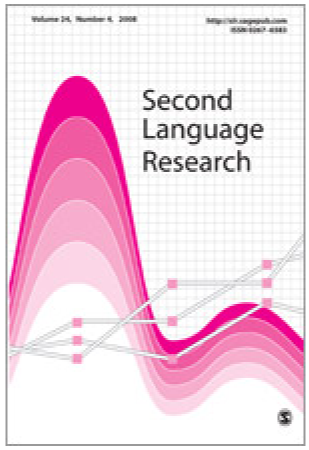


ERP sensitivity to subcategorization violations in second language learners

Journal:	*Second Language Research*
Manuscript ID	SLR-22-0194.R3
Manuscript Type:	Original Article
Keywords:	Verb subcategorization, Second language acquisition, Sentence processing, ERP, P600

Abstract:	Event concepts of common verbs (e.g., eat, sleep) can be broadly shared across languages, but a given language's rules for subcategorization are largely arbitrary and vary substantially across languages. When subcategorization information does not match between L1 and L2, how does this mismatch impact L2 speakers in real time? We hypothesized that subcategorization knowledge in L1 is particularly difficult for L2 speakers to override online. ERP responses were recorded from English sentences that include verbs that were ambitransitive in Mandarin but intransitive in English (*My sister listened the music). While L1 English speakers showed a prominent P600 effect to subcategorization violations, L2 English speakers whose L1 was Mandarin showed some sensitivity in offline responses but not in ERPs. This suggests that computing verb-argument relations, although seemingly one of the basic components of sentence comprehension, in fact requires accessing lexical syntax which may be vulnerable to L1 interference in L2. However, our exploratory analysis showed that more native-like behavioral accuracy was associated with a more native-like P600 effect, suggesting that with enough experience, L2 speakers can ultimately overcome this interference.

ERP sensitivity to subcategorization violations in second language learners

Abstract

Event concepts of common verbs (e.g., eat, sleep) can be broadly shared across languages, but a given language's rules for subcategorization are largely arbitrary and vary substantially across languages. When subcategorization information does not match between L1 and L2, how does this mismatch impact L2 speakers in real time? We hypothesized that subcategorization knowledge in L1 is particularly difficult for L2 speakers to override online. ERP responses were recorded from English sentences that include verbs that were ambitransitive in Mandarin but intransitive in English (**My sister listened the music*). While L1 English speakers showed a prominent P600 effect to subcategorization violations, L2 English speakers whose L1 was Mandarin showed some sensitivity in offline responses but not in ERPs. This suggests that computing verb-argument relations, although seemingly one of the basic components of sentence comprehension, in fact requires accessing lexical syntax which may be vulnerable to L1 interference in L2. However, our exploratory analysis showed that more native-like behavioral accuracy was associated with a more native-like P600 effect, suggesting that with enough experience, L2 speakers can ultimately overcome this interference.

Keywords: Verb subcategorization, Sentence processing, Second language acquisition, ERP, P600

1. Introduction

Second language (L2) acquisition has variable outcomes in adult learners. Many models have been proposed to discuss how neurocognitive mechanisms supporting L2 are similar to or different from that of the first language (L1) (Clahsen & Felser, 2006; Ullman, 2012, Steinhauer, 2014). Although the exact details of the models differ, they have generally agreed that the neural underpinnings of lexical/semantic processes could be qualitatively similar between L1 and L2. However, when it comes to rule-based grammatical processes, the results are less conclusive (Clahsen & Felser, 2006; Ullman, 2012). These studies provide an important overview of L2 acquisition, but the distinctions between lexical/semantic and rule-based grammatical computations could be oversimplified, as there are cases when structural information is encoded in a word. In the current study, we focus on lexical-syntax, investigating the processing of verb subcategorization information in L2. Can L2 speakers learn to rapidly access lexical-syntax information that differs from their L1 to parse a structurally simple sentence in L2?

Verb subcategorization refers to the grammatical restrictions that each verb places on the number and type of syntactic arguments that it appears with. The number of syntactic arguments grammatically required by a verb is linguistic knowledge, and it can differ from the number of entities conceptually required in an event. For example, English verb "steal" requires only a subject argument and an object argument, although stealing events involve a stealer, a victim, and a stolen item. Cross-linguistically, languages can differ in the subcategorization requirements of verbs that seem to refer to the same conceptual event (e.g., in the current study, whether *listen* takes as its argument a prepositional phrase headed by "to", as in "listen [to the music]," or a noun phrase argument as in "listen [the music]"). The relationship between a verb's grammatical subcategorization and meaning is not completely arbitrary. For example, it's unlikely that an event such as "sleep" that involves one participant will be realized with a verb that requires two syntactic arguments. However, because it is *somewhat* arbitrary, the learner cannot master grammatical subcategorization by acquiring any general rule; rather, grammatical subcategorization information has to be encoded for *individual* verbs of the language. Note that even basic structures of subcategorization can vary widely among languages, and these substantial typological differences could pose distinct challenges to learners of different language backgrounds. For instance, Mandarin features a highly productive system of compounding verbs, including coordinate verbs (e.g., X hit-scolded Y, meaning X hit and scolded Y) and resultative verbs (e.g., X bit-broke Y, meaning X bit Y such that Y was broken) (see Liao & Lau, 2020). To compute the representations of compounding verbs, the mechanism must combine the two verbs to form a single complex predicate and to derive the corresponding subcategorization constraints. Another example comes from restrictions on the coverb *ba* construction in Mandarin. According to Ye, Zhan and Zhou (2007), only verbs that denote a state of change can fit with the causation meaning of the *ba* construction. In these cases, for English speakers to learn Mandarin, they may not have any corollaries to their native rules to rely on, so they need to learn an entirely new system of subcategorization constraints. In brief, the complexity of subcategorization restrictions can vary across languages, and pose distinct challenges for language learners.

In the current study, we follow classic work from Odlen (1989) in defining transfer as "the influence resulting from similarities and differences between the target language and the language that has been acquired." Many studies have examined the impact of L1 on different aspects in L2 (for a systematic review, see Caffarra, Molinaroa, Davidsona, & Carreiras, 2015), but not many empirical studies have investigated L1 transfer on subcategorization knowledge in L2. Part of the reason may be Clahsen and Felser's influential work (2006) emphasizing that L2 speakers can successfully compute local verb-argument relations in canonical order even when they fail on more complex aspects of sentence structure. Since then, it is assumed that L2 speakers can correctly capture the thematic relations between arguments and verbs online, and research has been centered on L2 speakers' lack of hierarchical details and abstract components of syntactic structures. However, as verb-argument computation involves lexical syntax, evidence of native-like computation of verb-argument relations requires more than just getting the interpretation right. Specifically, while event concepts of common verbs are likely to be the same for speakers of different languages (e.g., eat, sleep), which arguments a verb subcategorizes for is linguistic knowledge, and could vary from language to language. When subcategorization information does not match between L1 and L2, how does this mismatch impact L2 speakers online?

The current study was designed to evaluate if online verb-argument computation is impacted by L1 transfer. In other words, subcategorization of verbs in L1 might have a significant impact on how learners parse sentences in L2. In the English example discussed earlier, the verb "listen" requires introducing the theme with a preposition ("My sister listened to the music"), but in Mandarin the corresponding verb does not ("My sister listened the music"). Therefore, English L1 speakers will reject sentences like "My sister listened the music" but those sentences might be accepted by Mandarin L2 speakers of English. Although the Revised Hierarchical Model (the RHM, Kroll & Stewart, 1994; Kroll, Van Hell, Tokowicz, & Green, 2010) is not uncontroversial (see Dufour & Kroll, 1995; Brysbaert & Duyck, 2010), it provides a useful framework for discussing how L2 speakers process subcategorization online (see Figure 1). According to this model, L2 comprehension drives strong and automatic associative retrieval of L1 lexical items. In framing the current study, we assume that after a verb is encountered, subsequent arguments are checked as they come in against the subcategorization information retrieved from the verb's lexical entry. The process of accessing the subcategorization information during online sentence comprehension could be more error-prone in L2, especially when the L1 verb has a substantially different subcategorization frame from the L2. For an L2 speaker, being able to run this check is dependent on having learned the subcategorization information for the verb. Even if subcategorization information has been acquired and constitutes part of the L2 speaker's knowledge, lexical association between L2 and L1 words could be so strong that L2 speakers might not able to immediately override the L1 information online.

[Figure 1 around here]

Figure 1: A schematic diagram of online processing upon reading a verb. The figure is adapted from the Revised Hierarchical Model (Kroll & Stewart, 1994).

1.1 Prior behavioral studies investigating the processing of subcategorization in L2

A handful of behavioral studies have investigated how processing of subcategorization is impacted in L2. Jiang (2007) asked whether Mandarin L2 speakers of English could detect subcategorization violations online in self-paced reading. Materials were sentences of a variety of structures, but the ungrammatical ones always involved a complement that the verb did not subcategorize for (e.g., "The mayor promised to offer/*keep the returning advisor a better position soon"). The results showed that the L2 had a similar processing profile as the L1. Although the reading time in the L2 was generally slower, both groups showed a slowdown at regions after the subcategorization violation occurred. Jiang (2007) thus argued that L2 speakers were sensitive to verb subcategorization errors online. However, Jiang (2007) did not address the discrepancies between L1 and L2 subcategorization frames. Therefore, although L2 speakers appeared to quickly detect subcategorization errors, these results leave the question open of whether native-like detection of subcategorization violations depends on facilitation through their L1.

Although Jiang's finding was (2007) was not conclusive about the impact from L1, several studies have suggested L1 transfer of subcategorization information by showing that subcategorization *preferences* could be carried over from L1 (Dussias & Scaltz, 2008; Dussias, Marful, Gerfen, & Bajo Monica, 2010). For example, Dussias, Marful, Gerfen, and Bajo Monica (2010) ran a norming task of 100 English verbs on late Spanish L2 speakers of English. Sentence frames, which contained only a subject and a verb, were given to the bilinguals. The authors looked at the structure of the continuations that the L2 speakers provided, and compared the results with the norming data collected from native English speakers in Garnsey, Pearlmutter, Myers and Lotocky (1997). The cross-study comparison showed that among the 100 verbs, 39 had between-group subcategorization differences. In addition, 10 of the 39 verbs showed a transfer effect from L2 speakers' L1. Although the proportion was not very high, the results suggested that L1 subcategorization could play a role in processing L2.

1.2 ERP studies of subcategorization violation in L1 and L2 speakers

EEG has high temporal resolution for tracking task-related computation online, and many previous studies have identified ERP responses that appear to be tied to the detection of subcategorization violations in native speakers. Although there is some variability, such violations have generally elicited N400 and P600 responses in ERP. Below we will first introduce the functional interpretations of the two ERP components, and then discuss how the L1 and L2 speakers process subcategorization violations in ERP. To begin with, the N400 response peaks between 300-600 ms after the onset of the stimulus presentation over the central-parietal sites (Kutas & Hillyard, 1984). N400 is suggested to reflect pre-activating conceptual features and/or pre-activating specific lexical items of an upcoming word (Brothers, Swaab, & Traxler, 2015, Thornhill &

Van Petten, 2012). As for the P600 response, it lacks a clear peak, but is usually more prominent in the 500-900 ms interval after the onset of a problematic word over the parietal sites (Osterhout & Holcomb, 1992; Hagoort, Brown, & Groothusen, 1993). The P600 effect has often been associated with grammaticality violations (Friederici & Frisch, 2000; Hagoort, Brown, & Groothusen, 1993; Silva, Folia, Hagoort, & Petersson, 2017) or structural complications (Gouvea, Phillips, Kazanina, & Poeppel, 2010; Ueno & Garnsey, 2008); other studies suggest that it is sensitive to rule-governed sequences, and could be a domain-general response (Hagoort & Brown, 2000; Koelsch & Jentschke, 2010; Sitnikova, Holcomb, Kiyonaga, & Kuperberg, 2008; Cohn & Kutas, 2015). Here, we will consider the P600 an index of difficulty in structure analysis (Kaan & Swaab, 2003; Tanner, Grey, & van Hell, 2017).

Friederici and Frisch (2000) was an early landmark study of the response to argument structure and subcategorization violations in ERP, which manipulated two kinds of mismatches between the verb and its arguments in German: Argument number and argument type. In their Experiment 1, two arguments preceded the verb ("Anna knows that the inspector$_{NOM}$ the banker$_{ACC}$ monitored…"). Argument number mismatches were created by substituting a sentence-final transitive verb with an intransitive one, such that the second argument became unlicensed (* "Anna knows that the inspector$_{NOM}$ the banker$_{ACC}$ departed…"). By contrast, argument type mismatches were created by replacing the verb that assigns accusative case with one that assigns dative case to the object (*"Anna knows that the inspector$_{NOM}$ the banker$_{ACC}$ helped…"). ERP responses were time-locked to the onset of the verb. The ERP results showed that compared with canonical sentences, argument number mismatches elicited an N400-P600 biphasic response, and argument type mismatches elicited a LAN-P600 response. As German has relatively free word order, in Experiment 2, they placed a verb before two arguments ("Today visited the cousin$_{NOM}$ the violinist$_{ACC}$ in the hospital"), and then manipulated the same kinds of mismatches between the verb and its arguments as in Experiment 1. ERP responses were time-locked to the onset of the second argument. Similar to Experiment 1, they found that compared with canonical sentences, argument number mismatches elicited an N400-P600 biphasic response, while here argument type mismatches only elicited a P600 effect. Similar results for argument number mismatches were observed as part of a Dutch study aimed at investigating the P600 across modalities (Hagoort & Brown, 2000), where they reported a P600 effect for these mismatches in both auditory and visual presentation, preceded by a small, frontal-central negativity, larger with auditory presentation. Kielar, Meltzer-Asscher and Thompson (2012) examined similar kinds of subcategorization violations of intransitivity in English ("John visited/*sneezed the doctor"). They reported an N400-P600 effect to the determiner and the noun in both young and healthy elder adults, but only a P600 effect in agrammatic aphasia patients. In a different kind of subcategorization violation paradigm, Osterhout, Holcomb and Swinney (1994) studied the ERP responses to an auxiliary verb in a subordinate clause, which was introduced either by a complement verb or by a transitive verb that did not take a finite complement clause ("The doctor hoped/*forced the patient was lying"). They also found an N400-P600 biphasic response at the auxiliary verb.

How should we understand the ERP responses to subcategorization violations? Studies discussed above all reported a P600 effect, and they all considered it a reflection of syntactic processes that involve reanalysis and repair (Friederici & Frisch, 2000; Hagoort & Brown 2000; Kielar, Meltzer-Asscher, & Thompson, 2012; Osterhout, Holcomb, & Swinney, 1994). The negativity occurred before P600 varied in terms of strength and distributions. Hagoort and Brown (2000) noticed that in their data, the distribution of the negativity was more prominent over left frontal cites. They suspected that this effect was a LAN, which reflected word category violations (Friederici, Hahne, & Mecklinger 1996). Different from Hagoort and Brown (2000), other studies found that the distribution of their negativity was more widespread, and was more prominent over the central-parietal cites, they thus interpreted it as an N400 (Friederici & Frisch, 2000; Kielar, Meltzer-Asscher, & Thompson, 2012; Osterhout, Holcomb, & Swinney, 1994). Friederici and Frisch (2000) treated this as an indication of difficulties in integrating lexical information into the context, but for Kielar, Meltzer-Asscher, and Thompson (2012), it reflected problems in lexical access. From our perspective, if the N400 effect was real, it could also be interpreted as encountering a word whose conceptual and semantic features were not pre-activated. Consider the example "John visited/*sneezed the doctor" from Kielar, Meltzer-Asscher, and Thompson (2012), given the intransitive verb "sneezed," comprehenders would not expect an argument to come next. By contrast, by getting a transitive verb "visited," the parser may start to expect an argument, and pre-activate features that could be related to that argument. It may be for this reason that the N400 amplitude to the subsequent noun phrase was larger in the intransitive context relative to the transitive one.

To our knowledge, only a handful of studies have investigated L2 subcategorization violation with EEG. Karatas (2019), in a manipulation of lexical case violations in Turkish, reported a widespread negativity and a P600 effect on the verb at which the violation can be detected in the L1 group, but the P600 effect was not observed in advanced L2 speakers. Guo, Guo, Yan, Jiang, and Peng (2009) asked a different question than Karatas (2019); they wondered if L2 speakers relied on a lexical-semantic strategy to process problematic syntactic structures such as subcategorization violations. In a series of prior studies, Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, and Molinaro (2006) showed that L2 speakers often demonstrate N400 effects to syntactic violations such as agreement that would normally elicit P600 effects in L1 speakers. Therefore, Guo et al. (2009) predicted that L2 speakers would show an N400 effect to subcategorization violations, whereas the L1 group a P600 effect. They adapted the materials from Jiang (2007) as discussed in the section above. In both studies, they compared verbs of different subcategorization frames, with the ungrammatical sentences always involving a complement that the verb did not subcatecorize for (e.g., "The mayor promised to offer/*keep the returning advisor a better position soon"). ERP responses were time-locked at a critical word at which subcategorization violation can be detected, which collapsed across a variety of lexical categories, including a noun phrase, a verb infinitive, a prepositional phrase, or an adjective. The results confirmed their prediction: While L1 speakers showed a P600 effect to subcategorization violations, Mandarin L2 speakers of English showed an N400 effect. However, it is worth noting that, as acknowledged by the authors, the negativity obtained in L2 speakers was only prominent over lateral electrodes, which was very different than a traditional central-parietal N400 distribution. The results of Guo et al. (2009), did not speak directly to our main question about whether subcategorization restrictions in L1 interfere with L2 verb-argument computation, because like Jiang (2007), Guo et al. (2009) did not consider subcategorization restrictions between L1 and L2 in constructing the materials and analyzing the results.

1.3 The current study

Our goal was to test if L2 verb-argument computation is subject to substantial L1 transfer online. Here we provided a stronger case for evaluating subcategorization effects than Guo et al. (2009), as we took subcategorization restrictions between L1 and L2 into consideration, and the lexical category of target words remained the same across all trials. We chose to look at the case of L1 Mandarin and L2 English, because there is a considerable number of English intransitive verbs whose Mandarin translation could be either transitive or intransitive (i.e., ambitransitive). For example, *listened the music* is acceptable in Mandarin but unacceptable in English. If L2 speakers are sensitive to English subcategory restriction online, they should show an ERP violation response parallel to native speakers when the unacceptable noun phrase is encountered. We predict to obtain a P600 effect to subcategorization violations in L1, because this result has been very reliable across studies (Friederici & Frisch, 2000; Hagoort & Brown, 2000; Kielar, Meltzer-Asscher, & Thompson, 2012; Osterhout, Holcomb, & Swinney, 1994). By contrast, the negativity observed before P600 in prior studies varied in strength and distribution, and could be dependent on the contextual expectations afforded by the design. Although for this reason the earlier response appears less ideal for testing the transfer hypothesis, we include this "N400" window in our analysis to provide a point of comparison with prior research.

One concern in L2 research aimed at a specific element of processing, is that the L2 may simply show insensitivity or non-native-like responses to all of the experimental manipulations, raising questions about the specificity of the results. Therefore, to show that our L2 speakers of English were able to parse English sentences online and recognize grammatical violations that are not subject to L1 transfer, we included a control comparison in the experiment. Prior work has showed that Mandarin L2 speakers were sensitive to phrase structure violations in English, such as "a proof <u>of the</u> theorem" vs. "Max's <u>of proof</u> the theorem" (Weber-Fox & Neville, 1996). In particular, those who were immersed in an English-speaking environment before puberty, showed a native-like LAN-P600 effect to phrase structure violations. By contrast, those who were exposed to English at an older age showed a wide-spread negativity in an early window, and the P600 effect was much diminished. Since L2 speakers of different learning conditions can detect phrase structure violations, we included such contrasts as a control comparison in the current study, and predicted the L2 to show ERP sensitivity to the violation on any account.

2. Methods

2.1 Participants

21 native English speakers (9 females, M_{age}: 20.4, range: 18-26) and 30 Mandarin L2 learners of English (19 females, M_{age}: 23.3, range: 19-32) participated in the study. English native speakers were recruited in the US, and none had ever been exposed to Mandarin before. L2 English learners were recruited in Taiwan. On average, the L2 speakers reported starting to learn English around age 7 (SD = 2). None of them had been exposed to an English-only environment for studying English. All of them considered themselves proficient in English, with the following self-reported English proficiency in different skills (1 = not fluent at all; 7 = very fluent): Listening: 5.6 (SD = 0.7); Speaking: 5.2 (SD = 1.0); Reading: 5.7 (SD = 0.9); Writing: 5.1 (SD = 0.8). As a more objective measure of proficiency, all participants had passed a standardized English proficiency test beyond the intermediate-advanced level, or the B2 level of the Common European Framework of Reference (CEFR). According to the CEFR, "(learners at this level) can achieve most goals in the target language, and they can express themselves on a range of topics." Both groups of participants were right-handed and did not have a history of neurological or psychiatric disorders. All of them consented to participate in the experiment. The experiment protocol was approved by the Institutional Review Board Office at a university in the US.

2.2 Materials

The critical subcategorization stimuli were sentences of Subject-Verb-Object structure, with the verb being varied between two conditions: (1) Grammatical and (2) Ungrammatical subcategorization. Verbs in Grammatical condition were transitive in both English and Mandarin (e.g., record), whereas verbs in Ungrammatical condition usually do not take a direct object in English but are ambitransitive in Mandarin (e.g., listen). Note that although verbs in Ungrammatical condition were intransitive, they can introduce a subsequent argument with the insertion of a preposition (e.g., listen to the music). The selection of the verbs was based on the intuition of the first author (a Mandarin native speaker with English as her second language), cross-checked with the online Cambridge English Dictionary (https://dictionary.cambridge.org/zht/) and another Mandarin native speaker. We matched the lexical frequency (Grammatical/Transitive: 17491, Ungrammatical/Intransitive: 17300; $t(59) = .46$, $p = .65$) and word length of verbs (Grammatical/Transitive: 6, Ungrammatical/Intransitive: 6; $t(59) = .55$, $p = .58$) between conditions. Except for the verbs, the rest of the sentences remained identical. Sixty pairs of critical sentences were created, and were proofread and edited by three native English speakers. To ensure that not all sentences with an intransitive English verb were ungrammatical and vice versa, we added two filler conditions with grammatical intransitive verbs and with ungrammatical transitive verbs (Table 1). Each filler condition consisted of 30 sentences. The materials are available online at https://figshare.com/s/9a4cabc0dfba71ca71c5.

Table 1: Example stimuli in each condition

Conditions		**Example stimuli**
Subcategorization	**Grammatical**	My sister recorded <u>the music</u> I played.
	Ungrammatical	My sister listened <u>the music</u> I played.

Conditions		Example stimuli
Phrase structure	**Grammatical**	The scientists criticized Max's proof of the theorem.
	Ungrammatical	The scientists criticized Max's of proof the theorem.
Filler	**Grammatical**	The singer sneezed during the concert.
	Ungrammatical	The leader should impose by next week.

To show that L2 speakers were able to parse English sentences, we adapted sentences with phrase structure violations from Neville, Nicol, Barss, Forster, and Garrett (1991) as our control items. In particular, Weber-Fox and Neville (1996) reported that L2 speakers, even with delayed exposure to English (above age 11), were able to show ERP sensitivity to phrase structure violation sentences ("The scientists criticized a proof of the theorem" vs. "*The scientists criticized Max's of proof the theorem"). We slightly revised the "standard" sentences by replacing "a proof" with "Max's proof" ("The scientists criticized Max's proof of the theorem"), to ensure that not all the proper names occurred in an ungrammatical context. We adapted 30 sentences from Neville et al. (1991), and wrote another 30 sentences of the same structure in order to create 60 pairs of sentences for a controlled comparison.

Two experiment lists were constructed such that no sentence context or critical verb was repeated within the same list. Each list consisted of 60 subcategorization sentences (30 Grammatical and 30 Ungrammatical), 60 filler sentences and 60 phrase structure sentences (30 Grammatical and 30 Ungrammatical). The presentation order was randomized within each list. Participants were randomly assigned to one of the two lists.

2.3 Procedure

Participants sat in front of a computer screen and put their hands on a keyboard.

Sentences were presented one word at a time in a black font on a white background at the center of the screen. Each sentence was preceded by a fixation cross that appeared for 600 ms. Each word appeared on the screen for 400 ms, with a 200-ms inter-stimulus interval, for a stimulus-onset asynchrony (SOA) of 600 ms. At the end of each sentence, participants performed a grammaticality judgment task via button pressing. Prior to the experimental session, participants were presented with six practice trials with feedback to familiarize themselves with the task. Including set-up time, an experimental session lasted around 90 minutes.

2.4 Data acquisition and analysis

The L1 data were collected in the US. We used Matlab to present the experiment stimuli, record participants' behavioral data, and send the event codes to the digitization computer (MathWorks, Incorporated). EEG was recorded from 29 tin electrodes mounted in an electronic cap (Electro-Cap International Incorporated) according to the 10/20 system (FP1, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FCZ, FC4, FT8, T7, C3, CZ, C4, T8, TP7, CP3, CPZ, CP4, TP8, P7, P3, PZ, P4, P8, O1, O2), with an additional electrode placed anterior to Fz as ground. Bipolar electrodes were placed above and below the left eye, and on the outer canthus of each eye to monitor blinks and horizontal eye movements. Each electrode was referenced to the right mastoid online and re-referenced to the average of the left and right mastoids offline. The impedance of all the electrodes was kept below 10 kΩ. The EEG signals were amplified by a SynAmps amplifier (Model 5083, NeuroScan Incorporated) with a bandpass of 0.05-100 Hz and was continuously sampled at 500 Hz during online data recording.

The L2 data were collected in Taiwan. We used E-prime 2.0 (Psychology Software Tools Incorporated) to present the experiment stimuli, record participants' behavioral data, and send the event codes to the digitization computer. EEG was recorded from 30 Ag/AgCl electrodes mounted in an electronic cap (Quick-cap, Neuroscan Incorporated) according to the 10/20 system: the same 29 scalp positions as in the L1 cap, plus OZ. The ground electrode was also placed anterior to Fz. Similar to the L1 group, bipolar electrodes were placed to monitor blinks and horizontal eye movements. Each electrode was referenced to the average of the left and right mastoids online and offline. Other subject preparation criteria were kept the same as the L1 group. The EEG signals were amplified by a NuAmps amplifier (NuAmps, NeuroScan Incorporated) with a bandpass of 0.05-100 Hz and was continuously sampled at 500 Hz during online data recording.

The EEG data were processed with EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) in Matlab (MathWorks, Inc.). A linear derivation file was first imported to convert the four monopolar eye-movement monitoring channels to two bipolar channels (VEOG and HEOG). We applied a notch filter at 60 Hz and an Infinite Impulse Response (IIR) filter with the band-pass value set between 0.1 to 30 Hz, 12 dB/oct. Then we extracted epochs of length -100 to 1200 ms, from the onset of the determiner until the end of the noun phrase for Subcategorization conditions, and from the onset of the preposition to the next word for Phrase structure conditions. Baseline correction was applied with the pre-stimulus -100 to 0 ms interval. After baseline correction, artifact rejection was carried out by reviewing the epochs both automatically and manually. At each electrode, a 200-ms window was moved across the entire epoch in 100-ms increments and any epoch where the peak-to-peak voltage exceeded 70 μV was rejected. We then reviewed the data, and if needed, adjusted the voltage threshold for individual subjects. Epochs contaminated by excessive blinking, body movements, skin potentials, and amplifier saturation were rejected. The overall rejection rates were 21±16% for the L1 group and 11±11% for the

L2 group (mean±SD); participants with greater than 40% trials rejected were excluded from further analysis.

We focused our hypotheses on the ERP responses occurring within the 300-500 ms and 600-900 ms time intervals, specifically examining the determiner in the case of Subcategorization conditions and the preposition in the case of Phrase structure conditions. We selected six electrodes over the frontal area (F3, FZ, F4, FC3, FCZ, FC4), six electrodes over the parietal area (CP3, CPZ, CP4, P3, PZ, P4) and averaged them as our clustered region of interest (ROI) for both Subcategorization and Phrase structure comparisons. Analyses were conducted using mixed-effects models with crossed random effects for subjects and items using the lme4 package (version 1.1-31) of R (version 4.2.0). For Subcategorization conditions, the analysis included contrast coded fixed effects for Subcategorization conditions (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front) and Group (-.5 = L2 group, .5 = L1 group) in a 2x2x2 factorial design. Random effects were fit using a maximal random effects structure (Barr, 2013), including random intercepts for subjects and items, by-subject random slopes for Subcategorization, Region and their interaction, and by-item random slope for group. For Phrase structure conditions, the analysis included contrast coded fixed effects for Phrase structure (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front) and Group (-.5 = L2 group, .5 = L1 group) in a 2x2x2 factorial design. Random effects include random intercepts for subjects and items, by-subject random slopes for Phrase structure, Region and their interaction, and by-item random slope for group. In both cases, models were fit using a maximum likelihood technique. A fixed effect was considered significant if the absolute value of the t was greater than or equal to 2.0 (Gelman & Hill, 2007).

2.5 Results

2.5.1 Behavioral data

The overall accuracy rate of the grammaticality judgment task was 92% for the L1 group, and 81% for the L2 group, showing that participants were paying attention in the experiment. The accuracy rate of each condition in the two groups is reported in Figure 2. The D-prime score of each factor is summarized in Table 2.

[Figure 2 around here]

Figure 2: Accuracy rate and standard deviation of each condition in L1 and L2 groups

Table 2: Mean D-prime score for each factor in L1 and L2 groups, range in parentheses

	Subcategorization	Phrase structure	Filler
L1	2.88 (1.48 – 4.02)	3.75 (3.16 – 4.28)	3.09 (1.55 – 4.28)
L2	1.39 (0 - 3.26)	3.06 (1.17 – 4.28)	2.22 (0.7 – 3.42)

Since Filler conditions were not the focus of the current study, below we will focus on the contrast between Subcategorization and Phrase structure. To examine whether there were differences in accuracy rates between L1 and L2 for the two manipulations, we conducted a Type III repeated measure ANOVA with the within-subject factors of Sentence type (Subcategorization, Phrase structure), Grammaticality (Grammatical, Ungrammatical) and between-subject factor of Group (L1, L2). We also conducted a separate Type III repeated measure ANOVA with the within-item factors of Grammaticality (Grammatical, Ungrammatical) and Group (L1, L2), and between-item factor of Sentence type (Subcategorization, Phrase structure) to account for any potential variability across items. We did observe a three-way interaction among Sentence type, Grammaticality and Group ($F_1(1,49) = 32.785$, $p < .001$, $\eta^2_p = .401$;

$F_2(1,118) = 59.508$, $p < .001$; $\eta^2_p = .335$). Follow-up analyses showed a Grammaticality by Group interaction for Subcategorization conditions ($F_1(1,49) = 33.563$, $p <.001$, $\eta^2_p = .407$; $F_2(1,59) = 91.967$, $p <.001$, $\eta^2_p = .609$). Pair-wise comparisons revealed that the accuracy rates for Grammatical and Ungrammatical conditions differed significantly in both groups (L1: $t_1(20) = 3.23$, $p < .05$, $t_2(59) = 3.493$, $p <.05$; L2: $t_1(29) = 10.234$, $p <.005$, $t_2(59) = 11.884$, $p <.005$). The accuracy rates for Ungrammatical conditions also differed significantly between groups ($t_1(49) = 6.575$, $p <.005$, $t_2(59) = 13.858$, $p <.005$). However, the subject-level and item-level analyses revealed different patterns in Grammatical conditions ($t_1(49) = 2.791$, $p <.05$, $t_2(59) = 2.121$, $p = 0.152$). For Phrase structure conditions, we only observed a main effect of Group ($F_1(1,49) = 14.086$, $p <.001$, $\eta^2_p = .223$; $F_2(1,59) = 33.322$, $p <.001$, $\eta^2_p = .361$), which did not interact with Grammaticality ($F_1(1,49) = .526$, $p = .472$, $\eta^2_p = .011$; $F_2(1,59) = 0.082$, $p =.766$, $\eta^2_p = .001$).

Overall, the results suggested that Subcategorization was particularly difficult to the L2 group. The d-prime scores indicated that some L2 participants displayed lower sensitivity to Subcategorizations. Consistent with prior work, we found a ‘yes bias’: both the L1 and L2 exhibited a higher accuracy rate in accepting the Grammatical condition compared to rejecting Ungrammatical condition.

2.5.2 ERP data

2.5.2.1 Verb subcategorization violations

Figures 3A and 3B show the grand average ERPs from the determiner to the noun to Subcategorization conditions for the L1 and L2 groups. Visual inspection suggested that there was no N400 difference to the determiner in both groups. At approximately 500 ms after the onset of the determiner, there was a prominent P600 effect in the L1, while there was little difference between conditions in the L2.

[Figure 3A around here]

Figure 3A: Top: Grand average ERPs of Subcategorization conditions from the determiner to the nouns at representative electrodes in L1 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the determiner in L1 speakers.

[Figure 3B around here]

Figure 3B: Top: Grand average ERPs of Subcategorization conditions from the determiner to the nouns at representative electrodes in L2 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the determiner in L2 speakers.

During the 300-500 ms time window, we did not obtain any significant group interactions (Subcategorization x Group x Region: Estimate = -.16, *SE* = .82, *t* = -.19; Subcategorization x Group: Estimate = .44, *SE* = .56, *t* = .79), nor did we find a significant main effect (Subcategorization: Estimate = .09, *SE* = .28, *t* = .34; Group: Estimate = .68, *SE* = .61, *t* = 1.12). The statistics confirmed the visual inspection that there was no N400 effect in both groups. The results of the model are reported in Supplementary file (Table 1).

During the 600-900 ms time window, the mixed effects analysis demonstrated a significant Subcategorization by Group interaction (Estimate = -1.46, *SE* = .68, *t* = -2.13). Follow-up pairwise analyses revealed that that there was a significant difference between Grammatical and Ungrammatical in the L1 ($t(20) = -2.653$, $p < .05$, $d = -.579$, 95% CI for d [-1.036, -.109]), but not the L2 ($t(29) = -.885$, $p = .384$, $d = -.161$, 95% CI for d [-.520, .200]). We did not obtain a three-way interaction among Subcategorization x Group x Region (Estimate = .58, *SE* = .85, *t* = .68), suggesting that the P600 effect in the L1 was a widespread effect. The results of the model are reported in Supplementary file (Table 2).

2.5.2.2 Phrase structure violations

Figures 4A and 4B show the grand average ERPs from the preposition to the following word to Phrase structure conditions for the L1 and L2 groups. Visual inspection suggested that there was a widespread negativity to the preposition during the 300-500 ms time window for both L1 and L2 groups. In the later 600-900 ms time window, only the L1 showed a P600 deflection. Below we report statistical tests for the two windows.

[Figure 4A around here]

Figure 4A: Top: Grand average ERPs of Phrase structure conditions from the preposition to the following word at representative electrodes in L1 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the preposition in L1 speakers.

[Figure 4B around here]

Figure 4B: Top: Grand average ERPs of Phrase structure conditions from the preposition to the following word at representative electrodes in L2 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the preposition in L2 speakers.

During 300-500 ms, the mixed effects analysis showed a main effect of Phrase structure (Estimate = 1.81, *SE* = .35, *t* = 5.28), which was not modulated by Group (Estimate = -.79, *SE* = .69, *t* = -1.14), nor by Region (Estimate = -.29, *SE* = .42, *t* = -.69). We did not obtain a Phrase structure x Group x Region interaction either (Estimate = .62, *SE* = .84, *t* = .74). The results confirmed our visual inspection that both groups elicited a widespread negativity during this time window. The results of the model are reported in Supplementary file (Table 3).

During 600-900 ms, the mixed effects analysis demonstrated a significant Phrase structure by Group interaction (Estimate = -2.62, *SE* = .71, *t* = -3.71). Follow-up pairwise analyses revealed that that there was a significant difference between Grammatical and Ungrammatical in the L1 (t(20) = -4.231, p < .001, d = -.886, 95% CI for d [-1.386, -.371]), but not the L2 (t(29) = -.520, p = .607, d = -.095, 95% CI for d [-.453, .265]). We did not obtain a three-way interaction among Phrase structure x Group x Region (Estimate = .73, *SE* = .84, *t* = .87), suggesting that the P600 effect in the L1 was a widespread effect. The results of the model are reported in Supplementary file (Table 4).

2.5.3 Exploratory analyses

The focus of the current study was whether L2 learners could detect subcategorization violations in online comprehension. Given the great variation in L2 participants' behavioral performance, it was not surprising that manipulating Subcategorization did not elicit a P600 effect in the L2 group. However, previous studies suggest proficiency is an important factor modulating the P600 contrast (Tanner, McLaughlin, Herschensohn, & Osterhout, 2013; Steinhauer, White, & Drury, 2009; Mickan & Lemhofer, 2020). Therefore, we conducted an exploratory analysis on whether participants' language proficiency could predict the P600 effect over the parietal area in Subcategorization conditions. We used overall behavioral accuracy rate as an objective measure of language proficiency. The Pearson product-moment correlation coefficient showed a positive correlation between them (r = .329, p < .05, see Figure 5), suggesting that as L2 speakers' proficiency improves, it is possible for them to show native-like P600 sensitivity to subcategorization violations, even if the subcategorization constraints of the verbs conflict with those in their L1.

[Figure 5 around here]

Figure 5: Correlation between participant's behavioral accuracy rate and the P600 contrast over the parietal area in Subcategorization.

As a supplement to the results of correlation analysis, below we present the averaged ERPs of Subcategorization conditions in the L2 group, divided by a median split of accuracy. This additional step allows us to visualize the potential differences between high-accuracy and low-accuracy groups. As can be seen in Figure 6, Ungrammatical and Grammatical conditions did not differ in the low-accuracy group. By contrast, there was a numerical tendency towards a P600 contrast in the high-accuracy group.

[Figure 6 around here]

Figure 6: Grand average ERPs of Subcategorization in the L2 group, divided by a median split of accuracy. Left: Low-accuracy group (N =14); Right: High-accuracy group (N=16).

3 General Discussion

The current study investigated if L1 subcategorization knowledge impacts verb-argument computation in L2. Although previous studies suggested that L2 speakers could compute verb-argument relations quickly online, we proposed that such computations could be subject to L1 transfer. To test the hypothesis, we selected verbs that were ambitransitive in L1 Mandarin but intransitive in L2 English. We focused on ERP responses to the argument immediately following the verb and predicted that L2 speakers would be insensitive to subcategorization violation at the argument.

3.1 ERP responses to subcategorization violations in L1 and L2

Our results showed that subcategorization manipulations elicited a P600 effect in L1, but this effect was absent in the L2. The L1 response was consistent with many existing studies in L1 subcategorization violations (Osterhout, Holcomb, & Swinney, 1994; Friederici & Frisch, 2000; Guo et al., 2009). When the L2 subcategorization was in conflict with the L1, our data revealed that the L2 speakers were not sensitive to the problem online. Specifically, the critical verbs in our experiment were intransitive in English (L2) but could be transitive in Mandarin (L1). The L2 speakers, despite their proficiency being above the intermediate level, appeared to parse it as a transitive verb, such that the ERP responses to Ungrammatical condition did not differ from Grammatical baseline. Our null effect finding for L2 subcategorization violations may seem to conflict with Guo et al. (2009), who reported an N400 effect in their L2 group. Based on these results, the authors concluded that L2 speakers used a semantic strategy to process sentences online. One important difference between these studies is that Guo et al. (2009) did not manipulate L1-L2 verb subcategorization discrepancies. If many of the L2 items used in their study shared the same subcategorization restrictions between L1 and L2, then this overlap could have allowed more immediate detection of subcategorization violations in the L2, in contrast to the current study where the verbs were always ambitransitive in the L1 but intransitive in the L2.

There are two classes of explanations for the observed L2 insensitivity: non-native knowledge, and/or non-native processing. In particular, it could be that these L2 speakers did not have complete knowledge of L2 verb subcategorization, and thus to compute the verb-argument relations online, they mainly relied on L1 subcategorization knowledge. Alternatively or additionally, it could be that these speakers did have native-like L2 verb subcategorization knowledge, but during processing, they had difficulties in inhibiting L1 subcategorization information in time. Therefore, when the subcategorization frames of the verbs in the two languages competed, the effect of L1 would be observed earlier. Below we first walk through the incremental process of how

L1 speakers process subcategorization violations, and then we turn to the processing profiles of the L2 speakers.

Let's first consider the steps that L1 speakers go through to process verb-argument relations. When reading the sentence "My sister listened ____", L1 English speakers can quickly identify the verb "listen" as an intransitive verb, and access its subcategorization information, which is encoded in the verb. At this point, all the open syntactic dependencies in the sentence are completed. Therefore, one possibility is that when the subcategorization violation is encountered at the noun phrase ("the music"), the processing problem that the comprehender is faced with is the fact that there is no attachment site for the noun phrase. They attempt to reanalyze the sentence to see whether they made an error (e.g., was there any alternative subcategorization frame in the verb's lexical entry?) and this reanalysis process generates the P600. An alternative possibility is that L1 speakers take the absence of a period after the intransitive verb as a cue to do further processing: since they know the sentence will continue, they generate an expectation of the most likely continuation. Although a number of categories are likely (adverb, coordinator, preposition…), the statistics of the experimental items might bias towards the preposition expectation. In this case, the problem in Ungrammatical condition is not just that there is no attachment site for the noun phrase, but that the determiner violates the *prediction* for a preposition. The effort to reanalyze the structure of the sentence and/or understand why their expectation was violated would generate the P600.

Then we can turn to the question of why and how for our L2 speakers, L1 Mandarin knowledge could interfere with processing of these violations. The great variation in behavioral performance from the Subcategorization conditions indicates that subcategorization knowledge was certainly far from native-like for some L2 speakers in the current data set. It is possible that some of the L2 lexical items might not have been completely learned in the first place. Still, even when they have such knowledge, we suggest that when encountering the L2 wordform ("listen") during online processing, L2 speakers immediately (but unconsciously) map it to the L1 lexical entry. All the lexical information from the L1 entry thus becomes available, including its subcategorization information, and can thus be (erroneously) incorporated into the current parse of the L2 sentence. For example, instead of accessing only an intransitive L2 subcategorization frame at "listen" and concluding that all dependencies had been fulfilled, the L2 speakers may have distributed syntactic predictions across both the transitive and intransitive possibilities in their L1 and L2 entries, such that a noun phrase object was predictively projected with some probability. Therefore, when a subsequent noun phrase was presented ("the music"), it would then be slotted into this object position, predicted based on the L1 subcategorization frame. This could lead to the null ERP effect in the L2 speakers in the current study.

Our exploratory analysis suggested that as the L2 speakers' proficiency improves, they may be able to override the influence from L1 and to recognize the subcategorization violations in a native-like timescale. Similar findings are reported in White, Genesee and Steinhauer (2012), which showed that highly proficient L2 speakers elicited a larger P600 to grammatical errors, especially in correct (but not incorrect) behavioral responses. This line of research finding supports many studies that aim at delineating the developmental stages of L2 syntactic processing, including Tanner, McLaughlin, Herschensohn, and Osterhout (2013), Steinhauer, White, and Drury (2009), and Mickan and Lemhofer (2020). Although the exact details differ across studies, they all argued that L2 speakers progress through different stages of learning: At an early stage, L2 speakers tended to focus more on lexical semantics during sentence processing. As their proficiency improved, they were able to compute more complicated grammatical rules online.

One possibility suggested by a reviewer was that L2 speakers might have the subcategorization knowledge that a verb requires *some* preposition, but the difficulty is in which preposition they associate with the subcategorization frame, as which preposition is used to introduce an argument is notoriously variable cross-linguistically. Although it is true that correctly associating prepositions with subcategorization frames is difficult for L2 speakers, and that it is not uncommon for L2 speakers to make prepositional mistakes (e.g., *"punch on the face" vs. "punch in the face), in the current study, all the subcategorization violations were intransitive verbs *missing* a preposition. If L2 speakers did associate some preposition with the verb, regardless of whether it was the right one or the wrong one (e.g. "listen in the music"), a lexical category violation effect should have been observed in the subcategorization violation condition, where L2 speakers predicted to see a preposition, but were presented with a determiner ("listened the music). Given that they did not show this sensitivity, associating the wrong preposition seems unlikely to be the source of the non-native like responses in the current study. Future work, however, could further examine whether L2 speakers notice the wrong preposition in subcategorization violations like "listened to the music" vs. "listened in the music."

We note that it has been debated to what degree L2 ultimate attainment can be native-like. Clahsen and Felser (2006) took a more pessimistic stand, proposing that syntactic representations constructed by L2 speakers are shallower and contain less structural detail. Interestingly, although Clahsen and Felser (2006) argued that it was challenging for L2 speakers to construct hierarchical details and more abstract elements of sentence structures, they believed that L2 speakers could compute verb-argument relations with ease. The subcategorization case studied here reminds us that to some extent, verb-argument relations still depend on language-specific syntactic knowledge prone to L1 transfer and therefore might not be completely straightforward for L2 speakers either. However, if their proficiency improved, they would be able to select the L2 subcategorization frames more quickly, and perform more native-like computations.

3.2 ERP responses to Phrase structure violations in L1 and L2

We included Phrase structure violations adapted from Neville et al. (1991) and Weber-Fox and Neville (1996) as our control comparison. Consistent with these studies, we found a P600 effect in the L1 but not in the L2 group. Slightly different than these studies, the wide-spread negativity observed in our early time window for both groups appears to resemble an N400 effect rather than an LAN effect. Note that in the phrase structure violation used by Neville et al. (1991) ("Max's proof <u>of the</u> theorem" vs. "Max's <u>of proof</u> the theorem") has since been criticized for confounding sentence position and different pre-target word baseline (Steinhauer & Drury, 2012). Meta-analyses by Caffarra, Mendoza, and Davidson (2019) suggest that in certain instances, the overall LAN effect can be modulated by individual participants, items, and trials. With respect to the current study, we remain neutral about whether the negativity observed here was an N400 or LAN effect, and what exactly should be the functional interpretation of this response. Our primary reason for including phrase structure violations was to set up a control comparison; the fact that both groups showed the same early response to phrase structure violations allows us to rule out several possible alternative explanations for the null effect of subcategorization violations in the L2 group: It is not the case that the L2 participants were insensitive to all kinds of grammatical violations online, nor that we were unable to record grammatically sensitive ERP responses from that group. Therefore, the lack of sensitivity to L2 subcategorization manipulations cannot be attributed to general insensitivities of grammatical violations or a failure to measure grammatical-related neural responses from the L2. To avoid concerns about the baseline problem, future work could adopt a different control comparison that fully crosses critical words and sentence context ("They wanted to leave/*about yesterday"; "She was thinking about/*leave you yesterday").

Although we did not observe a P600 effect in the L2 group, we noticed that the size of P600 effect was modulated by age of L2 exposure in Weber-Fox and Neville (1996). The P600 was larger for those with exposure before age 10, and the effect was attenuated in those whose exposure came later. All our L2 participants had some exposure to English before age 10, but they were all recruited in Taiwan, where English was taught through formal education, and chances to use English daily were more restricted, in contrast with Weber-Fox and Neville's participants recruited in the US, who had been immersed in an English environment for at least five years. We speculate that our L2 participants may have had less certainty about their knowledge of the grammar, and thus may not have attempted to reanalyze the ungrammatical sentences as those who had more exposure to the language would.

4. Conclusion

We hypothesized that native language subcategorization knowledge is particularly difficult for L2 speakers to override in online processing. The results of the ERP experiment supported this hypothesis: the L1 showed a prominent P600 effect to subcategorization violations, but the L2 was showed a null effect, even as they showed ERP sensitivity to phrase structure violations in a control comparison. Both deficits in L2 knowledge (not having the right information encoded in the lexicon), and deficits in

L2 processing (not being able to override L1 subcategorization information online) likely contribute jointly to the insensitivity observed here. While the processing deficit may reflect interference associated with automatic access of conflicting L1 lexical information at the verb, our exploratory analysis suggested that overriding such L1 information could be possible as the L2 speaker's proficiency improves. Together, our data serve as a reminder that computing verb-argument relations, although a seemingly simple task relative to other grammatical phenomena, in fact requires accessing lexical syntax which may be vulnerable to L1 interference in L2 until higher proficiency is attained.

References

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology, 4*, 328.

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2015). Effects of prediction and contextual support on lexical processing: Prediction takes precedence. *Cognition, 136*, 135-149.

Brysbaert, M., & Duyck, W. (2010). Is it time to leave behind the Revised Hierarchical Model of bilingual language processing after fifteen years of service? *Bilingualism: Language and cognition*, *13*(3), 359-371.

Caffarra, S., Mendoza, M., & Davidson, D. (2019). Is the LAN effect in morphosyntactic processing an ERP artifact? *Brain and Language, 191,* 9-16.

Caffarra, S., Molinaro, N., Davidson, D., & Carreiras, M. (2015). Second language syntactic processing revealed through event-related potentials: An empirical review. *Neuroscience & Biobehavioral Reviews, 51*, 31-47.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied psycholinguistics*, *27*(1), 3-42.

Cohn, N., & Kutas, M. (2015). Getting a cue before getting a clue: Event-related potentials to inference in visual narrative comprehension. *Neuropsychologia, 77,* 267-278.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21.

Dufour, R., & Kroll, J. F. (1995). Matching words to concepts in two languages: A test of the concept mediation model of bilingual representation. *Memory and Cognition*, *23*(2), 166-180.

Dussias, P. E., & Scaltz, T. R. C. (2008). Spanish–English L2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta psychologica*, *128*(3), 501-513.

Dussias, P. E., Marful, A., Gerfen, C., & Bajo Molina, M. T. (2010). Usage frequencies of complement-taking verbs in Spanish and English: Data from Spanish monolinguals and Spanish—English bilinguals. *Behavior Research Methods*, *42*(4), 1004-1011.

Friederici, A. D., & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and language*, *43*(3), 476-507.

Friederici, A. D., Hahne, A., & Mecklinger, A. (1996). Temporal structure of syntactic parsing: early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1219.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language*, *37*(1), 58-93.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and cognitive processes, 25*(2), 149-188.

Guo, J., Guo, T., Yan, Y., Jiang, N., & Peng, D. (2009). ERP evidence for different strategies employed by native speakers and L2 learners in sentence processing. *Journal of Neurolinguistics*, *22*(2), 123-134.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia*, *38*(11), 1531-1549.

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (SPS) as an ERP measure of syntactic processing. *Language and cognitive processes*, *8*(4), 439-483.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language learning*, *57*(1), 1-33.

Kaan, E., & Swaab, T. Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of cognitive neuroscience*, *15*(1), 98-110.

Karatas, N. B. (2019). *The comparison of L1 and L2 case processing: ERP evidence from Turkish. [Unpublished doctoral dissertation]. University of Maryland.*

Kielar, A., Meltzer-Asscher, A., & Thompson, C. K. (2012). Electrophysiological responses to argument structure violations in healthy adults and individuals with agrammatic aphasia. *Neuropsychologia*, *50*(14), 3320-3337.

Koelsch, S., & Jentschke, S. (2010). Differences in electric brain responses to melodies and chords. *Journal of cognitive neuroscience, 22*(10), 2251-2262.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of memory and language, 33*(2), 149-174.

Kroll, J. F., Van Hell, J. G., Tokowicz, N., & Green, D. W. (2010). The Revised Hierarchical Model: A critical review and assessment. *Bilingualism: Language and Cognition*, *13*(3), 373-381.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161.

Liao, C. H., & Lau, E. (2020). Enough time to get results? An ERP investigation of prediction with complex events. *Language, Cognition and Neuroscience, 35(9)*, 1162-1182.

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience, 8*, 213.

Mickan, A., & Lemhöfer, K. (2020). Tracking syntactic conflict between languages over the course of L2 acquisition: A cross-sectional event-related potential study. *Journal of Cognitive Neuroscience*, *32*(5), 822-846.

Neville, H., Nicol, J. L., Barss, A., Forster, K. I., & Garrett, M. F. (1991). Syntactically based sentence processing classes: Evidence from event-related brain potentials. *Journal of cognitive Neuroscience*, *3*(2), 151-165.

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning*. Cambridge University Press.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, *31*(6), 785-806.

Osterhout, L., Holcomb, P. J., & Swinney, D. A. (1994). Brain potentials elicited by garden-path sentences: evidence of the application of verb information during parsing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 786.

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning, 56,* 199-230.

Silva, S., Folia, V., Hagoort, P., & Petersson, K. M. (2017). The P600 in implicit artificial grammar learning. *Cognitive science, 41*(1), 137-157.

Sitnikova, T., Holcomb, P. J., Kiyonaga, K. A., & Kuperberg, G. R. (2008). Two neurocognitive mechanisms of semantic integration during the comprehension of visual real-world events. *Journal of cognitive neuroscience, 20*(11), 2037-2057.

Steinhauer, K. (2014). Event-related potentials (ERPs) in second language research: A brief introduction to the technique, a selected review, and an invitation to reconsider critical periods in L2. *Applied Linguistics, 35*(4), 393-417.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and language*, *120*(2), 135-162.

Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research*, *25*(1), 13-41.

Tanner, D., Grey, S., & van Hell, J. G. (2017). Dissociating retrieval interference and reanalysis in the P600 during sentence comprehension. *Psychophysiology, 54*(2), 248-259.

Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, *16*(2), 367-382.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology, 83*(3), 382-392.

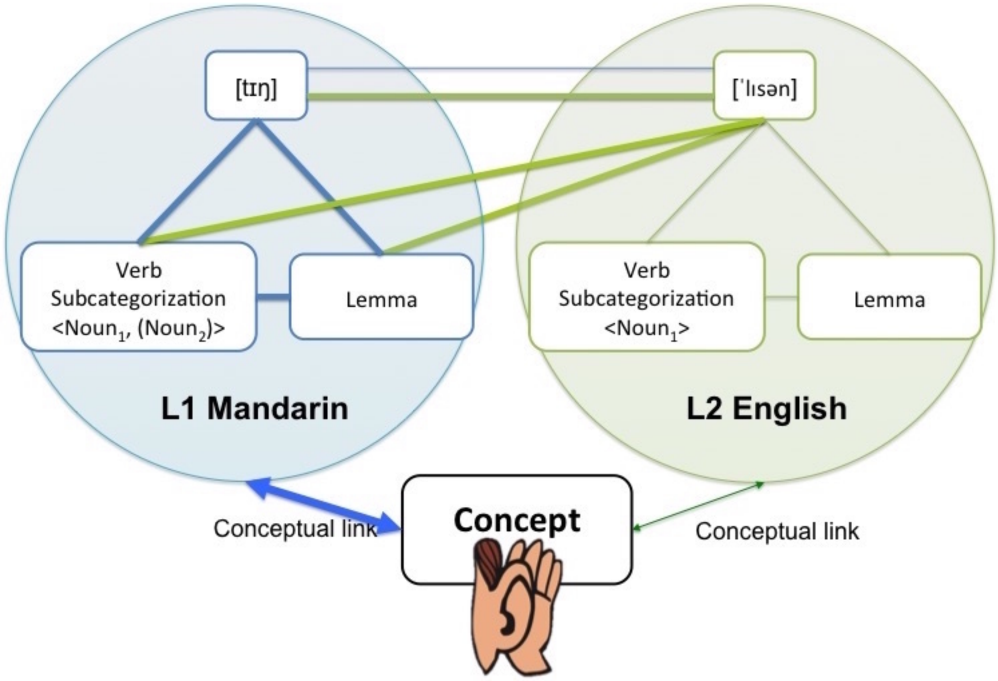
Ullman, M. T. (2012). The declarative/procedural model. In: P. Robinson (Ed.), *Routledge encyclopedia of second language acquisition* (pp. 160–164). New York & London: Routledge.

Ueno, M., & Garnsey, S. M. (2008). An ERP study of the processing of subject and object relative clauses in Japanese. *Language and Cognitive Processes, 23*(5), 646-688.

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of cognitive neuroscience*, *8*(3), 231-256.

White, E. J, Genesee F, & Steinhauer, K. (2012) Brain Responses before and after Intensive Second Language Learning: Proficiency Based Changes and First Language Background Effects in Adult Learners. *PLoS ONE 7*(12): e52318. doi:10.1371/journal.pone.0052318

Ye, Z., Zhan, W., & Zhou, X. (2007). The semantic processing of syntactic structure in sentence comprehension: An ERP study. *Brain research, 1142*, 135-145.



A schematic diagram of online processing upon reading a verb. The figure is adapted from the Revised Hierarchical Model (Kroll & Stewart, 1994)

63x43mm (300 x 300 DPI)

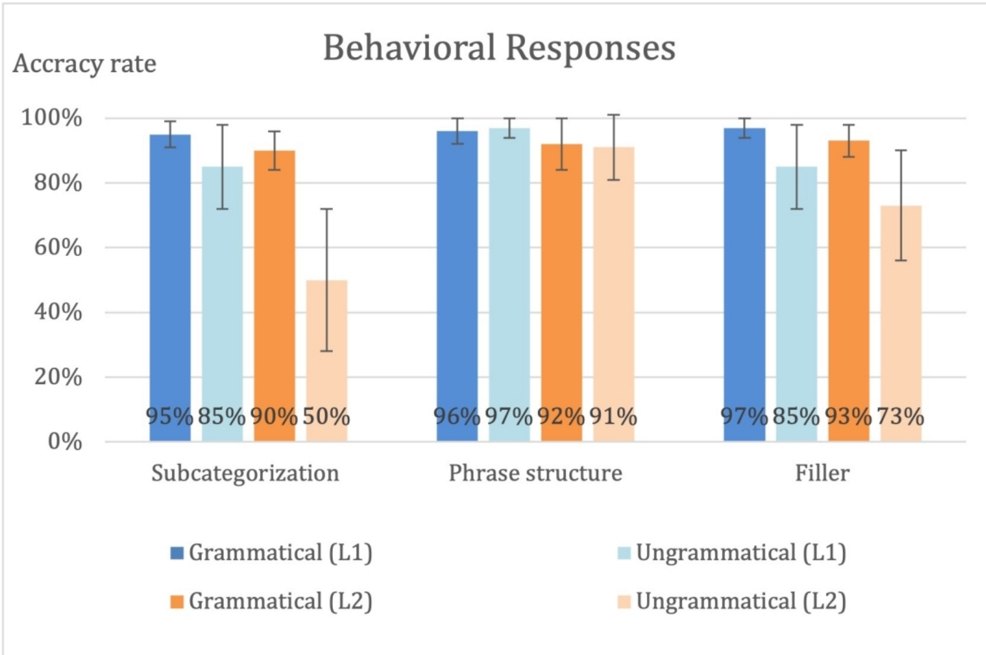


Figure 2: Accuracy rate and standard deviation of each condition in L1 and L2 groups

129x86mm (330 x 330 DPI)

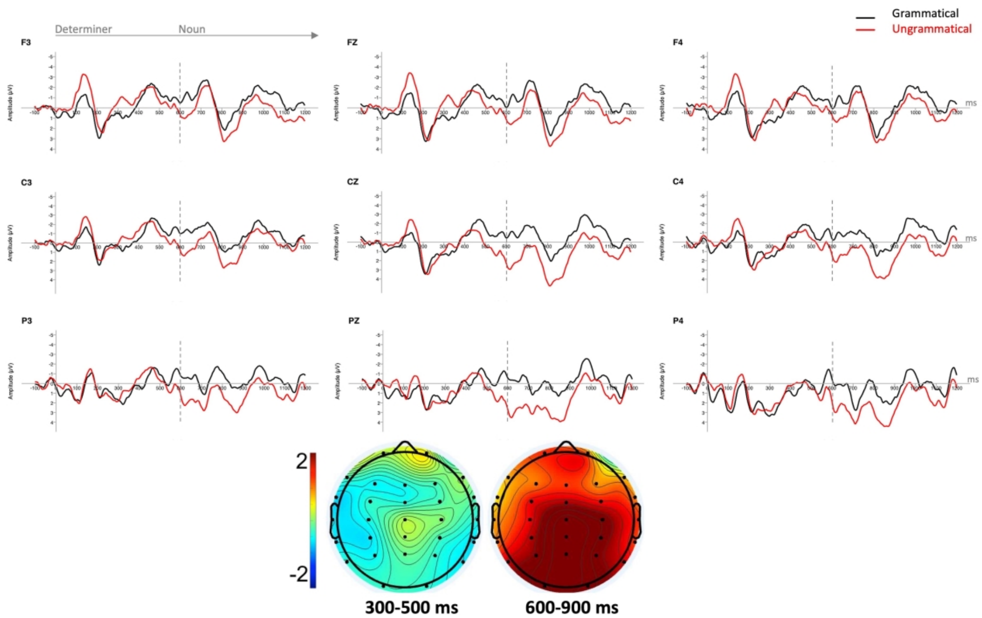


Figure 3A: Top: Grand average ERPs of Subcategorization conditions from the determiner to the nouns at representative electrodes in L1 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the determiner in L1 speakers.

550x347mm (330 x 330 DPI)

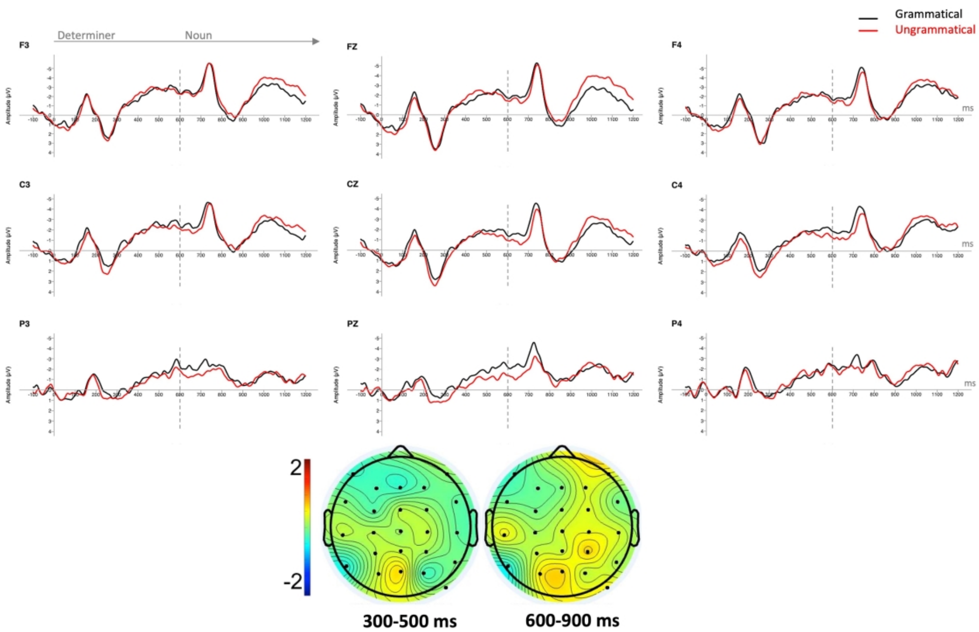


Figure 3B: Top: Grand average ERPs of Subcategorization conditions from the determiner to the nouns at representative electrodes in L2 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the determiner in L2 speakers.

550x354mm (330 x 330 DPI)

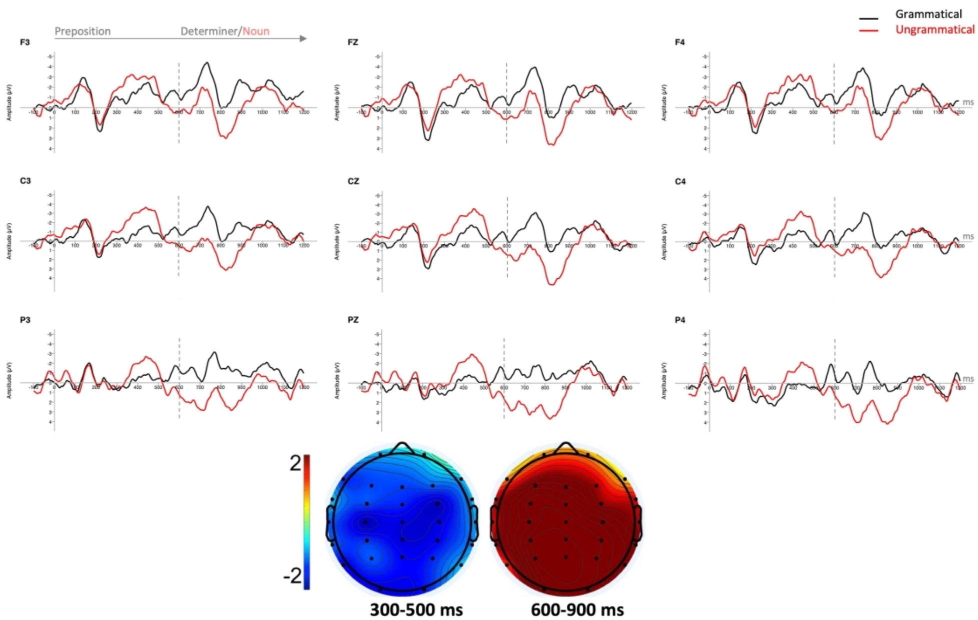


Figure 4A: Top: Grand average ERPs of Phrase structure conditions from the preposition to the following word at representative electrodes in L1 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the preposition in L1 speakers.

550x348mm (330 x 330 DPI)

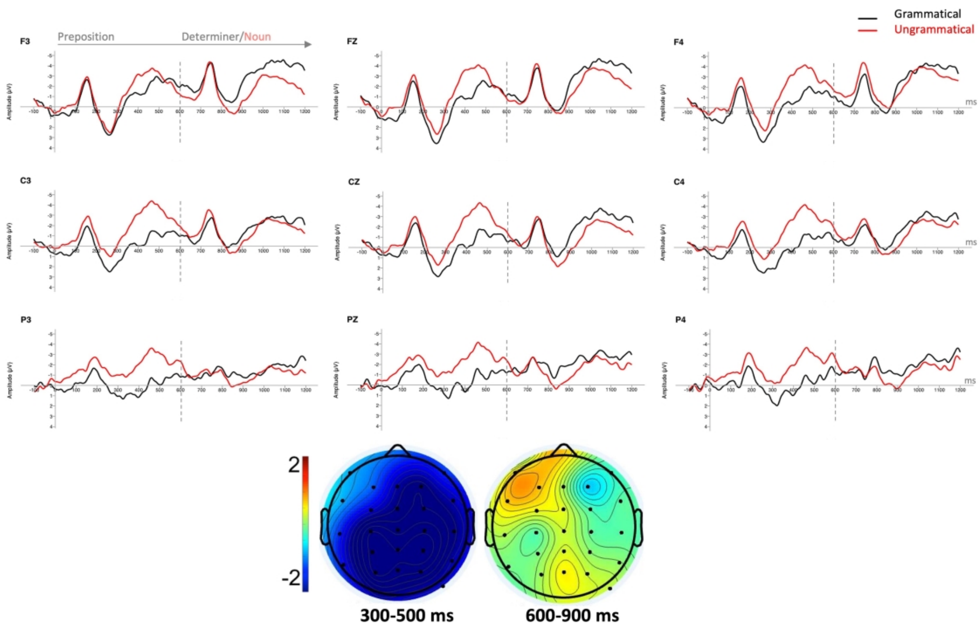


Figure 4B: Top: Grand average ERPs of Phrase structure conditions from the preposition to the following word at representative electrodes in L2 speakers. Bottom: Topographic distributions in the 300-500 and 600-900 ms intervals at the preposition in L2 speakers.

550x352mm (330 x 330 DPI)

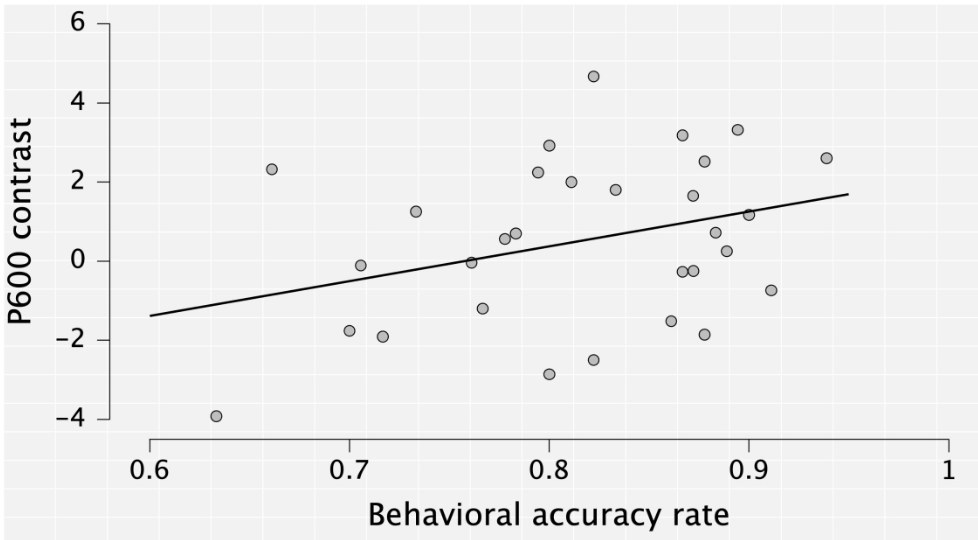


Figure 5: Correlation between participant's behavioral accuracy rate and the P600 contrast over the parietal area in Subcategorization.

232x128mm (330 x 330 DPI)

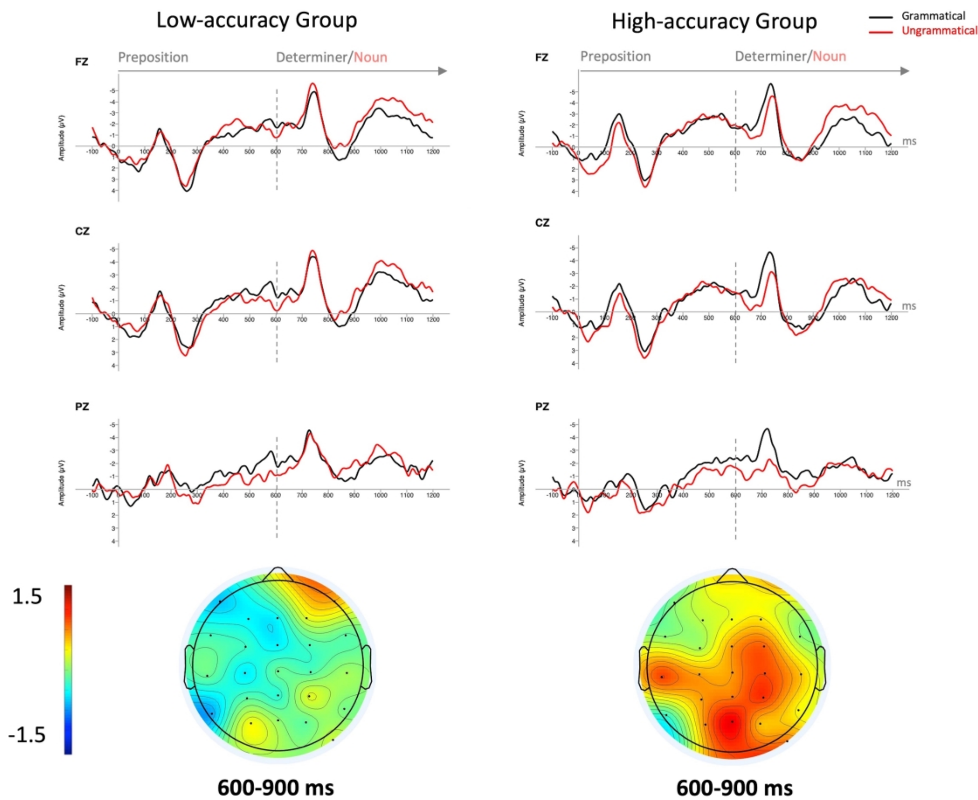


Figure 6: Grand average ERPs of Subcategorization in the L2 group, divided by a median split of accuracy. Left: Low-accuracy group (N =14); Right: High-accuracy group (N=16).

499x410mm (330 x 330 DPI)

Supplementary File

Table 1: Mixed effects model for Subcategorization violations (300-500 ms interval)

Parameters	*Fixed effects*			*Random effects*	
				By subject	By Item
	Estimate	SE	t	SD	SD
Intercept	-0.68	0.31	-2.20	1.79	1.08
Subcategorization	0.09	0.28	0.34	1.33	
Region	-0.84	0.20	-4.12*	0.15	
Group	0.68	0.61	1.12		1.98
Subcategorization x Region	0.16	0.41	0.38	0.24	
Subcategorization x Group	0.44	0.56	0.79		
Region x Group	-0.19	0.41	-0.47		
Subcategorization x Region x Group	-0.16	0.82	-0.19		

Note: All factors were coded using contrast coding, as follows: Subcategorization (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front), Group (-.5 = L2 group, .5 = L1 group). Model formula : N4 ~ Subcategorization*Region*Group + (Subcategorization*Region|Subject) + (Group|Item)

*|t| ≥ 2.0 indicates a significant effect (Gelman & Hill, 2007)

Table 2: Mixed effects model for Subcategorization violations (600-900 ms interval)

Parameters	*Fixed effects*			*Random effects*	
				By subject	By Item
	Estimate	SE	t	SD	SD
Intercept	-0.42	0.36	-1.16	2.16	1.13
Subcategorization	-1.05	0.34	-3.08*	1.87	
Region	-0.37	0.22	-1.69	0.45	
Group	2.45	0.71	3.45*		2.16
Subcategorization x Region	0.57	0.42	1.35	0.38	
Subcategorization x Group	-1.46	0.68	-2.13*		
Region x Group	-1.01	0.44	-2.30*		
Subcategorization x Region x Group	0.58	0.85	0.68		

Note: All factors were coded using contrast coding, as follows: Subcategorization (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front), Group (-.5 = L2 group, .5 = L1 group). Model formula : P6 ~ Subcategorization*Region*Group + (Subcategorization*Region|Subject) + (Group|Item)

*|t| ≥ 2.0 indicates a significant effect (Gelman & Hill, 2007)

Table 3: Mixed effects model for Phrase structure violations (300-500 ms interval)

Parameters	*Fixed effects*			*Random effects*	
				By subject	By Item
	Estimate	SE	t	SD	SD
Intercept	-1.28	0.25	-5.09*	1.29	1.06
Phrase	1.81	0.35	5.28*	1.94	
Region	-0.79	0.21	-3.70*	0.41	
Group	0.11	0.49	0.22		1.89
Phrase x Region	-0.29	0.42	-0.69	0.60	
Phrase x Group	-0.79	0.69	-1.14		
Region x Group	-0.56	0.43	-1.32		
Phrase x Region x Group	0.62	0.84	0.74		

Note: All factors were coded using contrast coding, as follows: Phrase (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front), Group (-.5 = L2 group, .5 = L1 group). Model formula : N4 ~ Phrase*Region*Group + (Phrase*Region|Subject) + (Group|Item)

*|t| ≥ 2.0 indicates a significant effect (Gelman & Hill, 2007)

Table 4: Mixed effects model for Phrase structure violations (600-900 ms interval)

Parameters	*Fixed effects*			*Random effects*	
				By subject	By Item
	Estimate	SE	t	SD	SD
Intercept	-0.40	0.34	-1.19	2.03	1.06
Phrase	-1.57	0.35	-4.45*	1.99	
Region	-0.42	0.21	-1.98	0.30	
Group	1.36	0.68	2.00*		2.26
Phrase x Region	0.51	0.42	1.21	0.20	
Phrase x Group	-2.62	0.71	-3.71*		
Region x Group	-1.15	0.43	-2.71*		
Phrase x Region x Group	0.73	0.84	0.87		

Note: All factors were coded using contrast coding, as follows: Phrase (-.5 = Ungrammatical, .5 = Grammatical), Region (-.5 = Back, .5 = Front), Group (-.5 = L2 group, .5 = L1 group). Model formula : P6 ~ Phrase*Region*Group + (Phrase*Region|Subject) + (Group|Item)

*|t| ≥ 2.0 indicates a significant effect (Gelman & Hill, 2007)