The role of lexico-syntactic features in noun phrase production and comprehension: insights from Spanish and Chinese in unilingual and bilingual contexts

Wu, R.

Chapter 4

Classifier Congruency Effects in Chinese Noun Phrase Production: Behavioral and Electrophysiological Evidence from Spanish−Chinese Bilinguals

---

A version of this chapter is under review: Wu, R., Parafita Couto, M. C., & Schiller, N. O. (under review). Classifier congruency effects in Chinese noun phrase production: behavioral and electrophysiological evidence from Spanish−Chinese bilinguals.

102 *The Role of Lexico-Syntactic Features in Noun Phrase Production and Comprehension*

Abstract

Mandarin Chinese employs a classifier system requiring classifiers obligatorily positioned between demonstratives or numerals and nouns in noun phrases (NPs; e.g., *一张桌子*, *[one + specific-classifier-zhang1 + table]*, "one table"). While classifier congruency effects have been documented in monolingual Chinese speakers, it remains unclear whether such effects extend to early Spanish−Chinese bilinguals. This study examined classifier congruency effects in NP production by 29 early Spanish−Chinese bilinguals and 30 Mandarin Chinese speakers using behavioral and EEG measures. Participants completed picture-word interference (PWI) tasks with four conditions, manipulating classifier congruency (congruent vs. incongruent) and semantic relatedness (related vs. unrelated). Both groups showed significantly longer naming latencies for classifier-incongruent and semantically related conditions. ERP results revealed N400-like effects for classifier incongruency in both groups and semantic interference, reflected by N400-like effects in bilinguals and P600-like effects in Mandarin Chinese speakers. These findings indicate competitive selection of classifiers during NP production across both groups, highlighting the robustness of classifier congruency effects and the flexibility of morphosynactic processing in bilinguals.

Keywords: language production; classifier congruency effect; semantic interference effect; bilingualism; picture-word interference (PWI); N400; P600

4.1 Introduction

Speech production in monolinguals involves a complex, multi-staged process, but how does this process differ for bilinguals? Theories of bilingual lexical access suggest a general three-stage process, regardless of whether the lexical selection mechanism is language-specific or language-non-specific (e.g., Costa et al., 1999a; De Bot, 1992; De Bot & Schreuder, 1993; Dijkstra & Van Heuven, 1998; Green, 1986; Roelofs, 1998). First, lexical entries from both the active (language-in-use) and non-active (language-not-in-use) languages are activated in parallel. Second, a lexicon-external mechanism or production rule determines which language will be used. Third, lexical entries from the target language receive the highest activation and are ultimately selected for production, while those in the non-target language are suppressed.

A widely used method to investigate these processes is the picture-word interference (PWI) paradigm (e.g., Rosinski et al., 1975). In the PWI paradigm, participants name a picture while ignoring a distractor word presented before, simultaneously with, or after the picture onset. The LRM speech production model (Levelt et al., 1999) provides a framework for understanding the process of naming pictures in noun phrase (NP) production. According to this model, both the target picture and the distractor activate their conceptual and lexical representations first. Next, their syntactic features, such as grammatical gender, classifiers, or number, are retrieved. All related nodes are then activated, with the target-related node receiving the highest activation. Finally, the most activated node is selected, encoded, and articulated as part of the NP.

While monolinguals consistently follow this process, NP production in bilinguals involves additional layers of complexity. During bilingual production, syntactic representations and conceptually related nodes of both the target and distractor from target and non-target languages are activated. However, nodes in the target language receive higher activation, and the target-related node in the target language receives the highest activation. This is because an external production rule determines which language is intended to be produced. Thus, selection is restricted to nodes within the target language, with the highest activated node ultimately selected for production (e.g., Costa et al., 1999a; Roelofs, 1998).

Recent research has increasingly examined the processing of classifiers in bare noun and NP productions (e.g., Wang et al., 2019; Huang & Schiller, 2021; Wang & Schiller, submitted; Wang et al., submitted) and established the mechanisms underlying these processes in monolinguals. However, little is known about how bilinguals produce classifiers. Thus, this raises a critical question: does classifier production in bilingual NP processing involve competition and selection mechanisms similar to those observed in monolinguals?

4.1.1 The effects of processing lexico-syntactic features in language production

4.1.1.1 The gender congruency effect

Experimental research on speech production has mostly investigated lexico-syntactic features, such as grammatical gender and classifiers, using the PWI paradigm (Glaser, 1992; Schriefers, 1993). In this paradigm, it has been found that the relationship between distractor words and target pictures significantly affects picture-naming latencies. For example, Schriefers (1993) initially used the PWI paradigm to investigate the selection of grammatical gender in NP production in Dutch (a language that features common and neuter gender). Faster naming latencies were obtained when the grammatical gender of the target (*boek*$_{NEU}$, "book") was congruent with the distractor word (*dak*$_{NEU}$, "roof") compared to the incongruent one (*tafel*$_{COM}$, "table"). Moreover, these faster naming latencies in gender-congruent conditions than gender-incongruent conditions were continuously observed in Germanic languages (e.g., German in Bürki et al., 2016; Schiller & Caramazza, 2003 and Schriefers & Teruel, 2000; Dutch in Schiller, 2013 and Schiller & Caramazza, 2006) and Romance languages (e.g., Portuguese in Sá Leite et al., 2021; French in Alario & Caramazza, 2002; Spanish in Costa et al., 1999b and Wu & Schiller, 2023), as well as other gender-marking languages (see Wang & Schiller, 2019 and Sá Leite et al., 2022 for an overview). Overall, this faster and more accurate processing in gender-congruent conditions is known as the *gender congruency effect*.

4.1.1.2 The classifier congruency effect

More recently, the PWI paradigm has been employed to investigate classifier processing in Mandarin Chinese (hereafter, "Chinese" refers to Mandarin Chinese). For example, Wang et al. (2006) examined the selection of classifiers in Chinese bare noun and NP production using the PWI paradigm with Chinese monolingual speakers. Shorter naming latencies were observed when the classifier of target pictures was congruent with distractor words in NP production, but not in bare noun production. A similar effect in naming latencies was also observed in Chinese NP productions by Huang and Schiller (2021), where the classifier congruency (congruent vs. incongruent) and semantic relatedness (related vs. unrelated) between target pictures and distractor words were manipulated. Monolingual Chinese speakers named pictures while simultaneously ignoring presented distractor words, with their naming latencies and electrophysiological (EEG) activity being recorded simultaneously. Behavioral results showed that classifier-congruent distractors elicited faster naming latencies, while semantically related distractors increased naming latencies. EEG results showed a significant N400 effect modulated by classifier incongruency and semantically unrelatedness. This finding aligns with Wang et al. (2019) who observed a comparable N400 effect in Chinese bare noun production using similar experimental designs. They interpreted this effect as evidence of classifier activation during lexical selection. Given the similar inherent properties and processing mechanisms, researchers have identified the observed effects in naming latencies and electrophysiological responses during Chinese NP production, as well as the electrophysiological responses during Chinese bare noun production, as manifestations of the *classifier congruency effect*.

These behavioral and EEG findings from bare nouns and NP production in Chinese monolingual speakers confirm the robustness of the classifier congruency effect in Chinese speech production. However, since no studies have examined how bilinguals (e.g., Spanish−Chinese bilinguals) process classifiers in Chinese NP production, it remains unclear whether similar effects would be observed in Spanish−Chinese bilinguals. Moreover, given that Chinese classifiers and Spanish grammatical gender function as lexico-syntactic features with similar characteristics in NPs and comparable processing mechanisms, it is particularly interesting to investigate how Spanish−Chinese bilinguals process, select, and produce classifiers in their speech.

4.1.2 The semantic interference effect

A well-established effect using the PWI paradigm is the *semantic interference effect*. Greater interference occurs when distractor words are semantically related to target pictures (e.g., PIG, lion), compared to unrelated distractors (e.g., PIG, cup) or nonsense trigrams (Rosinski, 1977). This results in longer naming latencies in semantically related than unrelated conditions. According to the lexical selection by competition theory (Roelofs, 1992, 1993; Levelt et al., 1999), when the target picture (e.g., *pig*) and the distractor (e.g., *lion*) are semantically related, the corresponding lemmas activate each other through the spreading of activation to representations in the lexical network. This semantically related distractor (unlike unrelated distractors) receives activation from the semantic category node activated by the target. This results in more competition during the lexical selection process and delays the selection of the target lexical node, producing longer naming latencies. In the unrelated condition, the target and the distractor are not activating each other, resulting in relatively less competition and faster naming latencies.

4.1.3 Electrophysiological correlates of classifier congruency and semantic interference effects

Electroencephalography (EEG) and event-related potentials (ERPs) are particularly suitable methodologies for monitoring the complex processing of speakers' NP production. In previous studies, two ERP components seem to have been especially in focus, i.e., the N400 and the P600. The N400 is an ERP component characterized by a negative voltage amplitude peak around 400 ms post-stimulus onset, typically within the 250 ms and 600 ms time window over central and posterior electrode sites (Kutas & Hillyard, 1980; Chwilla et al., 1995). It is typically associated with lexical-semantic integration and lexical co-activation processes (e.g., Lau et al., 2008; Kutas & Federmeier, 2011; Chen et al., 2017; Leckey & Federmeier, 2019). Previous studies on the gender congruency effect and the classifier congruency effect have frequently reported modulations of the N400. Specifically, more negative N400 amplitudes were mostly reported as elicited by gender- or classifier-incongruent trials compared to gender- or classifier-congruent trials (see an overview of gender congruency effects in Wang & Schiller, 2019 and Sá Leite et al., 2022; classifier congruency effects in Wang et al., 2019 and Huang & Schiller, 2021). Additionally, the N400 was also reported for picture-naming tasks in semantically unrelated conditions relative to related conditions, reflecting semantic integration processes (e.g., Greenham et al., 2000; Blackford et al., 2012; Wang et al., 2019; Huang & Schiller, 2021). It was concluded that the N400 effect is elicited by competition at the lemma level, triggered by strong activation at the conceptual level (Wang et al., 2019). Thus, in this study, we used the N400 to detect the co-activation of the classifier congruency effect and lexico-semantic integration of the semantic interference effect.

The P600 is an ERP component defined by a positive-going deflection primarily localized in the centroparietal regions, typically occurring between 500 and 800 ms post-stimulus onset and peaking around 600 ms (Osterhout & Holcomb, 1992). Typically, the P600 is sensitive to syntactic violations in syntactic processing (Friederici et al., 1993; Hagoort et al., 1993). However, the P600 is not only necessarily associated with syntactic violations, rather, it has also been reported to be elicited by orthographic violations (Münte et al., 1998), garden path sentences (Friederici et al., 1996), grammatical violations (Hagoort et al., 1993), following N400 effects of semantic violations (Osterhout & Nicol, 1999), generally improbable events (Coulson et al., 1998), and semantic anomalies involving implausible or unexpected events (e.g., **The cat that fled from the mice ran through the room*; Kolk et al., 2003). Furthermore, the P600 was challenged in the view of an exclusively syntactic violation response, for instance, Kim and Osterhout (2005) proposed a "semantic P600" effect, typically distributed in the centro–parietal regions with a maximal difference occurring within 600 – 800 ms (also see Kuperberg et al., 2003; Van Herten et al., 2005; Bornkessel-Schlesewsky & Schlesewsky, 2008, for a review). This perspective redefines the P600 as an electrophysiological response to "integration difficulties", reflecting the cognitive effort required to semantically, rather than syntactically, integrate an entity into the discourse context. Delogu et al. (2019) provided support for this view by observing a P600-like effect when words, though semantically related to the context, were difficult to integrate into the syntactically correct discourses (i.e., implausible events). Additionally, Puhacheuskaya (2021) also pointed out that the regions and time windows of the "semantic P600" effect are largely consistent with the centro–parietal P600 effect, which is primarily documented by most studies as the effect associated with syntactic reanalysis and repair. Given the view of the "semantic P600" effect, we also consider the P600 to detect the semantic interference effect.

4.1.4 The feature of classifiers and NPs in Chinese

Almost all languages incorporate some form of nominal classification in their grammar, with grammatical gender systems (e.g., in Spanish) and numeral classification systems (e.g., in Chinese) being the most common (Seifart, 2010). In Spanish, nouns inherently possess an identifiable grammatical gender, classified as either masculine or feminine (Roca, 1989;

Harris, 1991). Determiners and adjectives within Spanish NPs, while not inherently gender-marked themselves, must obligatorily agree in gender with nouns they modify (Harris, 1991). Thus, the grammatical gender choice within NPs primarily depends on the properties of nouns. In contrast, Chinese does not feature a grammatical gender classification system but instead employs a classifier system that is in some way comparable to gender division (Wang et al., 2019). Classifiers, as free morphemes, classify the category of nouns they accompany, with their selection determined by particular properties of nouns (Li & Thompson, 1981). In Chinese NPs, determiners, numerals, or quantifiers cannot directly modify nouns. Instead, a numeral classifier is obligatorily inserted as a mediating element (Li, 2013). Thus, a typical Chinese NP consists of demonstrative/numeral /quantifier + classifier + noun (e.g., *一只老虎*, /yi1[5]zhi1lao3hu3/ *[one classifier-zhi1 tiger]*, "one tiger") (Li & Thompson, 1981). In general, classifiers function similarly to grammatical gender in nominal systems, as they represent inherent properties of nouns and cannot be omitted in NPs.

In Mandarin Chinese, there are several dozen classifiers with approximately 150 commonly used (Erbaugh, 1986). According to Erbaugh (2006), these classifiers can be generally classified into five categories based on their lexical type and function: (1) measure classifiers, used for measuring weight or length (e.g., *公里* /gong1li3/ "kilometer", *两* /liang3/ "ounce"); (2) kind classifiers that describe categories or types of entities (e.g., *种* /zhong3/ "type", *类* /lei4/ "kind, category"); (3) collective classifiers, indicating arrangements of entities (e.g., *排* /pai2/ "row of" or *捆* /kun3/ "bundle"); (4) event classifiers that are used for events (e.g.,*班* /ban1/ "run of a bus or train", *场* /chang3/ "performance of a show"); (5) sortal classifiers, used for particular categories based on prope-

[5] In Mandarin Chinese, tones are essential for phonemic distinction. The language has four lexical tones: Tone 1 is high-level, Tone 2 is high-rise, Tone 3 is low-dip, and Tone 4 is high-falls (Chao, 1948). In this study, numerical markers indicate the corresponding tone in Pinyin transcriptions.

rties of entities (e.g., 棵 /ke1/ "classifier for plants", 块 /kuai4/ "chunk or square"). Mandarin Chinese has approximately 75 sortal classifiers, each typically used with 5 to 20 nouns (Erbaugh, 2002). Typically, classifier selection is determined by properties of nouns, such as animacy, shape, function, and size (Myers & Tsay, 2000). It should be noted that the relationship between nouns and classifiers is, in many cases, opaque, although choices of classifiers in NPs depend to a great extent on semantic properties (Shao, 1993; Tzeng et al., 1991). Additionally, there is a special so-called "general" or "default" classifier *ge* (个 /ge4/), which is the most frequently used classifier (Li & Thompson, 1981), with approximately 40% of Chinese nouns only taking this classifier, such as unique objects (e.g., "the sun") and abstractions (e.g., "dream", "idea") (Erbaugh, 2006). Moreover, compared to other classifier categories, Mandarin Chinese speakers often prefer sortal classifiers when referring to unfamiliar nouns or those with new information (Erbaugh, 2006). Therefore, both the general classifier and sortal classifiers were considered as the primary target classifiers in this study.

4.1.5 The current study

This study explores the processing of classifiers in Chinese NPs from behavioral and neural perspectives, focusing on Mandarin Chinese speakers who were born and grew up in China (hereafter "Mandarin Chinese speakers" refer to Chinese speakers born and raised in monolingual Mandarin-speaking environments in China) and early Spanish−Chinese bilingual speakers in Spain. In this study, we manipulate classifier congruency (classifier-congruent vs. classifier-incongruent) and semantic relatedness (semantically related vs. semantically unrelated) between target objects and distractor words. The study serves the main objective of examining whether classifier features are automatically activated and competitively selected during Chinese NP production in bilingual speakers. Specifically, we aim to determine whether this activation is evident in specific ERP components (e.g., N400) and whether competitive selections influence naming latencies, thereby resulting in a classifier congruency effect in both behavioral and neural measures. This leads to our primary research question: is there a classifier congruency effect in Chinese NP production by early Spanish−Chinese bilinguals?

4.1.5.1 Hypotheses

Behavioral hypotheses. We predict effects on classifier congruency and semantic interference on behavioral measures of naming latencies in both Mandarin Chinese speakers and early Spanish−Chinese bilingual speakers. Based on the lexical selection by competition theory (Roelofs, 1992, 1993; Levelt et al., 1999), lexical selections involve competitive processes among activated lemmas. Accordingly, for classifier-congruent targets and distractors, we predict a *classifier congruency effect*, reflected in less competition and shorter naming latencies in classifier-congruent compared to incongruent conditions.

For targets that are semantically related to distractors (i.e., belonging to the same semantic category), we expect a *semantic interference effect*, due to more competition and thus longer naming latencies for semantically related compared to unrelated conditions. However, according to the LRM (Levelt et al., 1999), lexico-semantic and lexico-syntactic information are independently activated and selected at the lemma level, with each following a distinct competitive selection process. Additionally, given that the mapping between classifiers and semantic features of nouns is often opaque, classifier selection is primarily driven by individual nouns rather than broad semantic categories (Shao, 1993; Tzeng et al., 1991). Consequently, no interaction is expected between the classifier congruency effect and the semantic interference effect.

EEG hypotheses. We predict a more negative N400 amplitude for classifier-incongruent conditions compared to congruent conditions between 250 – 600 ms after picture onset in Mandarin Chinese speakers. Given the similarities in processing mechanisms of both monolingual and bilingual language production, we also expect a similar N400 effect in early Spanish−Chinese bilinguals under classifier-incongruent conditions.

For the semantic interference effect, we similarly expect, based on Wang et al. (2019) and Huang and Schiller (2021), a more negative N400 amplitude elicited by semantically unrelated conditions compared to related conditions in two groups. However, considering prior research identifying the "semantic P600" effect as a response to "semantic integration difficulties", it is plausible that semantically unrelated conditions may also elicit a "semantic P600" effect, reflecting increased integration challenges when target pictures are semantically unrelated to distractors.

4.2 Methods

4.2.1 Participants

This study involved two groups of participants: Mandarin Chinese speakers and early Spanish−Chinese bilinguals. The Mandarin Chinese group consisted of 30 healthy, right-handed native Chinese speakers (21 females) with a mean age of 24.36 years (SD = 3.02), recruited from Pompeu Fabra University in Barcelona, Spain. All speakers in this group were born and raised in China (see details below). Participants in this group completed the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) prior to the experiments to assess their linguistic profiles, including language experience and usage. The Spanish−Chinese bilingual group included 29 healthy, right-handed early bilinguals (24 females) with a mean age of 21.01 years (SD = 3.01), also recruited at Pompeu Fabra University. These participants were required to complete a revised BCSP questionnaire developed based on the Bilingual Code-Switching Profile (BCSP; Olson, 2022), a Chinese Elicited Imitation (EI) test (Yan et al., 2020) and a short version of the Diplomas de Español como Lengua Extranjera (DELE) Spanish proficiency test (Instituto Cervantes, 2010) to evaluate their language proficiency, experience, and usage. All participants reported having no language disorders, psychological conditions, or visual impairments. The informed consent was filled out before the experiments, and compensation was provided upon completion of the experiments.

4.2.1.1 LEAP-Q: Mandarin Chinese group

Within the Mandarin Chinese group, all speakers were born and raised in monolingual Mandarin Chinese-speaking environments in China, ensuring native-level proficiency and natural exposure to Mandarin Chinese from birth. Of these, twenty-two participants were pursuing master's degrees, while the remaining were enrolled in bachelor's (n = 3) or doctoral programs (n = 5) in Barcelona, Spain. Although our initial efforts were to recruit monolingual Chinese speakers without any knowledge of Spanish or other languages (e.g., English), fourteen participants with Spanish proficiency below the B1 level were ultimately included, due to Spanish universities' requirements for master's students in Spain to attain basic Spanish proficiency. Additionally, given the global prevalence of English and its integration into higher education in China and abroad, all participants inevitably had an average English proficiency of approximately the B2 level. On average, participants reported using Chinese for approximately 70.18% (SD = 0.18) of their daily language use. Detailed linguistic profiles are provided in Appendix 4.A1.

4.2.1.2 Revised BCSP: early Spanish−Chinese bilingual group

All early Spanish−Chinese bilingual participants were either born in Spain (n = 25) or China (n = 4), and all were raised and currently reside in Barcelona, Spain. They grew up in households where Chinese served as the primary language spoken by their parents, acquiring it as a heritage language. Spanish, however, is the dominant language in their broader social and educational environments and is extensively used in daily life. Most participants learned Chinese in informal contexts, primarily at home or through heritage language schools (e.g., weekend or community Chinese language schools), where formal exposure to Chinese was typically limited to a few hours per week. The majority of participants (n = 22) were pursuing bachelor's degrees, while the remaining participants were either continuing master's studies (n = 4) or had recently graduated from high school (n = 3). On average, participants reported using Chinese for 32.55% (*SD* = 0.17) and Spanish for 44.87% (*SD* = 0.16) of their daily communication. All participants reported Chinese and Spanish as their native languages. Additionally, sixteen participants also identified Catalan as their native language, while the remaining participants either learned Catalan as a second language (L2) (n = 11) or reported no proficiency in Catalan (n = 2). Further details of the participants' linguistic profiles are provided in Appendix 4.A2.

4.2.2 Materials and design

The Mandarin Chinese group completed the picture naming task only, while the early Spanish−Chinese bilingual group was requested to first complete the picture naming task, then the DELE and EI tests to assess their proficiency in Spanish and Chinese.

4.2.2.1 Tasks and Stimuli

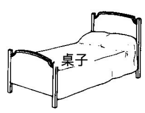
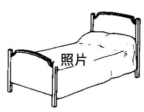
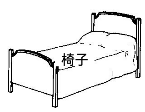
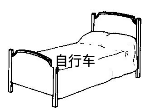
Picture naming task

Twenty-one black and white line drawings were obtained from Severens' picture database (Severens et al., 2005). The picture stimuli were selected based on two criteria: pictures should have easily recognizable features and represent concrete concepts with exact Chinese names. Each picture was associated with four distractor words, manipulated according to their classifier congruency (classifier-congruent vs. classifier-incongruent) and semantic relatedness type (semantically related vs. semantically unrelated) with target pictures. This yielded a total of 84 tokens, each consisting of one target picture and one distractor (see Appendix 4.B). The visual complexity of distractors ($F(3,80) = 0.466$, $p = 0.707$) and their stroke counts ($F(3,80) = 0.695$, $p = 0.558$) were balanced across four conditions. Additionally, the frequency of distractors, $F(3,80) = 0.228$, $p =$

0.877, based on the BCC corpus (Xun et al., 2016), was controlled to ensure uniformity across conditions. Distractors and target pictures were designed to avoid any phonological and orthographic similarities.

The experiment followed a fully factorial within-subjects design with two primary factors: classifier congruency (C) and semantic relatedness (S), resulting in a 2 by 2 design. The factor classifier congruency was manipulated based on whether the classifiers of targets and distractors were congruent (C+) or incongruent (C−). Similarly, the factor semantic relatedness distinguished between semantically related (S+) and semantically unrelated (S−) conditions, depending on whether or not targets and distractors belonged to the same semantic category. This design resulted in four experimental conditions for each target picture: C+S+, C+S−, C−S+, and C−S− (see Table 4.2.1).

Table 4.2.1. A sample set of target pictures paired with distractors for the picture naming task.

	Condition			
Target picture **BED** [床/chuang2/]	C+S+	C+S−	C−S+	C−S−
Target classifier 张 /zhang1/				
Distractors	table /zhuo1zi0/	photo /zhao4pian4/	chair /yi3zi0/	bicycle /zi4xing2che1/
Classifier of distractors	张 /zhang1/	张 /zhang1/	把 /ba3/	辆 /liang4/

Chinese Elicited Imitation (EI) test

The EI test of Chinese was programmed in E-prime 2.0 (Schneider et al., 2002), with identical instructions and stimuli used in Yan et al. (2020). This task aimed to assess Chinese as a second/foreign language (L2/FL) learners' general language proficiency, vocabulary, and grammatical knowledge (Yan et al., 2020). It consists of three sets of stimulus sentences, with each set including 24 sentences (k = 24*3 = 72). These three sets are targeted as follows: the first targets the beginning level, and the second and third sets target the intermediate level and advanced level, respectively. Each sentence consists of key vocabulary and/or grammatical knowledge that can be used for diagnostic assessment (Yan et al., 2020). For example, in (1), (2), and (3), the underlined words in the sentences are key vocabulary, and words in bold are target grammar. The scoring criteria are as follows: no points will be marked for minimal responses, such as complete silence, only function words, or one-word repetition; one point will be awarded for inadequate answers that demonstrate some understanding of the stimulus, such as half-sentence repetition with major grammatical errors, or few content words repetition but changing the main idea of the stimulus; two points will be marked for half repetition that includes more than half of sentences with more than one phrase containing grammatical mistakes; three points for repetition with minor deviation, repetitions with minor changes, or errors only; four points for exact repetition or appropriate paraphrased answers. The total score is max. 288 points.

1. 你是老师**还是**学生? (beginning level)
   Are you a teacher **or** a student?
2. 他去**过**上海很多次. (intermediate level)
   He went (**past tense**) to Shanghai a lot of times.
3. 那条裤子要**多少**钱? (advanced level)
   How much do those pants cost?

Spanish DELE proficiency test (short version)

The short version of the Spanish DELE proficiency test was developed based on the reading test of the traditional *Diplomas de Español como Lengua Extranjera* (*DELE*, Diplomas of Spanish as a Foreign language) proficiency test. The traditional test aimed to assess learners' general language proficiency in Spanish as a foreign language (FL) (Instituto Cervantes, 2010). The short version is designed as a multiple-choice and cloze test with identical test targets. It comprises 30 multiple-choice questions and a long cloze test (20 questions), with each question carrying one point, resulting in a total score of 50. Three distinct levels were distinguished based on the scores: beginning level (0–29), intermediate level (30–39), and advanced level (40–50).

4.2.3 Procedure

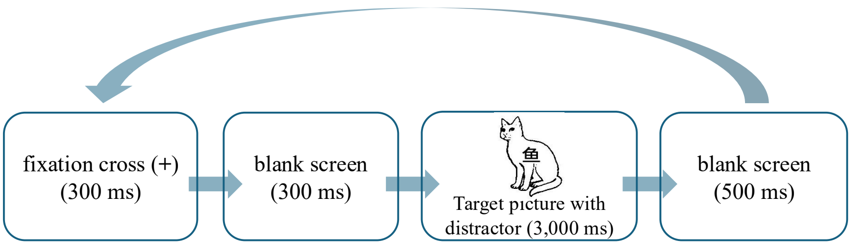
4.2.3.1 Picture naming task

All participants first completed the picture naming task. This task was programmed in E-prime 2.0 (Schneider et al., 2002) with a by-subject order design. This design was achieved using the Windows program Mix (Van Casteren & Davis, 2006) to generate a pseudo-random trial order, counterbalancing the effect of order. Thus, trials from the same condition or featuring the same target picture were not allowed to be presented consecutively, and no more than two consecutive trials could involve the same target classifier. As a result, the order of trials was randomized across participants.

The Mandarin Chinese and early Spanish−Chinese groups followed the same experimental procedure. Consistent with the traditional PWI paradigm, the experiment involved three key phases: familiarization, practice, and main experimental session. During the familiarization session, participants learned the exact names of 21 target pictures, each displayed on the screen for 3,000 ms along with its exact name underneath. Next, in the practice session, participants named target pictures using a bare noun while a meaningless string ("XX") appeared at the center of the screen. Each picture was displayed for 3,000 ms, and any unexpected responses were immediately corrected after all 21 pictures had been presented.

In the experimental session, each target picture randomly appeared four times, once with each of the four distractors, totaling 84 trials. In each trial, one or two identical pictures were presented simultaneously for 3,000 ms on the screen. Participants were instructed to name the target pictures quickly and accurately with Chinese noun phrases (e.g., *一只猫*, /yi1zhi1 mao1/ [one + classifier-zhi1 + cat], "one cat") while ignoring distractor words. Each trial followed a standard sequence (see Figure 4.2.1), i.e., a fixation cross at the center of the screen for 300 ms, followed by a blank screen for 300 ms. Subsequently, the target picture paired with a distractor was presented for 3,000 ms, followed by a blank screen for 500 ms. Vocal responses and EEG data were recorded during this session.

Figure 4.2.1. Illustration of the trial sequence in the experimental session.



4.2.3.2 EEG recordings

EEG data were collected with Ag/AgCl active electrodes via the BrainVision Recorder software by Brain Products GmbH. We used an actiCAP electrode cap (Easycap GmbH) with a BrainAmp standard amplifier (Brain Products GmbH) following a standard 32-electrode 10/20 montage (see Appendix 4.C). Of these 32 electrodes, FT9 was placed below the left eye to monitor the horizontal eye movement (HEOG).

Similarly, FT10 was placed at the outer canthus of the right eye to monitor vertical electrooculogram (VEOG). Additionally, TP9 and TP10 were placed on the left mastoid (M1) and right mastoid (M2), which were later reused as re-referencing channels. Thus, data were measured at twenty-eight channel locations at a sampling rate of 500 Hz. In addition, two external electrodes were plugged into the cap to use as a ground electrode (GND) and an online reference electrode (FCz/REF; see Appendix 4.C). The impedance was kept below 10 kΩ for all electrodes.

4.2.3.3 Chinese Elicited Imitation (EI) test

The second task was the Chinese EI test, which was only tested with the early Spanish−Chinese bilingual speakers to assess their Chinese proficiency. This task was programmed in E-prime with a by-participant design, with all sentences randomized across participants. Following instructions in Yan et al. (2020), procedures for each trial were as follows: first, each stimulus sentence was played to each participant only once; next, a blank screen with 2 seconds of silence was presented to avoid rote memory; third, a 0.5-second ringtone prompted the start of repetition, followed by participants' repetition of the sentences in 20 seconds. The voice response was recorded automatically.

4.2.3.4 Spanish DELE proficiency test

The third task was the Spanish DELE proficiency test, which was only taken by the early Spanish−Chinese bilingual speakers. This was done to evaluate their Spanish language proficiency. This task was completed in printed form, and participants were requested to answer all questions in 10 minutes. All answer sheets were collected for further calculations.

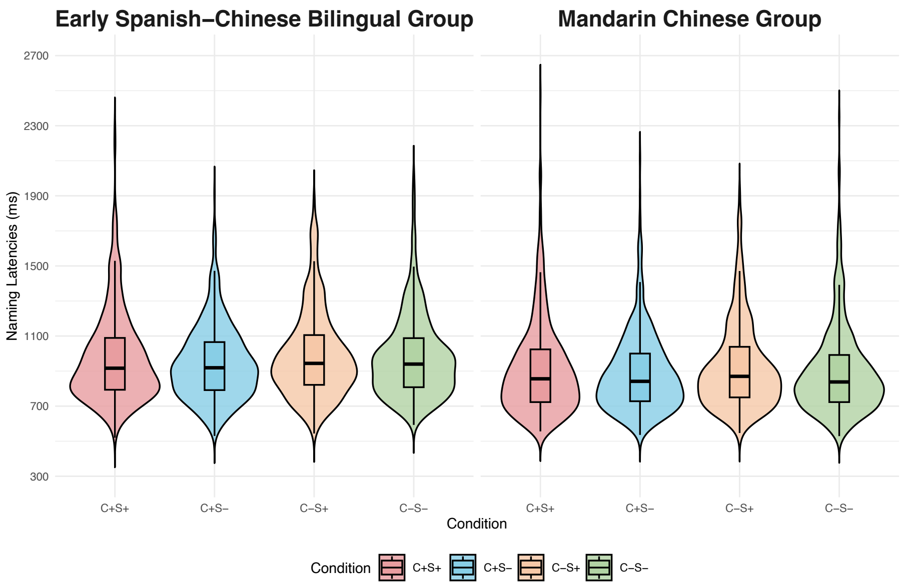
4.3 Results

4.3.1 Behavioral results

All recordings in two groups were first separately preprocessed using Praat (Boersma & Weenink, 2019), and naming latencies were then calculated and extracted for subsequent analysis (see Figure 4.3.1). Trials with incorrect, missed, or revised responses, and those marked by hesitation, were excluded from further analysis. Next, we analyzed the behavioral data for two groups separately using a generalized linear mixed model (GLMM) with a *gamma* distribution. This analysis was conducted via the *glmer* function from the *lme4* package (Bates et al., 2015) in Rstudio version 4.2.2 (*R Core Team*). To avoid the risk of inflating the Type I error rate, a top-down model selection approach was applied in the analysis of the behavioral data, starting with the theoretical maximum model (Barr, 2013). In order to fit the model to our data, the theoretical maximum model incorporated *classifier congruency* and *semantic relatedness* as fixed factors, with *item* and *participant* included as random factors.

For both groups, the full model structure remained consistent, incorporating *classifier congruency (congruent vs. incongruent)* and *semantic relatedness (related vs. unrelated)* as dummy-coded (1 vs. 0) fixed factors, random intercepts for participants and target items, by-participant random slopes for classifier congruency and semantic relatedness, as well as their correlation within the participant factor.

Figure 4.3.1. Mean naming latencies (ms) for each condition in the Mandarin Chinese group and the early Spanish−Chinese bilingual group.



4.3.1.1 Mandarin Chinese group

In the Mandarin Chinese group, a total of 2,520 trials were generated from the 21 picture stimuli. During preprocessing, 8.01% of the trials were excluded, and 0.79% were removed due to outliers, which were identified as naming latencies exceeding three standard deviations (SDs) from the mean response time of participants. This yielded 2,298 trials in total for subsequent statistical analysis. Considering our experiment design and model fit checks by plotting the model residuals against predicted values, we confirmed the full model was the best-fit model. Participants in the Mandarin Chinese group were significantly faster in classifier-congruent conditions with $\beta = -13.99$, $SE = 6.43$, $t = -2.18$, and $p = 0.03$ compared to the incongruent conditions. Moreover, participants responded significantly slower to semantically related conditions compared to semantically unrelated conditions, with $\beta = 22.71$, $SE = 6.09$, $t = 3.73$, and $p < .001$ (see Table 4.3.1).

Table 4.3.1. GLMM of best fit for naming latencies in the Mandarin Chinese group (n = 30), with classifier congruency and semantic relatedness as two predictors, including estimates, confidence intervals, and *p*-values.

Formula: Naming latency ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + (Classifier congruency + Semantic relatedness | participant) + (1 | item)

Predictors	**RTs**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	933.484	910.903 – 956.065	81.068	<0.001***
Classifier [Congruent]	−13.988	−26.596 – −1.380	−2.176	0.030*
Semantic [Related]	22.713	10.779 – 34.646	3.732	<0.001***
Random Effects				
σ^2	0.04			
$\tau_{00\ \text{Participant}}$	4845.50			
$\tau_{00\ \text{Item}}$	744.10			
$\tau_{11\ \text{participant. classifier-congruent}}$	305.89			
$\tau_{11\ \text{participant. semantic-related}}$	20.81			
$\rho_{01\ \text{participant. classifier-congruent}}$	− 0.19			
$\rho_{01\ \text{participant. semantic-related}}$	− 0.16			
ICC	1.00			
$N_{\text{participant}}$	30			
N_{Item}	21			
Observations	2298			
Marginal R^2/Conditional R^2	0.032/1.000			

4.3.1.2 Early Spanish−Chinese bilingual group

In the bilingual group, the 21 picture stimuli resulted in a total of 2,436 trials. Of these, 9.11% were removed during the preprocessing stage, and 0.49% were excluded due to outliers. Thus, a total of 2,202 trials were included for subsequent analysis. The best-fit model was generated as follows: first, the full model was constructed; second, the correlation of classifier congruency and semantic relatedness for the participant factor was removed in the case of singular fit. As a result, the best model was fitted, including classifier congruency and semantic relatedness as fixed effects, random intercepts for participants and target items, and by-participant random slopes for classifier congruency and semantic relatedness. The best-fit model demonstrated significantly shorter naming latencies in classifier congruent trials, with $\beta = -24.42$, $SE = 8.37$, $t = -2.92$, and $p = 0.004$ compared to incongruent trials. Additionally, early Spanish−Chinese bilinguals showed significantly longer naming latencies in semantically related than in semantically unrelated trials, with $\beta = 22.19$, $SE = 9.07$, $t = 2.45$, and $p = 0.014$ (see Table 4.3.2). Additionally, the average naming latencies for Mandarin Chinese and early Spanish−Chinese bilingual groups are presented in Table 4.3.3.

Table 4.3.2. GLMM of best fit for naming latencies in the early Spanish−Chinese bilingual group (n = 29), with classifier congruency and semantic relatedness as two predictors, including estimates, confidence intervals, and *p*-values.

Formula: Naming latency ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + (Classifier congruency| participant) + (Semantic relatedness | participant) +(1 | item)

Predictors	**RTs**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	994.597	960.662 – 1028.533	57.476	<0.001***
Classifier [Congruent]	−24.418	−40.836 – −8.000	−2.917	0.004**
Semantic [Related]	22.185	4.407 – 39.963	2.447	0.014*
Random Effects				
σ^2	0.05			
τ_{00} Participant. semantic-related	2901.42			
τ_{00} Participant. classifier-congruent	313.74			
τ_{00} Item	1319.24			
τ_{11} participant. semantic-related	886.43			
τ_{11} participant. classifier-congruency	825.88			
ρ_{01} participant. semantic-related	−0.09			
ρ_{01} participant. classifier-congruent	−0.78			
ICC	1.00			
$N_{participant}$	29			

N Item	21
Observations	2202
Marginal R^2/Conditional R^2	0.057/1.000

Table 4.3.3. Mean picture naming latencies by conditions in Mandarin Chinese and early Spanish−Chinese bilingual groups.

Conditions	Naming latencies (ms)			
	Mandarin Chinese group		Spanish−Chinese group	
	Mean	SD	Mean	SD
Classifier-congruent/Semantically Related (C+S+)	911	264	978	262
Classifier-congruent/Semantically Unrelated (C+S−)	888	223	945	212
Classifier-incongruent/Semantically Related (C−S+)	925	246	992	247
Classifier-incongruent/Semantically Unrelated (C−S−)	899	252	980	235

4.3.2 Chinese Elicited Imitation (EI) test results

The Chinese EL scores were calculated following the original scoring criteria. Almost all participants in the early Spanish−Chinese bilingual group accurately repeated or paraphrased sentences without altering their meanings (M_{score} = 283.41, SD_{score} = 3.01). Overall, all participants were classified within the advanced proficiency level.

4.3.3 Spanish DELE proficiency test (short version) results

The Spanish DELE proficiency scores were calculated by summing the correct answers from participants' answer sheets. Nearly all participants achieved scores within the advanced range (M_{score} = 45.73, SD_{score} = 2.95), indicating an advanced proficiency level in Spanish. Note that scores of either the Chinese EL test or the Spanish DELE test were not considered an absolute measure but rather an indicator of language proficiency.

4.3.4 EEG data results

4.3.4.1 EEG data analysis

EEG data from two groups were pre-processed separately in Brain-Vision Analyzer 2.1 (Brain Products GmbH) following standardized procedures. First, signals were visually inspected for quality. Next, recordings were re-referenced to the average of the left and right mastoid electrodes (M1 and M2). The data were then filtered using a high-pass filter of 0.1 Hz and a low-pass filter of 30 Hz. Ocular independent component analysis (ICA) with VEOG and HEOG as parameters was used to correct blink activity. Artifact rejection was subsequently performed to minimize noise. Finally, the signals were segmented to include only correct trials, generating epochs around picture onsets to analyze the voltage amplitude of ERP components of interest. Each epoch was defined as the interval spanning 200 ms before and 1,200 ms after trial onset, determined based on the average naming latency duration. Baseline correction was applied to the 200 ms prior to picture onset until picture onset, after which the epochs were extracted for statistical analysis. Before the statistical analysis, permutation tests were computed using the permutes package (Voeten, 2019) for each group to tentatively detect the locus of the effect of *classifier congruency* and *semantic relatedness* on voltage amplitudes. Specifically, we calculate F-values across all electrodes and the entire time window between 200 ms and 1,200 ms with respect to stimulus onset. Next, we performed a single-trial linear mixed model (LMM) method with *lmer()* function on each group separately for statistical analysis.

4.3.4.2 Mandarin Chinese group (EEG results)

Results of classifier congruency. The permutation test results for the Mandarin Chinese group revealed potential modulatory effects of classifier congruency (congruent vs. incongruent) in electrodes Pz, P8, P7, P4, P3, Oz, O2, O1, Fz, FC2, FC1, F8, F7, F4, F3, Cz, CP5, CP2, and CP1 in 400 ms – 500 ms time windows post-stimulus onset (see the left panel in Figure 4.3.2). Further, we did a visual inspection of these electrodes and observed a negative-going wave elicited by classifier-incongruent conditions between 400 ms and 500 ms, consistent with the topographic distribution of N400 (Kutas & Hillyard, 1980; Chwilla et al., 1995). Descriptively speaking, we observe the smaller amplitude for the classifier-incongruent condition with M = 4.61 (SD = 10.85) compared to the classifier-congruent condition with M = 5.40 (SD = 10.73) in the 400 – 500 ms time window. To do further amplitude analyses, we grouped the electrodes based on their location, which are left fronto–central, left centro–parietal, right fronto–central, right centro–parietal, and central regions. Furthermore, *location* of grouped electrodes was included as a covariate for amplitude statistical analyses. Additionally, although visual inspection suggests the presence of an N2 component around 200 ms, its underlying cause in this study remains unclear and falls outside the scope of the current study. Therefore, we do not further interpret this component.

The best-fitting model for voltage amplitudes included *classifier congruency (congruent vs. incongruent)* and *semantic relatedness (related vs. unrelated)* as dummy-coded (0 vs. 1) main effects, with *location* of grouped electrodes as a covariate. *Participant* and *item* were included as random effects, with a by-participant random slope for the effects of *classifier congruency* and *semantic relatedness* (see Table 4.3.4 in Appendix 4.D). Voltage amplitudes were significantly more negative for classifier-incongruent compared to congruent conditions with $\beta = -0.74$, $SE = 0.241$, $t = -3.064$, $p = 0.0046$[6].

[6] Note: The p-values in the main text were computed using Satterthwaite's method for better accuracy in mixed-effects models. These values may differ slightly from those

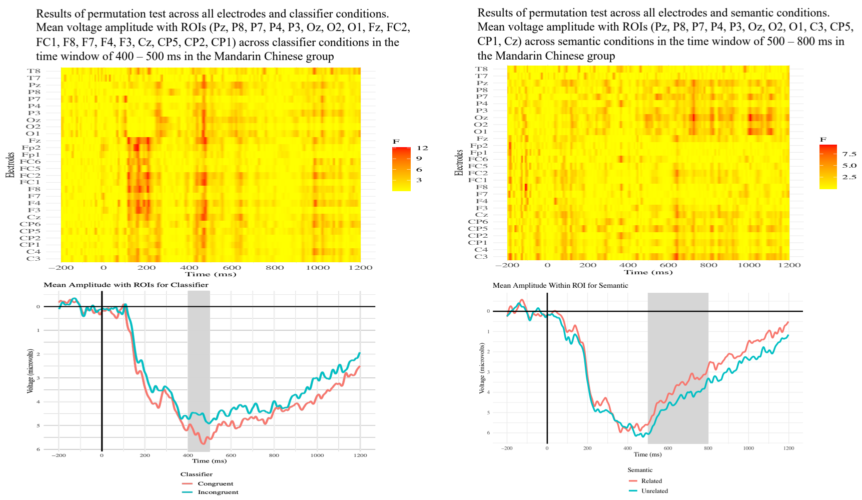
Results of semantic relatedness. The permutation test showed pote-ntial modulatory effects of semantic relatedness (related vs. unrelated) with electrodes Pz, P8, P7, P4, P3, Oz, O2, O1, C3, CP5, CP1, and Cz within the 500 – 800 ms time window (see the right panel in Figure 4.3.2). Visual inspection illustrated a positive-going wave elicited by semantically unrelated conditions between 500 ms and 800 ms, which is aligned with the topographic distribution of the typical P600 (Osterhout & Holcomb, 1992) and the "semantic P600" (see Kuperberg et al., 2003; Van Herten et al., 2005; Bornkessel-Schlesewsky & Schlesewsky, 2008, for a review). Specifically, the amplitude for semantically unrelated conditions (M = 4.51, SD = 11.08) is larger than for semantically related conditions with M = 3.89 (SD = 11.39) in the 500 – 800 ms time window. Electrodes within ROIs were grouped based on their location, that is, left centro–parietal, right centro–parietal, and central regions. Further, *location* of grouped electrodes was a covariate for amplitude statistical analyses. We found no indication of an N400 effect prior to the 600 ms based on the permutation test and visual inspections.

The best-fit model for our data was with *classifier congruency* and *semantic relatedness* as fixed effects, and *participant* and *item* as random effects. *Location* was included as a covariate, with a by-participant random slope for the effects of *classifier congruency* and *semantic relatedness* (see Table 4.3.5 in Appendix 4.D). Semantically unrelated conditions elicited a more positive amplitude than semantically related conditions with $\beta = 0.68$, $SE = 0.24$, $t = 2.81$, $p = 0.0086$.

---

presented in Table 4.3.4, 4.3.5, 4.3.6 and 4.3.7 (Appendix 4.D and 4.E), which were derived using Wald tests with *tab_model* function.

Figure 4.3.2. Results of the permutation test for classifier congruency (left panel) and semantic relatedness (right panel) across all electrodes from −200 ms to 1,200 ms, showing F-values and corresponding mean voltage amplitudes within ROIs and time windows in the Mandarin Chinese group.



4.3.4.3 Early Spanish−Chinese bilingual group (EEG results)

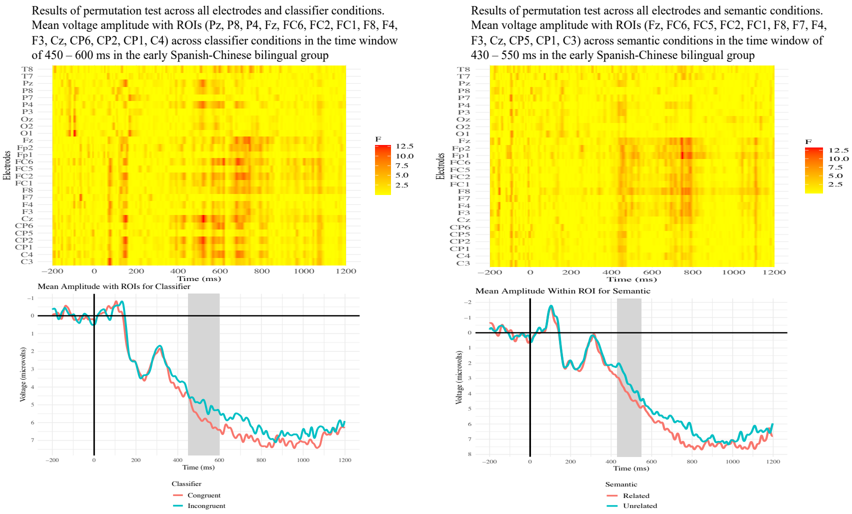
Results of classifier congruency. By following the same procedure as in the Mandarin Chinese group, we found a potential modulation effect of classifier congruency with electrodes Pz, P8, P4, Fz, FC6, FC2, FC1, F8, F4, F3, Cz, CP6, CP2, CP1, and C4, showing a negative-going wave in classifier-incongruent condition in the 450 – 600 ms time window (see left panel in Figure 4.3.3). This is consistent with the potential N400 effect found in the Mandarin Chinese group. Further, these electrodes were grouped into five ROIs: left fronto–central, left centro–parietal, right fronto–central, right centro–parietal, and central regions. In general, classifier-incongruent conditions (M = 4.99, SD = 12.2) showed a smaller amplitude than classifier-congruent conditions with M = 5.76 (SD = 12.0).

The best-fitting model incorporated *classifier congruency* and *semantic relatedness* as fixed effects, with *participant* and *item* as random effects.

Location was included as a covariate, and a by-participant random slope was specified for *classifier congruency* and *semantic relatedness* (see Table 4.3.6 in Appendix 4.E). The result demonstrated that classifier-incongruent conditions elicited a more negative amplitude than classifier-congruent conditions with $\beta = -0.77$, $SE = 0.33$, $t = -2.38$, $p = 0.0242$.

Results of semantic relatedness. Following the same steps, a potential modulation effect of semantic relatedness with a negative-going wave in the semantically unrelated condition was detected with electrodes Fz, FC6, FC5, FC2, FC1, F8, F7, F4, F3, Cz, CP5, CP1, C3 in the 430 – 550 ms time window (see the right panel in Figure 4.3.3). This effect aligned with the topographic distribution of the N400 effect (Kutas & Hillyard, 1980; Chwilla et al., 1995). The descriptive result showed that the amplitude in semantically unrelated conditions ($M = 3.26$, $SD = 11.6$) is smaller than in semantically related conditions ($M = 4.06$, $SD = 12.2$). For further statistical analysis, these electrodes were grouped into four ROIs, including left fronto–central, right fronto–central, left centro–parietal, and central regions. By fitting the *classifier congruency* and *semantic relatedness* as fixed factors, *participant* and *item* as random factors, *location* as a covariate, as well as by-participant random slopes for *classifier congruency* and *semantic relatedness*, the best model was generated. The result demonstrated that semantically unrelated conditions elicited a more negative amplitude than semantically related conditions with $\beta = -0.89$, $SE = 0.39$, $t = -2.30$, $p = 0.0287$ (see Table 4.3.7 in Appendix 4.E).

Figure 4.3.3. Results of the permutation test for classifier congruency (left panel) and semantic relatedness (right panel) across all electrodes from −200 ms to 1,200 ms, showing F-values and corresponding mean voltage amplitudes within ROIs for time windows of interest in the early Spanish–Chinese bilingual group.



4.4 Discussion

In this study, we explored the processing of classifier congruency and semantic relatedness in Mandarin Chinese NPs among early Spanish–Chinese bilinguals and Mandarin Chinese speakers using the PWI paradigm. We manipulated classifier congruency and semantic relatedness between targets and distractors by categorizing them as either classifier-congruent or classifier-incongruent and as either semantically related or unrelated. This allowed us to compare naming latencies between classifier-congruent and classifier-incongruent NPs to assess the behavioral representation of the classifier congruency effect. Meanwhile, we analyzed amplitude differences between these two conditions to detect potential ERP components (e.g., N400 or P600) to determine whether the classifier congruency effect was present electrophysiologically. We made a similar comparison between semantically related and semantically unrelated conditions to identify the semantic interference effect behaviorally and electrophysiologically.

Classifier congruency effect

In line with our expectations, we observed a significant classifier congruency effect behaviorally in both groups. Specifically, in both groups, behavioral results showed that participants' naming latencies were shorter when naming target pictures that were classifier-congruent with distractor words, compared to classifier-incongruent conditions. In other words, the classifier congruency effect was identified behaviorally in naming classifier-congruent targets and distractors. This aligns with findings in NP production observed in previous studies by Wang et al. (2006), Huang and Schiller (2021), Wang and Schiller (submitted) and Wang et al. (submitted), which reported longer naming latencies in classifier-incongruent conditions compared with congruent conditions in monolingual Chinese speakers' NP production.

Notably, the classifier congruency effect was not observed in the behavioral results during bare nouns production in Wang et al. (2019) and Wang et al. (2006). The main reason for the absence of such an effect in these studies is presumably that bare noun naming does not require the production of classifiers in participants' responses. According to the LRM speech production model (Levelt et al., 1999), when naming a picture, the conceptual and lexical representation of to-be-produced words need to be encoded first. Next, syntactic features, such as classifier features in Chinese, are automatically activated and eventually selected. Finally, the to-be-produced word is competitively selected and produced. In bare noun naming tasks, the classifier feature is not needed for production, and as a result, it is not selected. Therefore, no classifier congruency effect is observed in bare noun naming. In our study, the classifier is required for producing a Chinese NP, where it is automatically activated and eventually selected. As a result, the *classifier congruency effect* was observed in both groups.

Meanwhile, we observed that the classifier congruency effect was accompanied by a more negative amplitude elicited by classifier-incongruent conditions compared to congruent conditions in both groups. In the Mandarin Chinese group, negative amplitudes were observed within the 400 – 500 ms time window post-stimulus onset, with the largest effects observed at the fronto–central, central, and centro–parietal electrodes. Similarly, in the early Spanish–Chinese bilingual group, negative waves were maximal at the fronto–central, central, and centro–parietal regions within the 450 – 600 ms time window post-stimulus onset. This time window and regions closely align with the typical N400 effect (Kutas & Hillyard, 1980; Chwilla et al., 1995). Crucially, in both groups, amplitudes were significantly more negative in classifier-incongruent conditions compared to classifier-congruent conditions. This aligns with the electrophysiological result in NP production reported by Huang and Schiller (2021) and the bare noun production results reported by Wang et al. (2019). Both of these studies observed significantly more negative ERP waves elicited by classifier-incongruent compared to congruent conditions within the time windows of 370 – 430 ms or 385 – 585 ms in fronto–central and centro–parietal, respectively, reflecting an N400-like effect. Additionally, the more negative wave in classifier-incongruent conditions is consistent with previous EEG studies on classifier congruency (Chou et al., 2014; Zhang et al., 2012; Zhou et al., 2010). Given the consistency in time windows and electrode regions with previous studies, we also identified an N400-like effect for the classifier congruency effect in both the Mandarin Chinese and early Spanish–Chinese bilingual groups.

Semantic interference effect

On the other hand, behavioral results from semantic relatedness manipulation (i.e., semantically related vs. semantically unrelated) in both groups are in accordance with the well-established *semantic interference effect*. This is evidenced by longer naming latencies were observed in both groups when targets were named in the presence of distractors from the same semantic category (e.g., Rosinski, 1977; La Heij, 1988; Glaser, 1992;

Schriefers et al., 1990). Similar findings were reported by Wang et al. (2019), Huang and Schiller (2021), Wang and Schiller (submitted) and Wang et al. (submitted) with monolingual Chinese speakers, as well as Dutch late learners of Mandarin Chinese (L2 learners) in Wang and Schiller (submitted), where monolingual Chinese speakers and L2 learners exhibited prolonged naming latencies in semantically related targets and distractors in NP and bare noun production. The rationale is that when naming targets that are semantically related to the simultaneously presented distractors, the lemmas of both the target and distractor, along with their associated lexical nodes, are activated simultaneously. The lemmas of semantically related distractors are activated from the same semantic category as the target, leading to an increase in competition in lexical selection and making selection more difficult, thereby prolonging the lexical selection process (Roelofs, 1992, 1993; Levelt et al., 1999). This view was also used to explain the observation of semantic interference in other studies (Belke et al., 2005; Bloem & La Heij, 2003; Roelofs, 2003). The early Spanish–Chinese bilinguals, raised in heritage Chinese households and Spanish society, demonstrated advanced proficiency in both Spanish and Chinese (see Sections 4.3.2 & 4.3.3), which facilitated the observation of the semantic interference effect.

Along with the significant semantic interference effect observed in the behavioral data, we also identified two ERP components associated with this effect in the Mandarin Chinese group and the early Spanish–Chinese bilingual group, respectively, elicited by semantically unrelated conditions against related conditions. Specifically, in the Mandarin Chinese group, semantically unrelated conditions elicited a positive wave within the 500 – 800 ms time window after stimulus onset at centro–parietal regions. This seems to align with the "semantic P600" or centro–parietal P600 effect, which typically has a centro–parietal scalp distribution within a 500 – 800 ms time window (see Kuperberg et al., 2003; Van Herten et al., 2005; Bornkessel-Schlesewsky & Schlesewsky, 2008). In the early Spanish–Chinese bilingual group, we found a negative wave that peaked around 400 ms within the 430 – 550 ms time window elicited by semantically unrelated conditions in the fronto–central and centro–parietal regions. These results largely overlap with the typical N400 effect (Kutas & Hillyard, 1980; Chwilla et al., 1995) and are consistent with the results by Huang and Schiller (2021) and Wang et al. (2019), where they reported a more negative ERP wave around 400 ms evoked by the presence of target pictures with semantically unrelated distractors, reflecting an N400-like effect.

There appears to be a discrepancy in the ERP components observed in the semantic interference effect between the Mandarin Chinese group and the early Spanish–Chinese bilingual group. Specifically, the Mandarin Chinese group showed a semantic P600-like effect, while the early Spanish–Chinese group exhibited an N400-like effect. To address this discrepancy, it is important to consider the characteristics of the N400 and P600 components. As discussed in section 4.1.3, the N400 effect is typically semantically driven, reflecting lexico-semantic integration and lexical co-activation processes (Lau et al., 2008; Kutas & Federmeier, 2011; Chen et al., 2017; Leckey & Federmeier, 2019). In contrast, the P600 effect is typically syntactically driven, reflecting the syntactic integration process (Friederici et al., 1993; Hagoort et al., 1993), while the "semantic P600" effect has been proposed and identified as an electrophysiological response to "integration difficulties" in semantic processing of integrating an unexpected entity into the discourse context (see Kuperberg et al., 2003; Van Herten et al., 2005; Bornkessel-Schlesewsky & Schlesewsky, 2008, for a review), which can be interpreted as semantically driven within syntactic contexts. For the early Spanish–Chinese bilingual group, especially the fact that they have high proficiency in both languages, it is plausible that the semantic integration process is quicker and automatic, which would trigger the N400-like effect as a marker of semantic processing difficulty in semantically unrelated conditions. For the Mandarin Chinese group, given that the semantic P600 effect was proposed and linked to difficulties in semantic integration, the more positive wave observed could reflect a semantic P600-like effect.

However, the P600-like effect elicited by semantically unrelated conditions in the current study seems unexpected, particularly when compared to the N400 effect commonly reported in PWI-based ERP studies in Mandarin Chinese and other languages (e.g., Dell'Acqua et al., 2010; Wicha et al., 2003; Zhu et al., 2015), as few studies have explored or documented this effect in semantic interference. Yet, when considering the production of Chinese NPs, such effects are not entirely surprising. In classifier-NPs, the noun has a comparable lexico-syntactic feature marked by the classifier, and its processing is thought to involve both semantic and syntactic information (Wang et al., 2019). As demonstrated above, classifiers encode inherent properties of nouns, and their selection must be determined by and align with the properties of nouns. This implies that the processing of nouns within NPs inherently requires activation and retrieval of lexico-syntactic information, even when producing bare nouns. Evidence for this comes from the P600 effect observed in PWI-based Chinese bare noun naming, which was elicited by the activation of inherent lexico-syntactic features of nouns, suggesting syntactically driven processing of bare nouns (Wang et al., 2024). Given the role of lexico-syntactic features in Chinese NPs, producing a noun within an NP likely automatically requires the activation and retrieval of lexico-syntactic information. Additionally, since Mandarin Chinese speakers typically acquire classifier-noun combinations by the age of four to five (Erbaugh, 1986), and these combinations must be rotely memorized due to the opaque mapping between classifiers and nouns, the connection between the lexico-syntactic feature of classifiers and the noun remains relatively fixed and reliable. This stable association likely contributes to the P600-like effect observed in Mandarin Chinese speakers. Therefore, the observed P600-like effect may presumably reflect the processing of lexico-syntactic information within NPs.

Additionally, the absence of an N400 response to the semantic interference effect in our data could also be attributed to the inclusion of a relatively high proportion (i.e., 19%) of animate nouns (e.g., McRae et al., 2005; also see Wang et al., 2024 for exploring the role of animacy in bare noun production). Evidence from functional neuroimaging studies suggests that distinct brain regions are involved in processing living (animate nouns) and non-living (inanimate nouns) categories (Chao et al., 1999). In the processing of classifier-noun combinations, the animacy information is not immediately used for semantic integration between nouns and classifiers but rather plays a role in the later processing stage, which would reflect in the P600 effect (Zhang et al., 2012). In contrast, studies that have reported an N400 to semantic interference included a lower proportion of animate nouns, such as 12% in Huang and Schiller (2021) and 15% in Wang et al. (2019). It should be noted that Blackford et al. (2012) dissociated the connection between the behavioral semantic interference effect and the electrophysiological N400-like effect. While the N400-like effect may also arise from semantic priming (Kreher et al., 2006; Blackford et al., 2012), it remains unclear whether it directly corresponds to the semantic interference effect observed in behavioral results. Additionally, Costa et al. (2009) found no significant correlation between naming latencies and the ordinal position of pictures from the same categories within the 400 ms range of the ERP component. Thus, further research is needed to clarify the electrophysiological mechanisms underlying semantic interference.

4.5 Conclusion

To conclude, the current study examined the processing of Chinese classifiers in noun phrase production by early Spanish–Chinese bilinguals and Mandarin Chinese speakers through overt picture naming experiments. Using a picture-word interference paradigm, we investigated naming latencies and electrophysiological correlates of Chinese NP production in four conditions, manipulating classifier congruency (congruent vs. incongruent) and semantic relatedness (related vs. unrelated). Behavioral results showed that both Mandarin Chinese speakers and early Spanish–Chinese bilinguals took a significantly longer time to name targets with classifier-incongruent and semantically related distractors relative to classifier-congruent and semantically unrelated conditions. Electrophysiological data revealed significant N400-like effects in response to classifier-incongruent conditions for both groups, with the bilingual group showing N400-like negativity and the Mandarin Chinese group exhibiting significant P600-like positivity in the semantically unrelated relative to semantically related conditions. Moreover, bilinguals exhibited longer naming latencies and a delayed N400-like effect in classifier-incongruent conditions compared to the Mandarin Chinese group. Overall, this study provides behavioral and electrophysiological evidence that early Spanish–Chinese bilinguals exhibit robust classifier congruency effects in Chinese noun phrase production, mirroring the processing patterns of monolingual speakers in previous studies. These findings suggest that early and sustained bilingual exposure enables the acquisition of morphosyntactic features absent in one of bilinguals' first languages (L1s), supporting models of competitive selection in bilingual language production. Future research should explore the generalizability of these effects across different bilingual populations and morphosyntactic domains, further illuminating the mechanisms underlying bilingual language processing.

Acknowledgements

We sincerely thank all participants for their time and commitment to this study. We are also grateful to the research team at the Center for Brain and Cognition, Universitat Pompeu Fabra (UPF), for their valuable assistance throughout the project. Our deepest appreciation goes to Prof. Núria Sebastián Gallés, head of the Speech Acquisition and Perception (SAP) group at UPF, for her kind support and hospitality. We further acknowledge the technical guidance and logistical support provided by Francesc Xavier Mayoral Terron and Silvia Blanch Rabat from the Department of Technology at UPF, whose contributions were instrumental during the experimental setup and data collection phases.

Declaration of conflicting interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This work was partially funded by the China Scholarship Council (CSC) NO. 202107720015. Niels O. Schiller is supported by CityUHK start-up grant no.9380177 from City University of Hong Kong.

Data Availability Statement

The data supporting the findings of this study are openly available in Open Science Framework at https://osf.io/qwe7p/?view_only=a232bf276df54febbb162b46d940790d (view-only link).

Appendix

Appendix 4.A

1. Overview of linguistic profile for the Mandarin Chinese group (N = 30) according to the LEAP-Q (Marian et al., 2007).

Mandarin Chinese speakers	Mean (SD)/Distribution
Number of female/male participants	21/9
Mean age in years (SD; range)	24.36 (3.05; 22-30)
Number of participants pursuing a Bachelor degree (BA)	3
Number of participants pursuing a Master degree (MA)	22
Number of participants pursuing a Doctoral degree (PhD)	5
Number of participants with Spanish proficiency (A1)	4
Number of participants with Spanish proficiency (A2)	7
Number of participants with Spanish proficiency (B1)	3
Daily use of Chinese (frequency)	70.17% (SD = 0.17)
Daily use of English (frequency)	25.6% (SD = 0.16)
Daily use of Spanish (frequency)	4.23% (SD = 0.13)

2. Overview of linguistic profile for the early Spanish–Chinese bilingual group (N = 29) according to the revised BCSP (Olson, 2022).

Spanish-Chinese bilinguals	Mean (SD)/Distribution
Number of female/male participants	24/5
Mean age in years (SD; range)	21 (3.05; 18-29)
Number of participants born in Spain	25
Number of participants born in China	4
Number of participants with a high school diploma	3
Number of participants pursuing a Bachelor's degree (BA)	22
Number of participants pursuing a Master's degree (MA)	4
Daily use of Spanish (frequency)	44.86% (SD = 0.16)
Daily use of Chinese (frequency)	32.56% (SD = 0.17)
Daily use of Catalan (frequency)	12.17% (SD = 0.11)
Daily use of English (frequency)	10.17% (SD = 0.09)
Daily use of French (frequency)	0.21% (SD = 0.002)
Daily use of German (frequency)	0.03% (SD = 0.0003)
Number of participants with Catalan proficiency (C1/B2)	16/11
Number of participants with French proficiency (below B1)	14
Number of participants with German proficiency (A2)	1
Number of participants with Slovak proficiency (A2)	1
Number of participants with Italian proficiency (A1)	1

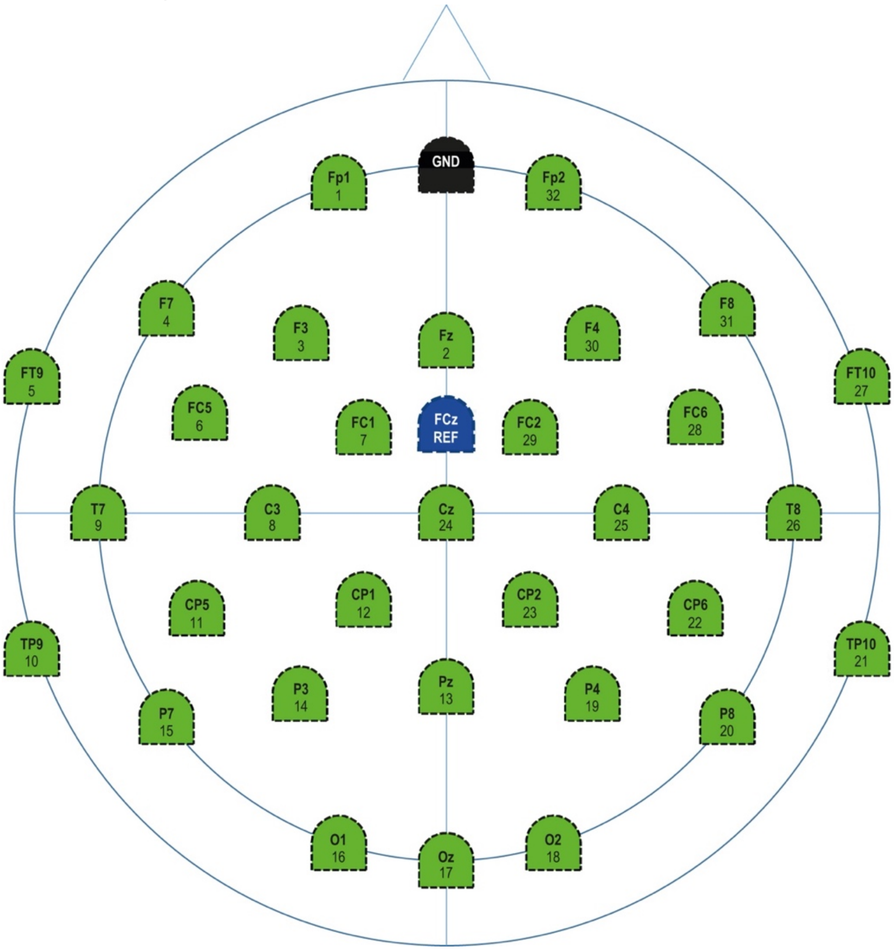
Appendix 4.B Stimuli used in the picture-naming task in Chapter 4.

Target picture	**Classifier**	**Distractor type**			
		Classifier-congruent		Classifier-incongruent	
		Semantically related	Semantically unrelated	Semantically related	Semantically unrelated
腿 (tui3) leg	条 (tiao2)	胳膊 (ge1bo0) arm	毛巾 (mao2jin1) towel	脚 (jiao3) foot	石头 (shi2tou0) stone
马 (ma3) horse	匹 (pi3)	狼 (lang2) wolf	布 (bu4) cloth	牛 (niu2) cow	刀 (dao1) knife
老虎 (lao3hu3) tiger	只 (zhi1)	猴子 (hou2zi0) monkey	船 (chuan2) boat	狮子 (shi1zi0) lion	骨头 (gu3tou2) bone
床 (chuang2) bed	张 (zhang1)	桌子 (zhuo1zi0) table	照片 (zhao4pian4) photo	椅子 (yi3zi0) chair	自行车 (zi4xing2che1) bicycle
耳朵 (er3duo1) ear	只 (zhi1)	眼睛 (yan3jing1) eye	鸽子 (ge1zi0) dove	鼻子 (bi2zi0) nose	棉花 (mian2hua1) cotton
叉子 (cha1zi0) fork	把 (ba3)	勺子 (shao2zi0) spoon	伞 (san3) umbrella	筷子 (kuai4zi0) chopsticks	牛奶 (niu2nai3) milk
围巾 (wei2jin1) scarf	条 (tiao2)	领带 (ling3dai4) tie	绳子 (sheng2zi0) rope	手套 (shou3tao4) glove	飞机 (fei1ji1) airplane
苹果 (ping2guo3) apple	个 (ge4)	桔子 (ju2zi0) mandarin	杯子 (bei1zi0) cup	香蕉 (xiang1jiao1) banana	竹子 (zhu2zi0) bamboo
花生 (hua1sheng1) peanut	粒 (li4)	玉米 (yu4mi3) corn	沙子 (sha1zi0) sand	小麦 (xiao3mai4) wheat	手指 (shou3zhi3) finger
梨 (li2) pear	个 (ge4)	桃 (tao2) peach	盘子 (pan2zi0) plate	葡萄 (pu2tao2) grape	报纸 (bao4zhi3) newspaper
裤子 (ku4zi0) pants	条 (tiao2)	裙子 (qun2zi0) skirt	尾巴 (wei3ba1) tail	衬衫 (chen4shan1) shirt	太阳 (tai4yang2) sun
嘴 (zui3) mouth	张 (zhang1)	脸 (lian3) face	表格 (biao3ge2) form	舌头 (she2tou0) tongue	熊猫 (xiong2mao1) panda
西红柿 (xi1hong2shi4) tomato	个 (ge4)	土豆 (tu3dou4) potato	书架 (shu1jia4) bookshelf	黄瓜 (huang2gua1) cucumber	云 (yun2) cloud
蛋糕	块	饼干	手表	面包	项链

(dan4gao1) cake	(kuai4)	(bing3gan1) cookie	(shou3biao3) watch	(mian4bao1) bread	(xiang4lian4) necklace
塔 (ta3) tower	座 (zuo4)	寺庙 (si4miao4) temple	雪山 (xue3shan1) snow mountain	楼 (lou2) building	西瓜 (xi1gua1) watermelon
包子 (bao1zi0) bun	个 (ge4)	饺子 (jiao3zi0) dumpling	网球 (wang3qiu2) tennis	面条 (mian4tiao2) noodle	剪刀 (jian3dao1) scissors
猫 (mao1) cat	只 (zhi1)	狗 (gou3) dog	袜子 (wa4zi0) sock	鱼 (yu2) fish	月亮 (yue4liang4) moon
鸭子 (ya1zi0) duck	只 (zhi1)	鸡 (ji1) chicken	手 (shou3) hand	蛇 (she2) snake	树 (shu4) tree
桥 (qiao2) bridge	座 (zuo4)	长城 (chang2cheng2) The great wall	岛 (dao3) island	地道 (di4dao4) tunnel	帽子 (mao4zi0) hat
猪 (zhu1) pig	只 (zhi1)	羊 (yang2) sheep	皮鞋 (pi2xie2) leather shoe	大象 (da4xiang4) elephant	书 (shu1) book
兔子 (tu4zi0) rabbit	只 (zhi1)	老鼠 (lao3shu3) mouse	行李箱 (xing2li3xiang1) suitcase	熊 (xiong2) bear	米 (mi3) rice

Appendix 4.C

Illustration of 32-electrode locations following a standard 32-electrode 10/20 montage.



Appendix 4.D

Table 4.3.4. LMM for Classifier congruency (N400) in the Mandarin Chinese group (n = 30).

Formula: Voltage amplitude ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + Location (left frontal central vs. left central parietal vs. right frontal central vs. right central parietal vs. central) + (Classifier congruency + Semantic relatedness | participant) +(1 | item)

Predictors	**Voltage amplitude**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	5.621	4.385 – 6.858	8.913	<0.001***
Classifier [Incongruent]	−0.739	−1.211 – −0.266	−3.064	0.002**
Semantic [Unrelated]	0.161	−0.331– 0.654	0.643	0.521
Location [left central parietal]	0.930	0.890 – 0.970	45.392	<0.001***
Location [left frontal central]	−2.136	−2.182 – −2.091	−91.567	<0.001***
Location [right central parietal]	0.356	0.314 – 0.398	16.480	<0.001***
Location [right frontal central]	−2.029	−2.075 – −1.984	−86.981	<0.001***
Random Effects				
σ^2	103.07			
$\tau_{00\ Participant}$	10.70			
$\tau_{00\ Item}$	0.86			
$\tau_{11\ participant.\ classifier\text{-}incongruent}$	1.74			
$\tau_{11\ participant.\ semantic\text{-}unrelated}$	1.89			
$\rho_{01\ participant.\ classifier\text{-}incongruent}$	0.00			
$\rho_{01\ participant.\ semantic\text{-}unrelated}$	− 0.24			
ICC	0.10			
$N_{participant}$	30			
N_{Item}	21			
Observations	2,098,474			
Marginal R^2/Conditional R^2	0.014/0.116			

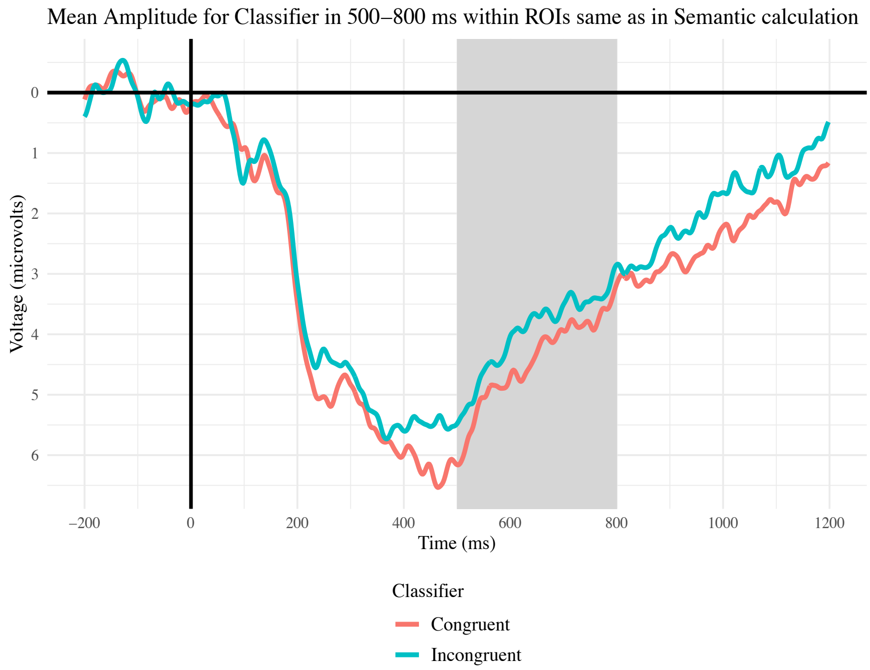
Table 4.3.5. LMM for semantic relatedness (P600) in the Mandarin Chinese group (n = 30).

Formula: Voltage amplitude ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + Location (left central parietal vs. right central parietal vs. central) + (Classifier congruency + Semantic relatedness | participant) +(1 | item)

Predictors	**Voltage amplitude**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	4.217	2.945 – 5.490	6.497	<0.001***
Classifier [Incongruent]	−0.474	−0.932 – −0.015	− 2.024	0.043*
Semantic [Unrelated]	0.675	0.204 – 1.146	2.810	0.005**
Location [left central parietal]	0.779	0.753 – 0.804	59.663	<0.001***
Location [right central parietal]	−1.842	−1.871 – −1.812	– 122.211	<0.001***
Random Effects				
σ^2	114.39			
$\tau_{00\ \text{Participant}}$	11.46			
$\tau_{00\ \text{Item}}$	0.83			
$\tau_{11\ \text{participant. classifier-incongruent}}$	1.64			
$\tau_{11\ \text{participant. semantic-unrelated}}$	1.73			
$\rho_{01\ \text{participant. classifier-incongruent}}$	− 0.15			
$\rho_{01\ \text{participant. semantic-unrelated}}$	− 0.48			
ICC	0.09			
$N_{\text{participant}}$	30			
N_{Item}	21			
Observations	4,030,152			
Marginal R^2/Conditional R^2	0.010/0.098			

Note: The classifier congruency is significant (p = 0.043) in the semantic model, however, it is not the classifier congruency effect here, as the voltage amplitude elicited by classifier incongruent trials has a negative-going direction in these ROIs and time windows of 500 – 800 ms. Please see Figure 4.3.4 for classifier congruency in these ROIs and time windows of 500 – 800 ms.

Figure 4.3.4. Mean amplitude for Classifier in 500 – 800 ms within the same ROIs as in Semantic calculations.



Appendix 4.E

Table 4.3.6. LMM for semantic relatedness (N400) in the early Spanish–Chinese bilingual group (n = 29).

Formula: Voltage amplitude ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + Location (left frontal central vs. left central parietal vs. right frontal central vs. central) + (Classifier congruency + Semantic relatedness | participant) +(1 | item)

Predictors	**Voltage amplitude**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	3.877	2.194 – 5.560	4.515	<0.001***
Classifier [Incongruent]	−0.500	−1.189 – 0.190	−1.420	0.156
Semantic [Unrelated]	−0.888	−1.643 – −0.132	−2.303	0.021*
Location [left central parietal]	2.801	2.745 – 2.857	98.526	<0.001***
Location [left frontal central]	−0.759	−0.812 – −0.706	−28.134	<0.001***
Location [right frontal central]	0.261	0.209 – 0.314	9.695	<0.001***
Random Effects				
σ^2	122.22			
$\tau_{00\ \text{Participant}}$	19.34			
$\tau_{00\ \text{Item}}$	1.47			
$\tau_{11\ \text{participant. classifier-incongruent}}$	3.58			
$\tau_{11\ \text{participant. semantic-unrelated}}$	4.30			
$\rho_{01\ \text{participant. classifier-incongruent}}$	− 0.03			
$\rho_{01\ \text{participant. semantic-unrelated}}$	− 0.70			
ICC	0.13			
$N_{\text{participant}}$	29			
N_{Item}	21			
Observations	1,638,312			
Marginal R^2/Conditional R^2	0.014/0.141			

Table 4.3.7. LMM for Classifier congruency (N400) in the early Spanish–Chinese bilingual group (n = 29).

Formula: Voltage amplitude ~ Classifier congruency (congruent vs. incongruent) + Semantic relatedness (related vs. unrelated) + Location (left frontal central vs. left central parietal vs. right frontal central vs. right central parietal vs. central) + (Classifier congruency + Semantic relatedness | participant) +(1 | item)

Predictors	**Voltage amplitude**			
	Estimate	*95% CI*	*t-value*	*Pr(>|z|)*
(Intercept)	6.283	4.701 – 7.866	7.781	<0.001***
Classifier [Incongruent]	−0.772	−1.408 – −0.136	−2.379	0.017*
Semantic [Related]	−0.656	−1.336 – 0.023	−1.893	0.058
Location [left central parietal]	2.130	2.066 – 2.195	64.725	<0.001***
Location [left frontal central]	− 2.501	−2.552 – −2.450	−96.108	<0.001***
Location [right central parietal]	1.290	1.249 – 1.331	61.970	<0.001***
Location [right frontal central]	−1.405	−1.447 – −1.362	−64.528	<0.001***
Random Effects				
σ^2	128.42			
$\tau_{00\ \text{Participant}}$	17.14			
$\tau_{00\ \text{Item}}$	1.27			
$\tau_{11\ \text{participant. classifier-incongruent}}$	3.05			
$\tau_{11\ \text{participant. semantic-unrelated}}$	3.48			
$\rho_{01\ \text{participant. classifier-incongruent}}$	− 0.03			
$\rho_{01\ \text{participant. semantic-unrelated}}$	− 0.65			
ICC	0.11			
$N_{\text{participant}}$	29			
N_{Item}	21			
Observations	2,370,960			
Marginal R^2/Conditional R^2	0.017/0.128			

Note: The p-values presented in the *tab_model* output were derived from Wald tests, which may slightly differ from those reported in the main text, as the latter were computed using Satterthwaite's method for better accuracy in mixed-effects models.