estimated means, standard error, confidence intervals and t-values.*

Formula: naming latency ~ semantic (related vs. unrelated) + classifier (congruent vs. incongruent) + (1| participant) + (1| item)					
Predictors	*Estimates*	*std. Error*	*CI*	*Statistic*	*p*
(Intercept)	1.074	0.072	0.934 – 1.215	15.008	<0.001
Classifier [incongruent]	0.027	0.027	-0.025 – 0.080	1.029	0.304
Semantic [unrelated]	-0.003	0.027	-0.056 – 0.049	-0.120	0.905
Random Effects					
σ^2	0.06				
$\tau_{00\ Item}$	0.01				
$\tau_{00\ Participant}$	0.00				
ICC	0.14				
$N_{\ Participant}$	5				
$N_{\ Item}$	18				
Observations	287				
Marginal R^2 / Conditional R^2	0.003 / 0.147				