Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

L2 processing of filled gaps
Non-native brain activity not modulated by proficiency and working memory

Zhiyin Renee Dong, Chao Han,
Arild Hestvik, and Gabriella Hermon

University of Delaware

Abstract

This paper investigates how late L2 learners resolve filler-gap dependencies (FGD) in real-time and how proficiency and working memory (WM) modulate their brain responses in an event-related potential (ERP) experiment. A group of intermediate to highly proficient Mandarin Chinese learners of English listened to sentences such as "The zebra that the hippo kissed *the camel on the nose ran far away," in which the extra noun phrase "the camel" created a 'filled-gap' effect. The results show that although L2 behavioral responses are comparable to native speakers and are positively correlated with proficiency and WM span, the brain responses to the filled gap are qualitatively different. Importantly, L2 processing patterns did not become more nativelike with higher proficiency levels or greater WM capacity. Specifically, while the native speakers exhibited a P600 typically observed for syntactic violations and repair, the L2 group produced a prefrontal-central positivity. Similar ERPs have previously been reported to reflect domain-general attentional and non-structural-based processes, suggesting that the L2 group has a reduced sensitivity to structural requirements for gap positing in the online resolution of FGDs. Our findings are discussed in light of various proposals accounting for L1-L2 processing differences, including the Shallow Structure Hypothesis.

Keywords
L2 sentence processing, filler-gap dependencies, ERP, individual factors, shallow structure hypothesis

1. Introduction

An important research topic in Second Language Acquisition (SLA) is whether adult second language (L2) learners adopt a fundamentally different parsing mechanism from native speakers in real-time sentence processing. Some

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

researchers maintain that first language (L1) and L2 processing are qualitatively the same (e.g., Sabourin & Stowe, 2008) and that non-native L2 processing can be explained by individual difference factors. These factors include proficiency (e.g., Ojima, Nakata & Kakigi, 2005; Dallas, DeDe & Nicol, 2013), L1 influence (e.g., Weber & Cutler, 2004), the higher demand of L2 processing for cognitive resources such as working memory (WM) (e.g., McDonald, 2006), lexical access effectiveness (e.g., Hopp, 2017), susceptibility to memory retrieval interference (Cunnings, 2017), and reduced ability to predict upcoming information during L2 parsing (e.g., Kaan, 2014). By contrast, others argue that these factors alone cannot account for the observed L1-L2 processing differences; online L2 processing may remain non-nativelike even with ultimate attainment (e.g., Pakulak & Neville, 2011; Hawkins & Chan, 1997). In particular, the shallow structure hypothesis (SSH) (e.g., Clahsen & Felser, 2006, 2018) states that adult L2 sentence processing, unlike that of L1, "prioritizes semantic, pragmatic or other types of non-grammatical information" and that "even highly proficient L2 speakers tend to have problems building or manipulating abstract syntactic representation in real-time" (Clahsen & Felser, 2018, p. 3). While numerous studies have examined how and why L2 sentence processing both resembles and differs from L1, the findings hitherto remain inconclusive (e.g., Omaki & Schulz, 2011; Ojima et al., 2005; Dowens, Guo, Guo, Barber & Carreiras, 2011). This is especially true of L2 studies on structurally hierarchical constructions, including filler-gap dependencies (FGDs) (e.g., Marinis, Roberts, Felser & Clahsen, 2005). Furthermore, how individual factors such as proficiency and WM impact L2 processing require further clarification (e.g., Van Hell & Tokowicz, 2010; Hopp, 2017). To address these gaps in our current knowledge, this paper presents an event-related potential (ERP) experiment examining how late Mandarin Chinese (hereafter Chinese) learners of English process sentences in which an extra noun phrase (NP) creates a 'filled gap' effect (Stowe, 1986) (i.e., "The zebra that the hippo kissed *the camel on the nose ran far away"). Specifically, we seek to investigate (1) whether online L2 processing of filled gaps is nativelike, in particular regarding the use of abstract syntactic information, and (2) to what extent individual difference factors, such as proficiency and WM capacity, can explain observed L1-L2 processing differences.

1.1 L2 processing of FGD: Is structural information underused in gap positing?

A 'filler-gap dependency' refers to the relationship between a dislocated sentence constituent (known as the 'filler') and its originating position (the 'gap'), where the verb typically assigns a thematic role to the filler (e.g., Clifton & Frazier, 1989), as in sentences such as "The lady (filler) that the doctor treated__ (gap) yester-

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

day for a minor cut was from England." FGD is ideal for testing L2 online use of detailed syntactic information due to its hierarchical structure. Long-distance FGD such as the stimuli used in the current study is also complex and memory-taxing, rendering it suitable to examine the impact of factors such as working memory (e.g., Johnson, Fiorentino & Gabriele, 2016). Current L2 FGD processing findings indicate that L2 learners have access to complex structural representations, are sensitive to certain structural constraints (e.g., Juffs, 2006; Omaki & Schultz, 2011), and can use many L1 parsing routines, including the active filler strategy (Clifton & Frazier, 1989) in nativelike ways (e.g., Williams, Möbius & Kim, 2001). However, it has also been found that L2 speakers resolve FGDs by associating the filler directly with the verb that subcategorizes for it, rather than by computing detailed, hierarchical structures (e.g., Felser & Roberts, 2007; Marinis et al., 2005). For example, in a self-paced reading experiment by Marinis et al. (2005), intermediate-high to advanced learners of English from various L1 backgrounds read sentences such as the following:

(1) a. The manager who$_i$ the secretary claimed ***t$_i$ [intermediate gap site]*** that the new salesman had pleased t$_i$ ***[base position]*** will raise company salaries.
b. The manager who$_i$ the secretary's claim ***[no intermediate gap]*** about the new salesman had pleased t$_i$ ***[base position]*** will raise company salaries.

Assuming the movement account of generative grammar (Chomsky, 1986), the filler "the manager" originated from the base position after "had pleased," leaving behind a trace. Because a one-step movement from the base position is prohibited by principles such as subjacency (Chomsky, 1986), an intermediate step is required; hence the intermediate gap site in (1a). Marinis et al. (2005) reported increased reading time, indicative of filler integration for both groups at the base position in (1a) and (b) in comparison to the controls sentences without movement. However, only the native speakers showed increased reading time at the proposed intermediate site, which was taken to reflect the cost of filler/trace activation. As the L2 learners were from different L1 backgrounds and were highly nativelike in the offline comprehension test, such an L1-L2 reading time difference cannot be explained by either low proficiency or L1 influences. Marinis et al. (2005) concluded that L2 FGD is formed by linking the filler to a suitable subcategorizing verb based on semantic/pragmatic fitness, rather than building detailed structure with abstract elements (i.e., intermediate trace). Marinis et al. (2005)'s results were later called into question due to several methodological limitations, including, for instance, failure to examine the 'spillover' regions after the intermediate gap site (e.g., the complementizer "that") (e.g., Miller, 2015). Several subsequent studies addressed this concern and found elevated RT at the complementizer "that," which was taken to suggest that an intermediate gap was indeed

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

posited in online L2 computation of FGD (Miller, 2015). However, while increased RT reflects the cost of gap positing and filler integration, it reveals little about whether such processes are guided by structural constraints (e.g., subjacency) or verb-filler association. To shed new light on this issue, we replicated an L1 study designed specifically to examine how FGD is formed (Hestvik, Maxfield, Schwartz & Shafer, 2007; Hestvik, Bradley & Bradley, 2012), using a different paradigm (filled gap) and a different methodology (ERP).

1.2 ERP studies on the filled-gap effect and the present study

The ERP technique is widely used in language processing studies due to its excellent temporal resolution and high sensitivity to automatic and sometimes subconscious language processes than behavioral measures (e.g., Dowens et al., 2011). One of the most commonly found ERP indices is the N400, a centro-parietal negative-going voltage shift typically extends from 250–500 milliseconds (ms) and peaks at 400 ms after the violating item. The N400 is reliably linked to accessing semantic features from long-term memory, semantic incongruities, and violations associated with a verb's arguments (e.g., Kutas & Federmeier, 2011; Frisch, Hahne & Friederici, 2004). Another common component is the P600, a positive-going voltage wave obtained 600–900 ms post-onset of the stimulus in the parietal region of the scalp. The P600 has been consistently observed for various syntactic anomalies, including phrase structure violations (e.g., Hahne & Friederici, 1999), morphosyntactic violations (e.g., Hagoort, Brown & Groothusen, 1993), complex syntactic structures such as FGDs (e.g., Kann & Swaab, 2003), and the 'reanalysis and revise' processes triggered by syntactic ambiguities and processing difficulties (e.g., Gouvea, Phillips, Kazanina & Poeppel, 2010). The P600 has also been found in various non-syntactic contexts, including thematic role reversal (e.g., Kim & Osterhout, 2005), animacy violations (Kuperberg, Kreher, Sitnikova, Caplan & Holcomb, 2007), and certain strong semantic violations following an N400 (e.g., DeLong & Kutas, 2020). Finally, the left-anterior negativity (LAN) is also relevant here. The LAN is a negative voltage shift between 300–500 ms after the onset of the violation. It is obtained in the anterior position, usually on the left side but sometimes bi-laterally (e.g., Pakulak & Neville, 2011). The LAN is commonly found for morphosyntactic and syntactic category violations (e.g., Caffarra, Mendoza & Davidson, 2019). In the latter case, the LAN is often followed by a P600 (e.g., Bowden, Steinhauer, Sanz & Ullman, 2013). Syntactic category violations have also elicited the early left anterior negativity (ELAN), a similar ERP to the LAN, but with an earlier onset of 150–200 ms and is taken to denote first-pass phrase structure building (e.g., Friederici, 2002). The status of ELAN is somewhat controversial; some earlier ELAN effects were thought to be artifacts

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

rather than the results of experimental manipulations, as the materials preceding the manipulation were not the same in experimental and control conditions (e.g., Steinhauer & Drury, 2012). In the current study, we avoided such a problem by keeping the materials leading up to the critical region identical across conditions (see Section 2.3 for details).

Using the ERP method, Hestvik et al. (2007, 2012) tested whether structural trace was used in the online L1 resolution of FGD by comparing the test stimulus like (2a) to the control in (2b):

(2) a. *The zebra that the hippo kissed the camel on the nose ran far away.
b. The weekend that the hippo kissed the camel on the nose, it was hot.

Assuming the active filler strategy (Clifton & Frazier, 1989), the parser posits a gap immediately after the verb "kissed" and attempts to interpret it as corresponding to the filler "the zebra," only to find that the gap has been filled with an extra NP "the camel." Hestvik et al. (2007, 2012) hypothesized that if the verb and the filler were linked together based on argument structure, thematic role assignment, or other non-syntactic information, then the extra word should violate verb argument expectations and generate an N400 (e.g., Frisch et al., 2004). However, an anterior negativity, interpreted as an ELAN due to its early onset, was found instead of the N400. In Hestvik et al. (2012), this anterior negativity was also followed by a P600. As noted above, ELAN/LAN + P600 has been obtained for syntactic category violations, which is consistent with the filled gap manipulation; as the parser posits the gap and builds an upcoming syntactic position of a trace, it expects a word category other than an NP.[1] This projected structure clashes with the extra NP and registers as a syntactic category violation, hence the ELAN/LAN. The P600 found in Hestvik et al. (2012), in which the role of WM was examined, was taken to reflect the 'reanalysis' and 'repair' processes commonly encountered concerning issues in processing syntax (e.g., Kann & Swaab, 2003; Gouvea et al., 2010). Findings from Hestvik et al. (2007, 2012) thus show that L1 FGD formation is guided by detailed structural building with abstract trace rather than direct verb-filler association. Additionally, in line with evidence suggesting that WM capacity may affect aspects of filler-gap formation (e.g., Nakano, Felser & Clahsen, 2002; Nicenboim, Vasishth, Gattei, Sigman & Kliegl, 2015), Hestivk et al. (2012) found that the low-WM span participants were slower with the gap-filling

1. At that temporal juncture, the sentence is not technically ungrammatical, as it could continue as the "The zebra that the hippo kissed the camel for." However, these alternative continuations have a much lower probability than a simple direct object NP. We adhere to the position that the parser is a probabilistic device, and as such, any low-probability event will be unexpected, and ungrammaticality is the least probable event.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

process, as both of their (E)LAN and the P600 had delayed onset compared to the high-WM span participants.

Current L2 ERP studies on the processing of filled gaps in FGD are limited. Notably, an ERP study conducted by Jessen, Festman, Boxell, and Felser (2017) examined the filled indirect object effect by having highly proficient German speakers of English read sentences like the following:

(3) a. *Sarah tickled the monkey for which Peter arranged some class for it after the vacation.
b. Sarah tickled the monkey while Peter arranged some class for it after the vacation.

For the presumptive prepositional phrase (PP) "for it" in (3a), which fills the indirect object position intended for the filler "the monkey" relative to the control (2b), the L2 speakers produced a P600 in the classic central-parietal location, interpreted as indicative of grammatical violation and its repair (e.g., Friederici, Hahne & Saddy, 2002). The native speakers also generated a positivity in the 600–800 ms range, but with a frontal-central distribution. Such 'frontal P600' was previously found for well-formed sentences with discourse/syntactic complexity and temporary syntactic ambiguity (e.g., Kaan & Swaab, 2003). According to Jessen et al. (2017), this unexpected L1 ERP was possibly because the FG stimuli sentences were still globally acceptable, rendering the ungrammaticality less evident to the parser. Interestingly, the L2 positivity was more globally distributed and of higher amplitude than the L1. Although both positivities were interpreted as the P600, suggesting qualitatively similar L1-L2 processing patterns, the authors nevertheless maintain that the observed the L1 and L2 ERPs could index different processes and that the L2 speakers experienced greater difficulty than the L1 group in processing the filled gap.

To further investigate the nature and sources of L1-L2 sentence processing differences – particularly that of filled gaps in FGDs – the present study replicates Hestvik et al. (2012) in a group of L2 speakers and directly compares L1 and L2 brain responses. We hypothesize that if L1 and L2 filled-gap processing are the same in principle and the L2 speakers can effectively use abstract structural-based information in FGD processing, the same components (Early AN/AN and the P600) should be found for both groups. If, however, we found neither of these components for the L2 group or components indicative of semantic-based processing, then L2 FGD is formed by relying on non-structural information, as suggested by Marinis et al. (2005). Additionally, we examine the L2 ERPs in relation to the L2 speakers' WM capacities and proficiency levels. As mentioned in the introduction, it is argued that L2 processing of complex syntax can become native-like with high enough proficiency and large enough cognitive resource

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

(e.g., Dallas et al., 2013; McDonald, 2006). Working memory, for example, has been found to affect online FGD formation in both L1 and L2 contexts, such that L1 participants with higher WM resolve FGD faster (e.g., Hestvik et al., 2012) and produced greater gap-filling effects (e.g., Nakano et al., 2002). L2 evidence to date also shows that high-WM span participants can recover easier from certain misanalyzed FGDs than low-WM participants (Dussias & Pinar, 2010), and higher L2 WM could lead to greater, more native-like gap-filling effects (Johnson et al., 2016). We thus predict that if L1 vs. L2 processing differences result from proficiency or memory limitations, L2 processing patterns should become more native-like as WM capacity or proficiency increases. If, however, L1 vs. L2 processing differs qualitatively and is not affected by individual factors, we should find that L2 processing patterns remain non-native, even with higher proficiency and WM span.

2. Method

2.1 Participants

A total of 57 Chinese speakers of English participated in the experiment, and data from one participant was excluded due to an EEG data collection error. The average age of the remaining 56 L2 participants was 24 years ($SD = 2$, 38 females, and 18 males). Their English learning was mostly classroom-based, with an average first exposure age of ten years ($SD = 3$) and an average formal instruction duration of 12 years ($SD = 3$). None of the participants had lived in an all-English-speaking environment before age 14; the average length of residence in English-speaking countries was 36 months ($SD = 19$).

As a native control group for ERP responses, we used data from Hestvik et al. (2012), which has the same experimental design and stimuli as the current study (see Section 2.3 for details). We reprocessed the raw L1 data from scratch using the same parameters as for the L2 participants (see 2.4.2 for details). After preprocessing data and excluding six high-artifact participants, we obtained data from 45 L1 participants (29 females and 16 males), with an average age of 21.3 years ($SD = 3.4$). No L1 or L2 participants reported any history of neurological impairment or speech/language impairment.

To measure WM span, the L1 speakers took the reading span test by Daneman and Carpenter (1980) and scored an average of 2.98 of 5 ($Median = 3$, $SD = 0.85$). The L2 participants used an audio version of the Harrington and Sawyer (1992) reading span test – an L2 version of Daneman and Carpenter's (1980) test (e.g., Martin & Ellis, 2012). In this L2 WM test, sentences were manipulated for obvious

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

ungrammaticality, and participants made grammaticality judgments rather than the truth-value judgments as in the Daneman and Carpenter test (1980). L2 participants showed high accuracy for the grammaticality judgments ($M = 94\%$) and recalled 32 out of 42 words on average ($M = 31.6$, $SD = 4.2$). A potential problem with the L2 WM test is that it relied on grammaticality judgments of English sentences and, as such, could be highly correlated with proficiency scores; we return to this issue below. Both groups exhibited normal distribution of WM span scores based on the Kolmogorov-Smirnov test (L1: $d = .18$, $p = .10$; L2: $d = .15$, $p = .15$). WM scores were converted to z-scores to facilitate statistical comparisons between the L1 and L2 participants.

The L2 participants took the Versant English Test (Pearson Plc), a highly valid spoken test for English proficiency (e.g., Bernstein & Cheng, 2007). Versant aligns with established foreign language proficiency guidelines such as the Common European Framework of Reference (CEFR) and the American Council on the Teaching of Foreign Languages (ACTFL) (e.g., Bernstein & De Jong, 2001). The L2 participants scored an average of 62 out of 80 points on Versant ($SD = 8$), indicating that as a group they are Advanced-Low[2] speakers of English by the ACTFL guidelines. Three discrete proficiency groups were formed following the CEFR and ACTFL proficiency guidelines, as shown in Table 1 below:

Table 1. Versant English scores and L2 proficiency levels according to ACTFL and CEFR standards

Versant	CEFR	ACTFL	Proficiency index	Num. of participants
78–80	C2	Adv-High/Superior	HIGH	2
69–78	C1	Advanced-Mid	HIGH	11
58–68	B2	Advanced-Low	MID	27
48–57	B1	Intermediate-High	LOW	14
38–47	A2	Intermediate-Mid	LOW	2

2.2 Paper-and-pencil grammaticality judgment test

To measure L2 speakers' offline grammatical knowledge of FGD, we ran a paper-and-pencil rating study comparing the L2 participants to an additional and separate group of 37 monolingual native English speakers (22 female, 15 male,

2. Advanced-Low speakers are "able to handle various communicative tasks. They can participate in most informal and some formal conversations on topics related to school, home, and leisure activities. They can also speak about some topics related to employment, current events, and matters of public and community interest" (ACTFL, 2012).

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

M_{age} = 20, SD = 2). The participants rated 30 sentences' acceptability on a seven-point scale (1-completely unacceptable and 7-perfectly acceptable) after the EEG collection. The sentences were structurally identical to the stimuli used in the ERP tasks but with different vocabulary items. Six of them were ungrammatical with the filled gap violations like the FG condition in Table 2 below; the other six items were grammatical and structurally identical to the other three ERP test conditions (see Table 2). Additionally, 18 filler sentences of various degrees of acceptability taken from Sprouse and Almeida (2012) were incorporated.

2.3 ERP experimental materials

Stimuli sentences of the present study include four conditions, as summarized in Table 2 below.

Table 2. Example sentences from each experimental condition

Condition	Sample sentence
FILLED GAP (FG)	The zebra that the hippo kissed *the camel on the nose ran far away.
Adjunct (ADJUNCT)	The weekend that the hippo kissed the camel on the nose, it was humid.
Object (OBJECT)	The zebra said that the hippo kissed the camel on the nose and then ran far away.
Trace (TRACE)	The zebra that the hippo kissed on the nose ran far away.

Both FG and ADJUNCT conditions start with a relativized object filler, which triggers the search for a suitable gap. The difference is that in the ADJUNCT condition, the dependency extends beyond the post-verbal NP, whereas in the FG condition, the dependency is interrupted by the post-verbal extra NP, creating the filled-gap effect (e.g., Stowe, 1986). The OBJECT and TRACE conditions are included as distractors so that only half the object relatives had a filled gap, and 25% of the sentences were not relative clauses. Details regarding stimuli creation can be found in Hestvik et al. (2012). Each sentence was followed by a comprehension question. For example, for the stimulus sentence "The weekend that the hippo kissed the camel on the nose, it was humid," the participant answered a question such as "Who did the hippo kiss?" or "Did the camel kiss the hippo?".[3] Feedback was given on the accuracy of the participants' answers to encourage stimuli meaning processing.

3. A total of 32 comprehension questions of four different types were used. Please see Hestvik et al. (2012) for details.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

2.4 Procedures

The participants first took the WM test. After electrode net application, they were instructed to listen to each sentence and answer the following comprehension question by pressing a button on a response box. The set of stimulus sentences and questions was divided into four blocks of 32 sentences randomly presented to the participants. The EEG recording session lasted about one hour and fifteen minutes in total, after which the L2 participants also completed the paper-and-pencil judgment task.

2.4.1 *EEG acquisition*

The experiment was programmed using E-Prime software (Schneider, Eschman & Zuccolotto, 2002). EEG was recorded using a 128-channel EGI 300 system (Hydrocel HCGSN 100 v.1.0, Geodesics, U.S.A) with a sampling rate of 250 Hz. Eye movements and blinks were monitored with electrodes placed under each eye. Electrode Cz was used as a reference during recording, and the electrode impedances were maintained below 50 kΩ. After recording, the continuous EEG was divided into epochs of 1200 ms for each trial (including a 200 ms baseline period before the onset of the critical word), and 0 ms was time-locked to the onset of the article "the" before the critical word. Baseline correction was performed using a 200-ms baseline period from −200 ms to 0 ms. For artifact correction, bad channels were replaced by spline interpolation, and eye blinks and eye movement artifacts were subtracted using ICA with the ERP PCA toolkit (Dien, 2010). The ICA was run with EEGLAB's *runica* function (Delorme & Makeig, 2004). ICA components correlated at .9 or higher with the eyeblink template were marked as eyeblinks and were subtracted from the data. Trials containing eye activity artifacts took up 18% of the L1 group and 10% of the L2 group on average. After the artifact correction, at least 97% of trials remained in each condition for both participant groups. The ERPs were computed for each participant and condition by averaging all trials regardless of participants' response to the comprehension question following each trial. The data were then average re-referenced. A 40-Hz low-pass filter was applied to the waveforms shown in the graphs for easier visualization. However, the statistical analysis was performed on the data without applying the low-pass filter.

2.4.2 *PCA-constrained derivation of time windows and electrode regions*

We used the data from Hestvik et al. (2012) as the L1 control but followed the protocol of the present study to re-establish the findings and ensure that the data from both groups were analyzed in the same way. We first computed a difference wave for the contrast between the test and the control condition (FG

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

minus ADJUNCT), which was then submitted to a sequential temporospatial PCA (Dien, 2012). The PCA procedure first decomposes the ERP response into latent uncorrelated temporal factors. The factor score matrix extracted from the temporal PCA was further decomposed into independent spatial subfactors using ICA. Because the decomposition is based on the difference wave, it provides temporospatial factors that potentially reflect the experimental effect (i.e., when and where the critical effects occurred and how much of the total variance these combined factors accounted for). Following recommendations for the removal of researcher's bias in selecting temporal and spatial 'regions of interest (ROI)' (Luck & Gaspelin, 2017), we then used the temporospatial PCA solutions to determine time windows and electrode regions for statistical analysis. Specifically, the temporal factors that each account for at least 6% of the variance and the spatial subfactors showing a spatial distribution consistent with the expected component(s) were used to select time windows and electrode regions for generating mean voltage values from the raw data. These single time/space measures per participant were then entered in mixed-effects statistical models for statistical analysis. We hypothesized that if the L2 group differed from the L1 group, we should observe an interaction between the experimental effect and Group. The L1 and L2 data were also submitted to separate stand-alone analyses to determine group-specific ERPs.[4] Mixed-effects models were built using the R software. For each model, we started with the maximal random effects structure and gradually reduced the random effects until the model converged. Multiple comparisons were corrected for using the Tukey method.

3. Results

3.1 Paper-and-pencil grammaticality judgment test results

For the paper-and-pencil judgment test, the participants rated the grammaticality of the stimuli sentences on a 1–7 (7 being the most acceptable) scale. The L1 mean ratings for the grammatical and ungrammatical sentences were 6.1 ($SD = 0.48$) and 2.0 ($SD = 0.7$), respectively. The L2 mean ratings were 5.94 ($SD = 0.87$) and 2.6 ($SD = 1.25$), respectively. The advanced L2 speakers' ratings were highly native-

4. The L2 study had half as many trials (128) as the L1 study (256). Because L2 processing is generally slower and more resource-taxing (e.g., McDonald, 2006), we decided that listening to all 256 complex FGD sentences in addition to the WM and proficiency test would have caused fatigue and negatively affected the L2 participants' performance. To address the uneven trial number issue, we also ran an additional analysis that randomly sampled half of the trials from the L1 group and replicated the results reported below.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

like, exhibiting a clear trend of accuracy improvement as a function of proficiency increase, as shown in Figure 1 below.

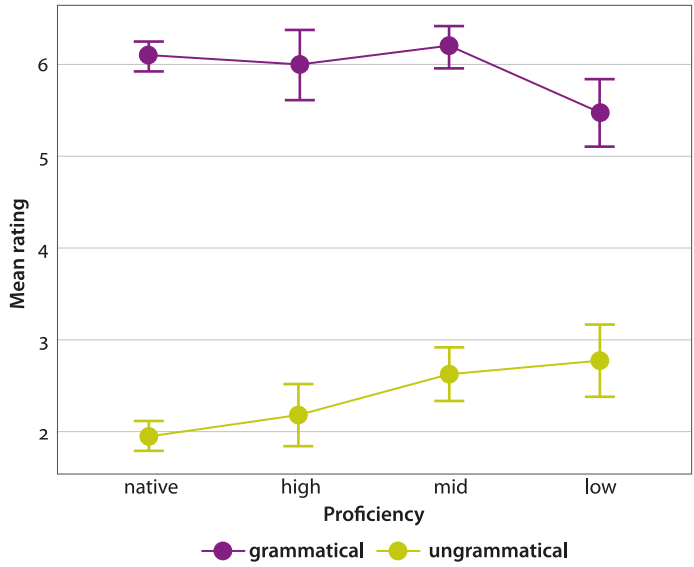


Figure 1. Paper-and-Pencil grammaticality task mean rating and proficiency (L2 and L1 groups)
Note. Vertical bars denote 95% confidence intervals (CI).

To examine the effect of proficiency on ratings, we performed a linear mixed-effects analysis using R package *lme4* (Bates, Maechler & Bolker, 2012).[5] The fixed effects were Grammaticality (Grammatical vs. Ungrammatical sentences), Proficiency (NATIVE vs. HIGH vs. MID vs. LOW), and the interaction between the two. The model included Participant and Sentence as random intercepts and revealed a significant interaction between Grammaticality and Proficiency (χ^2 (3) = 40.12, $p < .001$). We also found a significant effect of Proficiency for both the grammatical sentences (χ^2 (3) = 11.2, $p = .01$) and the ungrammatical sentences (χ^2 (3) = 9.99, $p = .002$). Post-hoc tests revealed that the L1 group differ significantly from the low-proficiency group (t (161) = 2.72, $p = .036$) for the grammatical sentence, and differ from both the low-(t (161) = 3.64, $p = .002$) and the mid-

5. We used a linear mixed-effects model instead of a cumulative profit mixed model for the Likert-scale data because previous studies demonstrated that the former model has lower Type-I error rates (Kizach, 2014) and stands valid for the Likert-scale data (e.g., Cunnings, 2012). We also ran an ordinal model using the R package *ordinal* (Christensen, 2019) and obtained the same results as the linear model.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

proficiency groups (t (161) = 3.52, p = .003) for ungrammatical sentences. However, no difference was found between the L1 group and the high-proficiency L2 group, suggesting that L1-L2 offline behavioral differences minimize as L2 proficiency increases and reaches advanced levels.

3.2 Comprehension questions results

To measure online behavioral performance, we calculated the participants' average accuracy score for comprehension questions presented after each stimuli sentence during the ERP task. Only questions for grammatical sentences were analyzed, as ungrammatical stimuli have uncertain interpretations. The L1 and L2 overall accuracy was 86% (SD = 6%, based on 192 questions, see footnote 5 for details) and 80% (SD = 8%, based on 96 questions), respectively. As Figure 2 A shows, there is a linear relationship between accuracy and proficiency.

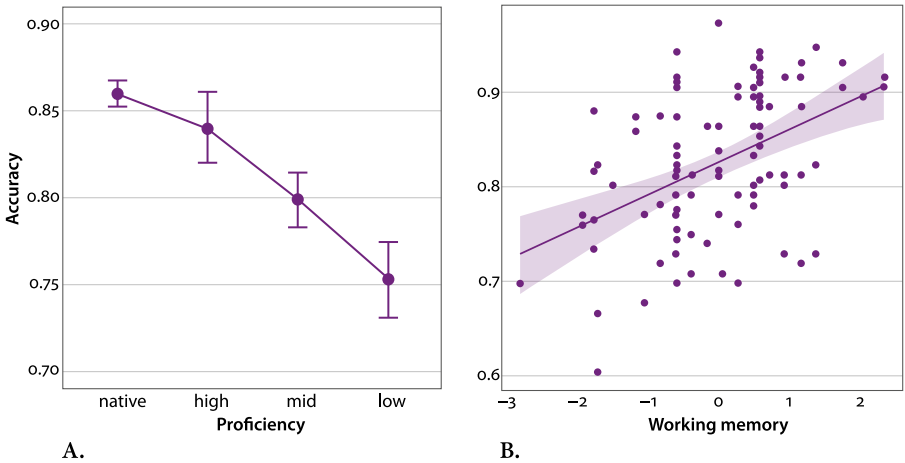


Figure 2. Effect of proficiency level (A) and WM Span(B) on online comprehension question accuracy for L1 and L2 participants
Note. Vertical bars (A) and Shades (B) denote 95% CI.

We performed a mixed-effects logistic regression to analyze the effect of proficiency and WM span on accuracy for both groups. The fixed effects were Proficiency (NATIVE (L1) vs. HIGH vs. MID vs. LOW), WM span, and the interaction between the two. The model included Participant and Sentence as random intercepts. Considering the possible correlation between proficiency and WM span, we evaluated the model's reliability by measuring the variance inflation factor, using the *vif* function from the *car* package. We found that all variance inflation factors were less than 2, suggesting multicollinearity is not a concern. The

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

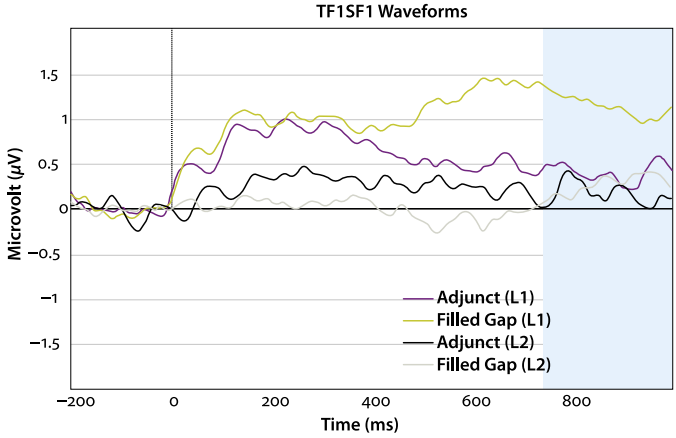
model revealed a main effect of Proficiency (χ^2 (3) = 37.31, $p < .001$), but a follow-up post-hoc test revealed that native speakers were significantly more accurate than the L2 mid-proficiency ($z = 4.29$, $p < .001$) and low-proficiency ($z = 5.37$, $p < .001$) groups, but not the high-proficiency group. A main effect of WM span (χ^2 (1) = 13.80, $p < .001$) was also observed, suggesting that the accuracy rate increased with WM span, as shown in Figure 2B above. The results confirm that L2 proficiency and WM span both predict online behavioral performance and offline grammatical knowledge accuracy. Additionally, we see that advanced L2 speakers' online behavioral performance was comparable to native speakers in both contexts.

3.3 ERP results

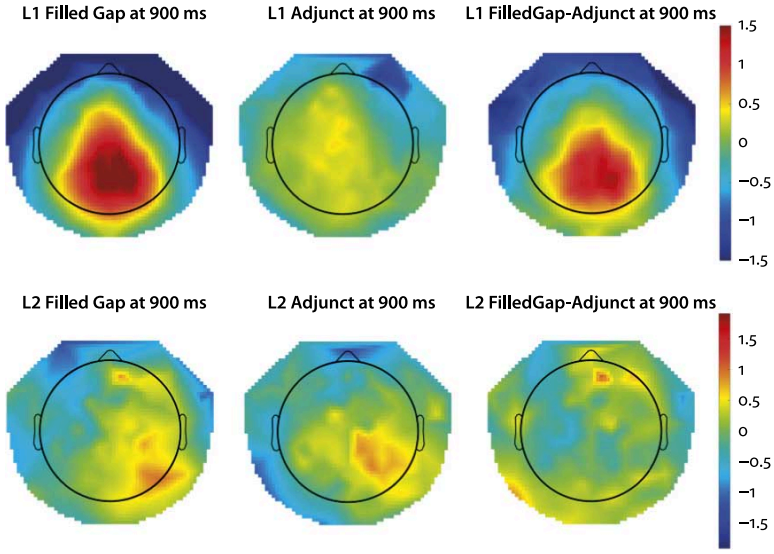
3.3.1 *Comparison of L1 and L2 speakers' brain responses*

Following the PCA analysis protocol specified above, we first analyzed ERP data from both groups to see whether there was an interaction between the groups and the ERP responses. PCA solution selected for analysis a time window of 740–996 ms and an ROI (region of interest) including 26 electrodes located in the central-parietal region (for details on how the time window and ROI are determined, please see Section 1 of the Appendix). The time window of 740–996 ms is constrained by a temporal factor peaking at 900 ms, later than a typical P600. This is because both L1 and L2 groups' data are included in the PCA, and the extracted temporal factors are affected by the L2 time series. As shown by the ERP waveforms averaged over the ROI in Figure 3A, the difference between the FG and ADJUNCT conditions of the L1 group starts around 400 ms and becomes prominent at around 700 ms, consistent with the time series of a typical P600. Figure 3B shows the topographical maps at the peaking time point (900 ms) of the selected temporal factor.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don



A.



B.

Figure 3. (A): ERP waveforms averaged over electrodes delimited by the PCA solution. The shaded blue area indicates the time window selected for analysis (740–996 ms); (B): L1 ERP topography at 900 ms for Filled Gap (upper-left), ADJUNCT (upper-middle), and the difference between the two conditions (upper-right); L2 ERP topography at 900 ms for Filled Gap (lower-left), ADJUNCT (lower-middle), and the difference between the two conditions (lower-right)

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

Using the mean voltage averaged over the selected time window and ROI, we performed a linear mixed-effects regression. The fixed effects included Group (L1 vs. L2), Condition (FG vs. ADJUNCT), WM span, and all possible interactions. The model converged when it included Participant as a random intercept. The model revealed a two-way interaction between Group and Condition (χ^2 (1) = 6.81, p = .009). To decompose the interaction, we looked into the effect of Condition within each language group using the *emmeans* function from the *emmeans* package. A significant effect of Condition was found for the L1 group (t (97) = 4.05, p = < .001) but not for the L2 group (t (97) = .61). Figure 4 suggests that the FG condition elicited a more positive response than the ADJUNCT condition for the L1 group and that the amplitude difference between ADJUNCT and FG conditions increases with WM span. However, the interaction between Condition and WM span did not reach statistical significance, possibly because the analysis time window delimited by the TF1 reflects only part of the P600 response of the L1 group. As Figure 3A shows, the most salient difference between the ADJUNCT and the FG condition in the L1 group occurred at around 600–800 ms and was only partially covered in the current analysis time window. Below we report a separate PCA for the L1 data, which better captures the P600 effect. To confirm that the absence of P600 for the L2 group applies to all the proficiency levels, we fitted a new model for the L2 group, including Proficiency (HIGH vs. MID vs. LOW) and Condition as the fixed effects and Participant as the random intercept. No interaction between Condition and Proficiency or any main effect was found (ps > .1), suggesting the absence of a P600 is invariant to the language proficiency (Figure 5).

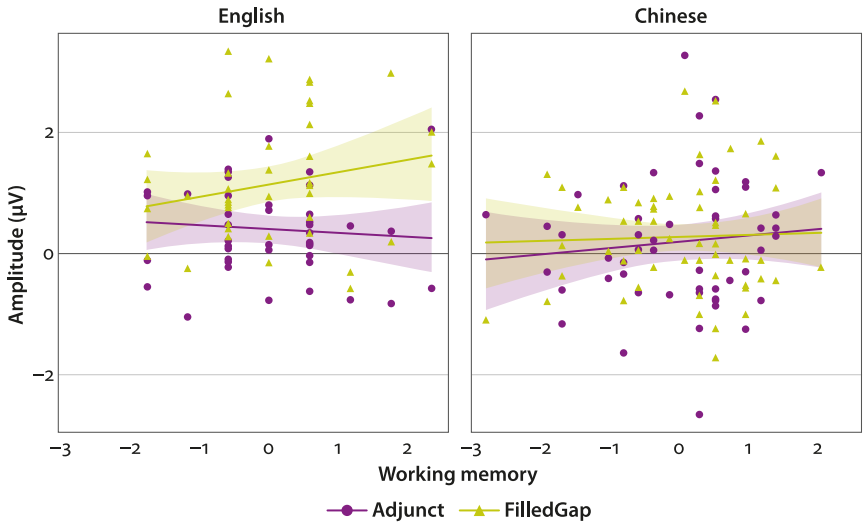


Figure 4. L1 and L2 ERP amplitude as a function of condition and WM
Note. Shade denotes 95% CI.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

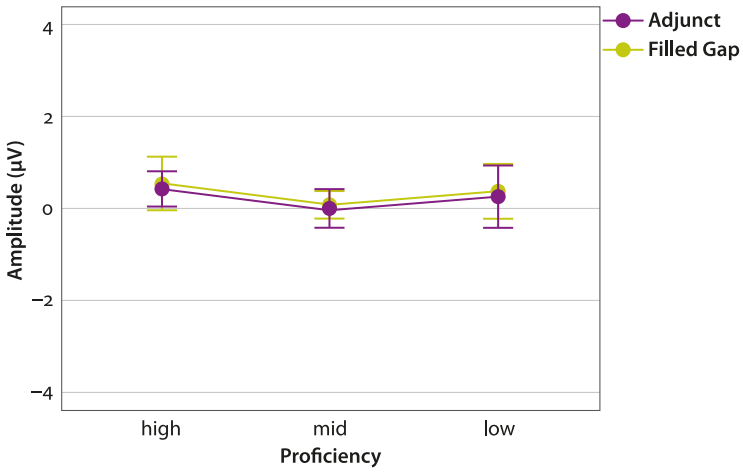
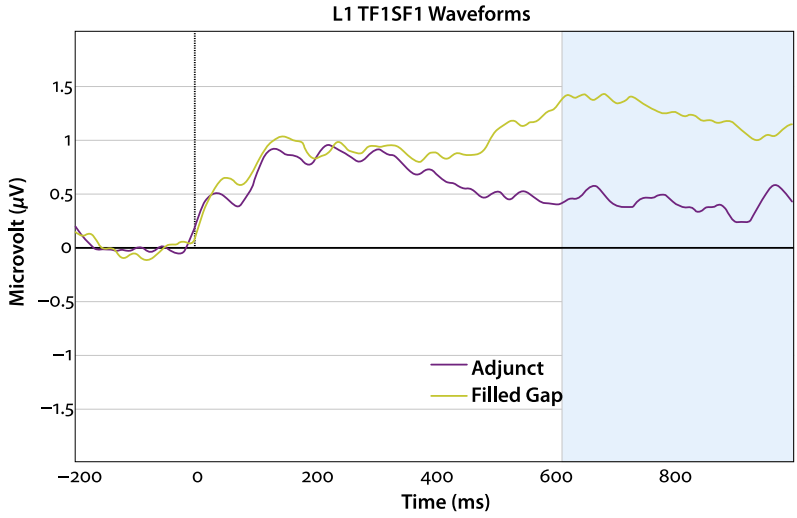


Figure 5. L2 participants ERP amplitude as a function of proficiency and condition
Note. Vertical bars denote 95% CI.

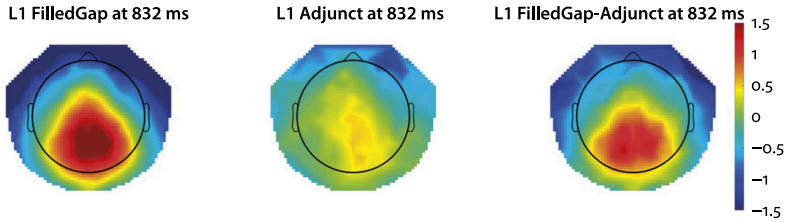
3.3.2 *L1 participants' brain responses to filled gaps: P600*

The PCA solution for the L1 group yielded a 612–996 ms window and an ROI including 33 electrodes from the central-parietal region (see Section 2 of the Appendix for details). The L1 effect is consistent with a typical P600, as shown by the ERP waveforms averaged over the ROI and the topographical maps at the peaking time (832 ms) of the selected temporal factor in Figure 6:



A.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don



B.

Figure 6. (A): L1 ERP waveforms averaged over electrodes delimited by the PCA solution. The shaded blue area indicates the time window selected for analysis (612–996 ms); (B): L1 ERP topography at 832 ms for Filled Gap (left), ADJUNCT (middle), and the difference between the two conditions (right)

For statistical analysis, we built a model including Condition (FG vs. ADJUNCT) and WM span as fixed effects and all interactions between these effects; the model converged when Participant was included as a random intercept. This analysis reveals a main effect of Condition (χ^2 (1) = 25.2, p < .001) and an interaction between Condition and WM span (χ^2 (1) = 4.34, p = .037). As shown in Figure 7 below, the amplitude difference between the conditions Filled Gap (FG) and ADJUNCT (the P600) increases as the working memory span increases, replicating the results from Hestvik et al. (2012).

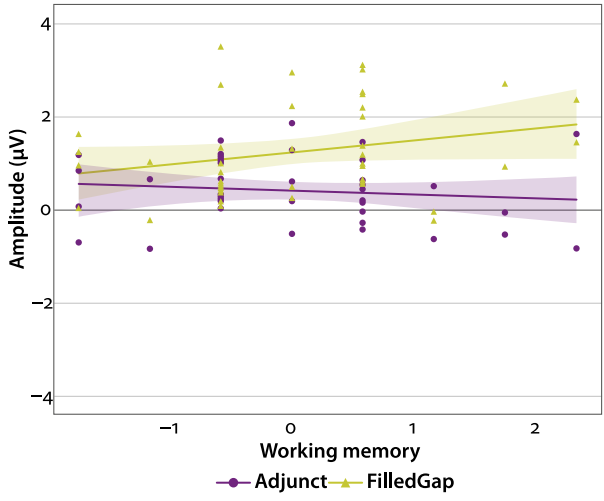
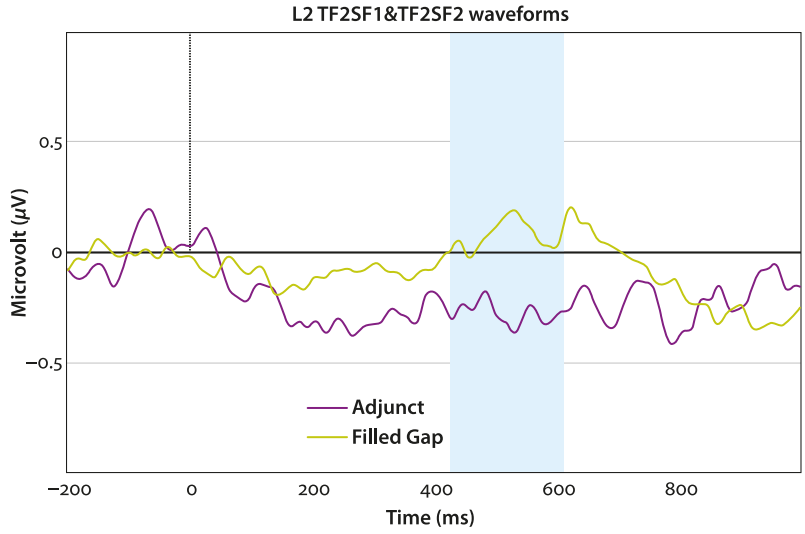


Figure 7. Effect of condition and working memory on the L1 ERP amplitude
Note. Shade denotes 95% CI.

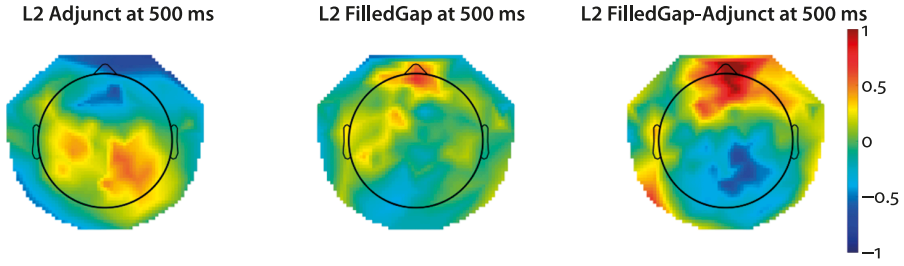
Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

3.3.3 *L2 participants' brain responses to filled gaps: Prefrontal-central positivity*

For the L2 group, the PCA solution yielded a window of 424–612 ms and an ROI including 20 electrodes in the frontal region (see Section 3 of the Appendix for details). The ERP waveform and the topographical maps in Figure 8 reveal that the FG condition elicited a greater prefrontal-central positivity than the ADJUNCT condition.



A.

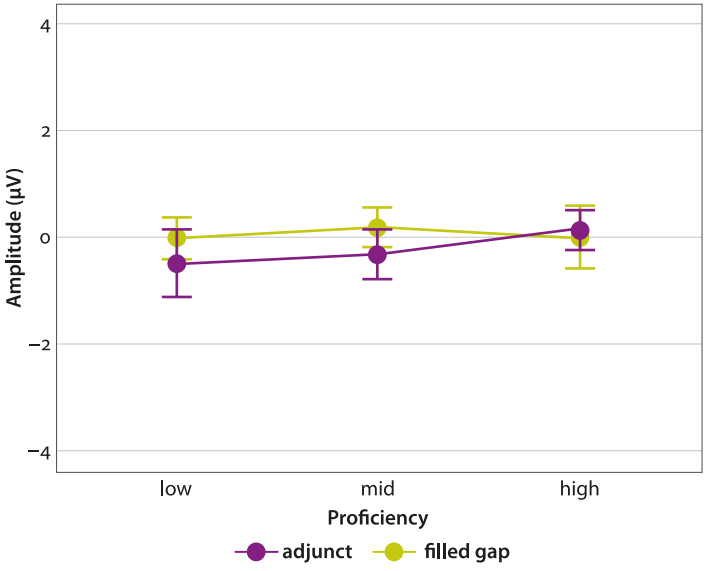


B.

Figure 8. (A): L2 ERP waveforms averaged over electrodes delimited by the PCA solution. The shaded blue area indicates the time window selected for analysis (424–612 ms); (B): L2 ERP topography at 500 ms for Filled Gap (left), ADJUNCT (middle), and the difference between the two conditions (right)

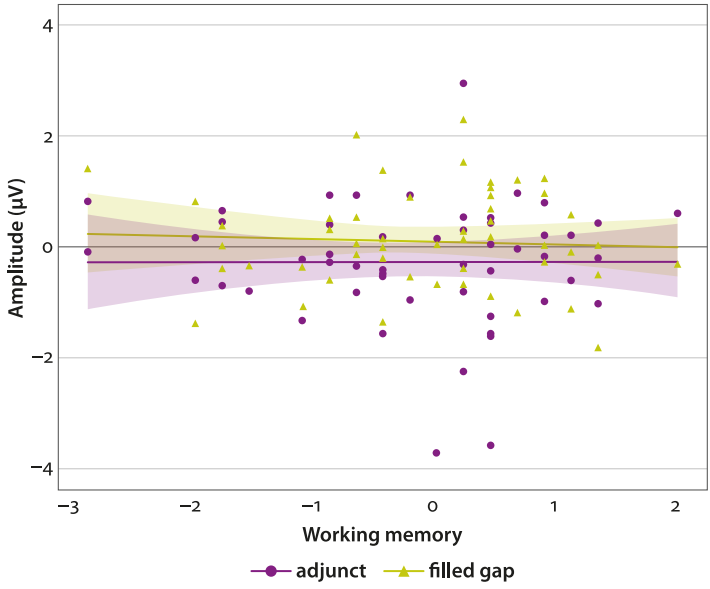
Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

The model for statistical analysis includes Condition (FG vs. ADJUNCT), Proficiency (HIGH vs. MID vs. LOW), WM span as fixed effects, and tests interactions between these effects; the model converged when Participant was included as a random intercept. This analysis revealed a main effect of Condition (χ^2 (1) = 5.48, p = .019) with no effect observed for Proficiency or WM span, as shown in Figure 9A and Figure 9B below, suggesting that the observed prefrontal positivity in L2 speakers stayed mostly the same across proficiency levels and WM differences. Since grouping the Versant scores into discrete proficiency levels might obscure the patterns that could otherwise be observed in the ungrouped data, we reran the model with the fixed effect Proficiency replaced with Versant scores as a continuous variable. The model revealed a main effect of Condition (χ^2 (1) = 9.61, p = .002), but still no effect of Versant scores or WM span, as shown in Figure 9 (C), confirming that Proficiency/WM does not affect the amplitude of the L2 ERP.

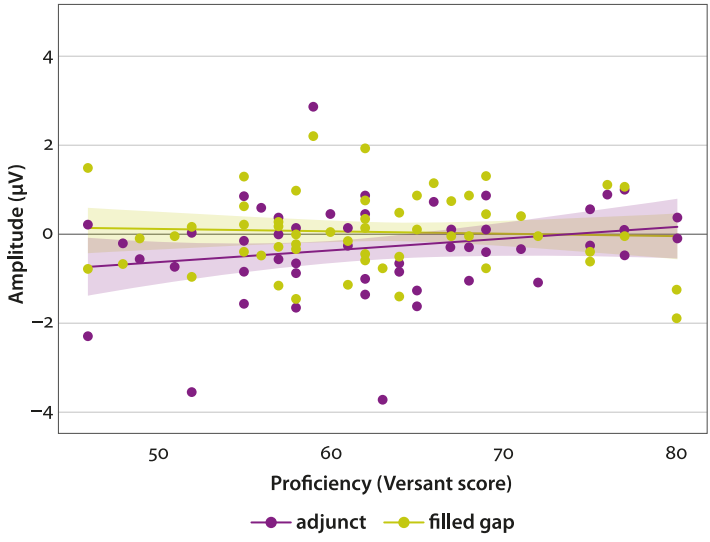


A.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don



B.



C.

Figure 9. (A): L2 ERP amplitude as a function of Condition and Proficiency level; (B): L2 ERP amplitude as a function of Condition and WM; (C): L2 ERP amplitude as a function of Condition and Proficiency score (continuous)
Note. Vertical bars and shades denote 95% confidence intervals.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

4. Discussion

Using the ERP method, this paper examines whether late L2 speakers process a filled gap in FGD in a nativelike way and to what extent L2 proficiency and WM capacity can explain any L1-L2 differences. Our results show that while the L2 speakers' behavioral performances were comparable to those of the native speakers in answering the online comprehension questions and demonstrated strong offline grammatical knowledge about FGDs, their brain responses differed categorically from those of the native speakers. First, the L2 group did not produce the L1 ERP P600 indicative of the repair attempt after syntactic processing difficulties triggered by the filled gap. Second, the only significant component for the L2 participants was a prefrontal-central positivity in the 424–612 ms time range, which was absent from the L1 speakers' brain responses. We will discuss the functional interpretations of this ERP below. Significantly, the L2 participants' brain responses were not modulated by either proficiency or WM span. Even for the L2 speakers who were highly proficient (i.e., Advanced-Mid to Superior as defined by ACTFL proficiency guidelines) and those who had large WM capacity, no brain activity remotely resembled those of the L1 speakers was found. Furthermore, the amplitude of the L2 component (e.g., prefrontal positivity) remains the same across WM span and proficiency levels. This pattern contrasts sharply with the behavioral findings, which show a clear trend of more native-like performance as proficiency and WM capacity increases, suggesting that while L2 speakers can function more and more like the native speakers as proficiency increases at the *behavioral* level, they nonetheless process the language non-natively at the *brain* level. Given that the individual difference factors tested in this study do not explain any L1-L2 brain activity discrepancies, our findings support proposals claiming drastic, persistent L1-L2 processing differences (Clahsen & Felser, 2006, 2018).

We now turn to the second question this study aims to address, namely, whether L2 gap positing involves structural cue use. To interpret the only L2 ERP found, the prefrontal positivity, we first consider the possibility of it being a frontally-distributed P600 (fP600), sometimes obtained for syntactic ambiguities and complexity (e.g., Kaan & Swaab, 2003). However, previous fP600 had occurred much later (e.g., 700–900 ms in Kaan & Swaab, 2003; 800–1100 ms in Friederici et al., 2002) and were in more central-frontal scalp regions (e.g., Friederici et al., 2002) rather than prefrontal sites like the current L2 component. Our L2 ERP can also be compared to L2 frontal positivities previously found for word order issues (e.g., Bowden et al., 2013; Andersson, Sayehli & Gullberg, 2020). For example, Bowden et al. (2013) presented written Spanish sentences in which an object NP was incorrectly placed before the verb to learners of different

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

proficiency levels. The violation elicited the LAN + P600 among the native speakers and advanced L2 speakers, but only a P600 for the less proficient L2 learners, who also produced a left-to-center frontal positivity at 300–425 ms. Additionally, Morgan-Short et al. (2012) found frontal and prefrontal positivity at 350–700 ms for similar word order violations for explicitly trained L2 speakers who attained high proficiency in an artificial language. Implicitly trained advanced speakers of the same language, however, produced the expected LAN + P600 complex. Given that the word order problem in these L2 studies also constitutes syntactic/word category violation similar to the FG effect in the present experiment, one might wonder if these two ERPs are triggered by the same processes. However, closer examination reveals notable differences between them: the word order prefrontal positivity was followed by a P600 and modulated by proficiency levels and learning environment (implicitly vs. explicitly taught), while the current L2 ERP is not. Additionally, the distribution and time course of these two L2 components are not entirely consistent.

Lastly, we compare our L2 ERP to the anterior Post-N400-Positivity (aPNP), an ERP evoked by a plausible but contextually unexpected word, such as the "mistake" relative to the control "splash" in "Bill jumped in the lake. He made a big? mistake/splash with his cannonball" (e.g., Delong & Kutas, 2020). Like our L2 ERP, the aPNP can have a prefrontal distribution and an early onset of 400 ms (e.g., Delong & Kutas, 2020; Thornhill & Van Petten, 2012) and has been replicated in at least one L2 study (e.g., Foucart, Martin, Moreno & Costa, 2014). However, the aPNP typically follows an N400, which may have a reduced amplitude (Delong & Kutas, 2020). Visual inspection of our L2 voltage results revealed a central negativity peaking at 356 ms, reminiscent of the N400, though its effect was too small to reach significance by the PCA analysis. We thus speculate that the current L2 prefrontal positivity might be an aPNP, although further research is clearly needed to verify this proposal. If our L2 component were indeed an aPNP, it would suggest that L2 speakers treated the extra NP as a contextually unexpected but plausible word rather than a violation of a syntactic/word category like the native speakers.

Although the nature of the current L2 ERP needs further clarification, it is unlikely to be triggered by any known structural-based language processes, such as those of the L1 speakers. Considering that our L2 speakers failed to produce any component indicative of structural-based processing and that L2 brain response did not become more native-like as proficiency and working memory capacity increases, our findings overall are consistent with accounts such as the SSH, which claims that L2 online parsing of certain structures underuses syntactic information even at highly advanced acquisition stages. However, we do not believe that L2 parsing is structurally shallow for all constructions; nativelike

processing patterns have been attested by numerous L2 ERP studies on sentence and morphosyntactic processing (see Tolentino & Tokowicz, 2011 for a review). Instead, we propose that the 'nativelikeness' of L2 processing varies by the type of linguistics form under investigation (and by extension the specifics of the rules involved), and nonnative processing is limited to only a few situations, perhaps where (1) underuse of structural cues does not compromise meaning computation, and (2) no overt surface feature is present to prompt the application of the relevant structural rules. In the current case of FGD, meaning can be accurately computed by directly associating the filler with the subcategorizing verb. Further, unlike morphological agreement, where overt cues (e.g., English third-person singular 's') are present in the input to relate to rule application, FGD has no surface feature to trigger computation of a full syntactic representation constrained by all its grammatical rules. Assuming that the parser and the grammar are a single system (Phillips & Lewis, 2013) and that acquisition *is* the process of parsing the input to construct the target grammar, which in turn affects how the parser processes the input (e.g., Gregg, 2003), an FGD with full syntactic details may not develop in the L2 grammar. Concurrently, the L2 parser may not – and need not – adopt a particular structural-based rule as native speakers do to process FGD-related input. In fact, the L2 learners might be motivated by efficiency to use a meaning-based, structurally shallow routine. Thus, the processing of constructions like FGDs may remain non-nativelike despite extensive language experience. This view is shared by recent accounts such as Felser (2019), who argues that when explaining any L1-L2 processing differences, it is essential to consider what linguistic cues must be extracted from the input for a given structural-sensitive constraint to be applied. We also agree with Felser (2019) that L1-L2 parsing difference should be attributed to singular causes *and* the *interactions* among different causes. Specifically, we believe that L2 use of information sources is modulated by the features of the structures being processed. To verify this point, future research could test L2 processing of other complex constructions, such as VP ellipsis and referential dependencies.

An important point that needs additional clarification is the role of L1 interference, which arguably may have caused the non-nativelike brain responses in our L2 learners. Mandarin Chinese is a *wh*-in-situ language (Huang, Li & Li, 2009), and whether movement is involved in the derivation of Mandarin *wh*-questions continues to be debated. However, most researchers agree that at least adjunct *wh*-elements in Chinese must undergo movement at LF (Huang et al., 2009). Moreover, Mandarin has overt filler-gap dependency structures with dislocated items, such as relative clauses (RCs) and topicalization. A movement analysis has been proposed for both structures (e.g., Shyu, 1995; Hsu, 2008), and existing experimental findings, though limited, indicate that the processing of

Chinese RCs and topicalization structures are similar to movement languages, including English (e.g., Dong, Rhodes & Hestvik, 2021; Lin & Garnsey, 2011). Furthermore, ERP evidence suggests that highly proficient Chinese English learners can process L2 grammatical features that are different or absent from their first language in a nativelike way (e.g., Liang & Chen, 2014). Although we do not believe that L1 transfer caused the current non-nativelike online L2 parsing patterns, replicating our study with learners of different language backgrounds is nevertheless warranted. Another promising future research direction to pursue concerns the effect of L2 instruction type. A distinction can be made regarding whether the learner received explicit, memorization-based, or implicit, immersion-style instruction. Our participants had only received the former kind of instruction, in which the target language grammatical rules are typically taught by explicit explanation in the learner's native language. However, when learners receive implicit instruction delivered with full contextualization in the target language, they may 'acquire' the grammatical rules of the language more effectively from processing a large quantity of naturistic input selected to model rule application. Recent studies (e.g., Morgan-Short, Steinhauer, Sanz & Ullman, 2012) have also produced evidence suggesting that the types of L2 exposure and training can shape L2 processing strategies and neuro-cognition.

To conclude, this paper offers novel neurophysiological evidence demonstrating distinctly different L1 and L2 brain responses to a filled gap when processing FGD sentences. Specifically, L2 results show reduced sensitivity to structural violations during gap positing, and such a nonnative pattern was not modulated by either proficiency levels or WM capacity. Our findings thus lend support to accounts such as the SSH, which claims that L2 processing prioritizes non-structural-related information, is less sensitive to grammatical constraints, and as such may be qualitatively different from L1 processing even with ultimate attainment. However, considering that nativelike L2 processing profiles have been observed for various other linguistic constructions, we propose that L2 persistent underuse of structural information only occurs when meaning can be successfully computed without detailed structure building, and in particular when the overt surface reflex is lacking to trigger related grammatical rule application.

References

Andersson, A., Sayehli, S., & Gullberg, M. (2019). Language background affects online word order processing in a second language but not offline. *Bilingualism: Language and Cognition*, *22*(4), 802–825. https://doi.org/10.1017/S1366728918000573

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Bernstein, J. & De Jong, J. H. A. L. (2001). An experiment in predicting proficiency within the Common Europe Framework Level Descriptors. In Y. N. Leung et al. (Eds.), *Selected Papers from the Tenth International Symposium on English Teaching* (pp. 8–14). Crane Publishing.

Bernstein, J. & Cheng, J. (2007). Logic and validation of fully automatic spoken English test. In M. Holland & F. P. Fisher. (Eds.), *The path of speech technologies in computer assisted language learning: From research toward practice* (pp. 174–194). Routledge.

Bowden, H. W., Steinhauer, K., Sanz, C., & Ullman, M. T. (2013). Native-like brain processing of syntax can be attained by university foreign language learners. *Neuropsychologia*, *51*(13), 2492–2511. https://doi.org/10.1016/j.neuropsychologia.2013.09.004

Caffarra, S., Mendoza, M., & Davidson, D. (2019). Is the LAN effect in morphosyntactic processing an ERP artifact? *Brain and Language*, *191*, 9–16. https://doi.org/10.1016/j.bandl.2019.01.003

Chomsky, N. (1986). *Knowledge of language: Its nature, origins, and use*. Greenwood Publishing Group.

Christensen, R. H. B. (2019). Regression Models for Ordinal Data [R package ordinal version 2019.12-10].

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369–382. https://doi.org/10.1177/0267658312443651

Cunnings, I. (2017). Interference in Native and Non-Native Sentence Processing. *Bilingualism: Language and Cognition*, *20*(04), 712–721. https://doi.org/10.1017/S1366728916001243

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42. https://doi.org/10.1017/S0142716406060024

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition*, *40*(3), 693–706. https://doi.org/10.1017/S0272263117000250

Clifton, C., & Frazier, L. (1989). Comprehending sentences with long distance dependencies. In G. M. Carlson & M. K. Tanenhaus. (Eds.), *Linguistic structure in language processing* (pp. 273–317). Kluwer. https://doi.org/10.1007/978-94-009-2729-2_8

Dallas, A., DeDe, G., & Nicol, J. (2013). An Event-Related Potential (ERP) Investigation of Filler-Gap Processing in Native and Second Language Speakers. *Language Learning*, *63*(4), 766–799. https://doi.org/10.1111/lang.12026

Daneman, M. & Carpenter, P.A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior 19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

DeLong, K.A., & Kutas, M. (2020). Comprehending surprising sentences: sensitivity of post-N400 positivities to contextual congruity and semantic relatedness. *Language, Cognition and Neuroscience*, 1–20. https://doi.org/10.1080/23273798.2019.1708960

Delorme, A., & Makeig, S. (2004). EEGLAB: an open-source toolbox for analysis of single-trial EEG dynamics, *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Dien, J. (2010). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods*, *187*(1), 138–145. https://doi.org/10.1016/j.jneumeth.2009.12.009

Dien, J. (2012). Applying principal components analysis to event-related potentials: a tutorial. *Developmental Neuropsychology*, *37*(6), 497–517. https://doi.org/10.1080/87565641.2012.697503

Dong, Z., Rhodes, R., & Hestvik, A. (2021). Active Gap Filling and Island Constraint in Processing the Mandarin 'Gap-Type' Topic Structure. *Frontiers in Communication 6*: 650659. https://doi.org/10.3389/fcomm.2021.650659

Dowens, M. G., Guo, T., Guo, J., Barber, H., & Carreiras, M. (2011). Gender and number processing in Chinese learners of Spanish–Evidence from event related potentials. *Neuropsychologia*, *49*(7), 1651–1659. https://doi.org/10.1016/j.neuropsychologia.2011.02.034

Dussias, P. E., & Piñar, P. (2010). Effects of reading span and plausibility in the reanalysis of wh-gaps by Chinese-English second language speakers. *Second Language Research*, *26*(4), 443–472. https://doi.org/10.1177/0267658310373326

Felser, C. (2019). Structure-sensitive constraints in non-native sentence processing. *Journal of the European Second Language Association*, *3*(1), 12–22. https://doi.org/10.22599/jesla.52

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, *23*(1), 9–36. https://doi.org/10.1177/0267658307071600

Foucart, A., Martin, C. D., Moreno, E. M., & Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(5), 1461–1469.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, *6*(2), 78–84. https://doi.org/10.1016/S1364-6613(00)01839-8

Friederici, A. D., Hahne, A., & Saddy, D. (2002). Distinct neurophysiological patterns reflecting aspects of syntactic complexity and syntactic repair. *Journal of Psycholinguistic Research*, *31*, 45–63. https://doi.org/10.1023/A:1014376204525

Frisch, S., Hahne, A., & Friederici, A. D. (2004). Word category and verb–argument structure information in the dynamics of parsing. *Cognition*, *91*(3), 191–219. https://doi.org/10.1016/j.cognition.2003.09.009

Gouvea, A. C., Phillips, C., Kazanina, N., & Poeppel, D. (2010). The linguistic processes underlying the P600. *Language and Cognitive Processes*, *25*(2), 149–188. https://doi.org/10.1080/01690960902965951

Gregg, K. (2003). SLA theory construction and assessment. In C. Doughty & M. Long. (Eds.). *Handbook of Second Language Acquisition* (pp. 831–865). Oxford: Blackwell. https://doi.org/10.1002/9780470756492.ch23

Hagoort, P., Brown, C. M., & Groothusen, J. (1993). The syntactic positive shift as an ERP measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Hahne, A., & Friederici, A. D. (1999). Electrophysiological evidence for two steps in syntactic analysis: Early automatic and late controlled processes. *Journal of Cognitive Neuroscience*, *11*(2), 194–205. https://doi.org/10.1162/089892999563328

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*(1), 25–38. https://doi.org/10.1017/S0272263100010457

Hawkins, R., & Chan, C. Y. (1997). The partial availability of universal grammar in second language acquisition: The "Failed functional features hypothesis." *Second Language Research*, *13*(3), 187–226. https://doi.org/10.1191/026765897671476153

Hestvik, A., Maxfield, N., Schwartz, R. G., & Shafer, V. L. (2007). Brain responses to filled gaps. *Brain and Language*, *100*(3), 301–316. https://doi.org/10.1016/j.bandl.2006.07.007

Hestvik, A., Bradley, E., & Bradley, C. (2012). Working Memory Effects of Gap-Predictions in Normal Adults: An Event-Related Potentials Study. *Journal of Psycholinguistic Research*, *41*(6), 425–438. https://doi.org/10.1007/s10936-011-9197-8

Hopp, H. (2017). Individual differences in L2 parsing and lexical representations. *Bilingualism: Language and Cognition*, *20*(4), 689–690. https://doi.org/10.1017/S1366728916000821

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hsu, C.-C.N. (2008). Revisit relative clause islands in Chinese, *Language and Linguistics*, *9*(1), 23–48.

Huang, J., Li, Y.A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166935

Jessen, A., Festman, J., Boxell, O., & Felser, C. (2017). Native and non-native speakers' brain responses to filled indirect object gaps. *Journal of Psycholinguistic Research*, *46*(5), 1319–1338. https://doi.org/10.1007/s10936-017-9496-9

Johnson, A., Fiorentino, R., & Gabriele, A. (2016). Syntactic constraints and individual differences in native and non-native processing of wh-movement. *Frontiers in psychology*, *7*, 549. https://doi.org/10.3389/fpsyg.2016.00549

Juffs, A. (2006). Grammar and parsing and a transition theory. *Applied Psycholinguistics*, *27*(1), 69–71. https://doi.org/10.1017/S0142716406060115

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?. *Linguistic Approaches to Bilingualism*, *4*(2), 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E., & Swaab, T.Y. (2003). Repair, revision, and complexity in syntactic analysis: An electrophysiological differentiation. *Journal of Cognitive Neuroscience*, *15*(1), 98–110. https://doi.org/10.1162/089892903321107855

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language*, *52*(2), 205–225. https://doi.org/10.1016/j.jml.2004.10.002

Kizach, J. (2014). Analyzing Likert-scale data with mixed-effects linear models: a simulation study. *Poster Presented at Linguistic Evidence*. Tübingen, Germany.

Kuperberg, G.R., Kreher, D.A., Sitnikova, T., Caplan, D.N., & Holcomb, P.J. (2007). The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. *Brain and Language*, *100*(3), 223–237. https://doi.org/10.1016/j.bandl.2005.12.006

Kutas, M., & Federmeier, K.D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Liang, L., & Chen, B. (2014). Processing morphologically complex words in second-language learners: The effect of proficiency. *Acta Psychologica*, *150*, 69–79. https://doi.org/10.1016/j.actpsy.2014.04.009

Lin, Y., & Garnsey, S.M. (2010). Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In *Processing and producing head-final structures* (pp. 241–275). Springer, Dordrecht. https://doi.org/10.1007/978-90-481-9213-7_12

Luck, S.J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, *27*(1), 53–78. https://doi.org/10.1017/S0272263105050035

Martin, K.I., & Ellis, N.C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379–413. https://doi.org/10.1017/S0272263112000125

McDonald, J.L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381–401. https://doi.org/10.1016/j.jml.2006.06.006

Miller, A.K. (2015). Intermediate Traces and Intermediate Learners: Evidence for the Use of Intermediate Structure during Sentence Processing in Second Language French. *Studies in Second Language Acquisition*, *37*(3), 487–516. https://doi.org/10.1017/S0272263114000588

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M.T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience*, *24*(4), 933–947. https://doi.org/10.1162/jocn_a_00119

Nakano, Y., Felser, C., & Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research*, *31*(5), 531–571. https://doi.org/10.1023/A:1021260920232

Nicenboim, B., Vasishth, S., Gattei, C., Sigman, M., & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology*, *6*, Article 312. https://doi.org/10.3389/fpsyg.2015.00312

Ojima, S., Nakata, H., & Kakigi, R. (2005). An ERP study of second language learning after childhood: Effects of proficiency. *Journal of Cognitive Neuroscience*, *17*(8), 1212–1228. https://doi.org/10.1162/0898929055002436

Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second language sentence processing. *Studies in Second Language Acquisition*, *33*(4), 563–588. https://doi.org/10.1017/S0272263111000313

Pakulak, E., & Neville, H.J. (2011). Maturational constraints on the recruitment of early processes for syntactic processing. *Journal of Cognitive Neuroscience*, *23*(10), 2752–2765. https://doi.org/10.1162/jocn.2010.21586

Phillips, C., & Lewis, S. (2013). Derivational order in syntax: Evidence and architectural consequences. *Studies in Linguistics*, *6*, 11–47.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabourin, L., & Stowe, L.A. (2008). Second language processing: When are first and second languages processed similarly? *Second Language Research*, *24*(3), 397–430. https://doi.org/10.1177/0267658308090186

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools, Inc.

Shyu, S.I. (1995). The syntax of focus and topic in Mandarian Chinese. (Doctoral dissertation). University of Southern California.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(3), 609–652. https://doi.org/10.1017/S0022226712000011

Steinhauer, K., & Drury, J.E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135–162. https://doi.org/10.1016/j.bandl.2011.07.001

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. https://doi.org/10.1177/0267658308098995

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227–245. https://doi.org/10.1080/01690968608407062

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. https://doi.org/10.1016/j.ijpsycho.2011.12.007

Tolentino, L. C., & Tokowicz, N. (2011). Across Language, space and time. *Studies in Second Language Acquisition*, *33*(1), 91–125. https://doi.org/10.1017/S0272263110000549

Van Hell, J. G., & Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research*, *26*(1), 43–74. https://doi.org/10.1177/0267658309337637

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176–190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Williams, J., Möbius, P., & Kim, C. (2001). Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, *22*(4), 509–540. https://doi.org/10.1017/S0142716401004027

Appendix

Section 1. *PCA procedures for determining the time window and regions of interest (ROI) for analysis: L1 and L2 data combined*

We pooled data from both groups and computed the difference wave (FG minus ADJUNCT) as the PCA input. We first ran a temporal PCA based on the covariance matrix constructed by treating each time point as a variable and each Participant-channel combination as an observation. Using the Parallel Test (Horn, 1965), we retained 19 temporal factors (TFs), which accounted for 92% of the total variance. Factor score matrices were constructed based on those 19 TFs, with each electrode as a variable and each participant as an observation. The factor score matrix was subsequently submitted to a spatial ICA. Six spatial factors (SFs) were retained for each TF following the same scree-test procedure for the temporal analysis. The combined temporospatial factors accounted for 61% of the total variance. To select the temporospatial factors that reflect a latent component, we first selected among the 19 TFs the ones accounted individually for more than 6% of the variance. Three TFs were thus selected: TF1 peaking at 900 ms (31% of the total variance), TF2 peaking at 512 ms (20% of the variance), and TF3 peaking at 292 ms (13% of the variance). We then examined the topographic maps of the data reconstructed from the three TFs. We found that TF1 topography showed a clear central-posterior positivity consistent with a spatial distribution of a P600 (Figure S1, left column).

Next, we examined the topographic maps of the six SFs (rescaled to microvolts) of TF1. We found that the first SF of TF1 (TF1SF1) successfully captured the central-posterior positivity of TF1 (Figure S1, right column). As a further step in confirming the temporospatial factors' relationship to the experimental manipulations, we tested each temporospatial factor with a t-test against 0 (as these factors reflect the difference waveforms), and TF1SF1 did come out signifi-

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

cant. We thus used TF1SF1 and its parent TF1 to delimit the ROI and the time window, respectively, for statistical analysis. We selected time samples that exceeded the 0.6 threshold in TF1 (which gave a time window of 740–996 ms) and the electrodes whose factor loadings exceeded the 0.6 threshold in TF1SF1. Twenty-six electrodes were selected, including: E52, E53, E58, E59, E60, E61, E62, E65, E66, E67, E70, E71, E72, E75, E76, E77, E78, E79, E83, E84, E85, E90, E91, E92, E96, and E97.

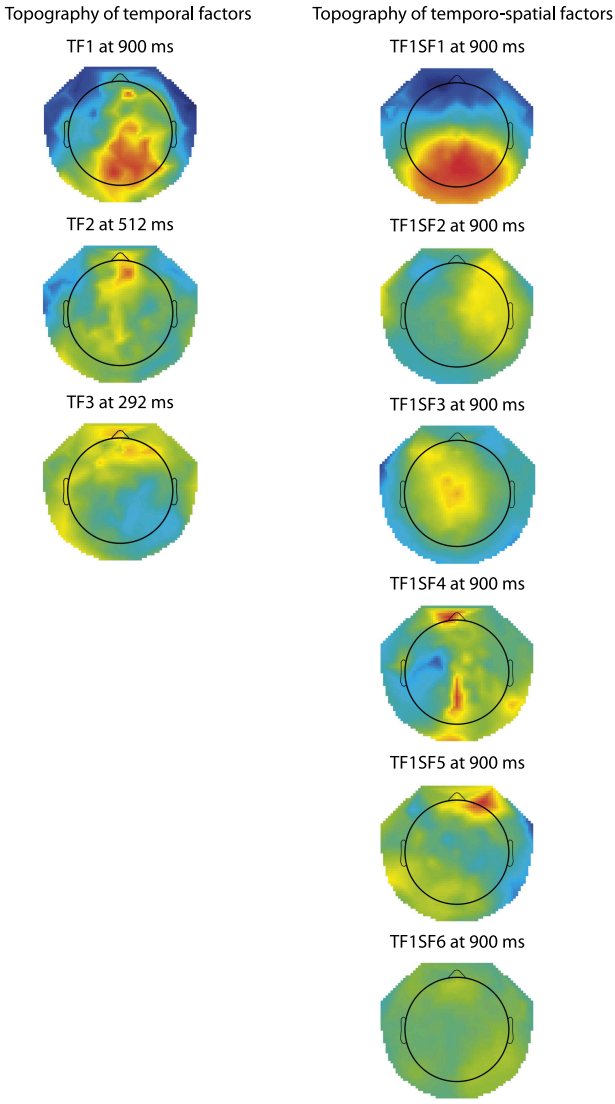


Figure S1. The Topography of the data reconstructed from the first three TFs at their peaking time (left column): The topography of the factor loadings (rescaled to microvolts) of each spatial factor extracted from TF2 at 900 ms

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

Section 2. *PCA procedures for determining the time window and regions of interest (ROI) for analysis: L1 data only*

For the L1 group, the temporal PCA yielded 18 factors, accounting for 92% of the total variance. The first three of these factors each accounted for over 6% variance: TF1 (accounts for 40%) peaks at 832 ms, TF2 (19%) at 376 ms, and TF3 (8%) at 180 ms. The spatial ICA on the 18 TFs extracted four SFs for each TF. The topography of TF1 shows a P600-like spatial distribution (Figure S2, left column), which was also captured in the topography of TF1SF1 (Figure S2, right column). As a further confirmation, the factor scores of each temporospatial factor tested against 0 also reached significance for TF1SF1. We thus retained TF1SF1 and TF1 to delimit the ROI and time window for statistical analysis, respectively. We selected the time samples with a factor loading of 0.6 or higher in TF1, which gave a time window of 612–996 ms, and electrodes whose factor loading exceeded the 0.6 threshold in TF1SF1. The selected 33 electrodes include E31, E52, E53, E54, E55, E59, E60, E61, E62, E65, E66, E67, E70, E71, E72, E75, E76, E77, E78, E79, E80, E82, E83, E84, E85, E86, E87, E90, E91, E92, E93, E96, and E97.

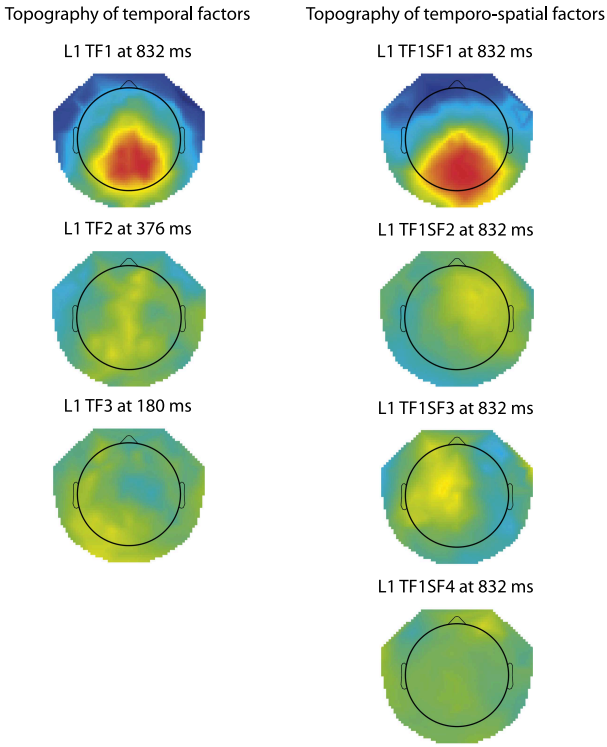


Figure S2. The Topography of the data reconstructed from the first three TFs at their peaking time (left column): The topography of the factor loadings (rescaled to microvolts) of each spatial factor extracted from TF2 at 832 ms

Section 3. *PCA procedures for determining the time window and regions of interest (ROI) for analysis: L2 data only*

For L2 participants' brain responses to the filled-gap manipulation, the initial temporal PCA retained 18 temporal factors, accounting for 91% of the variance. The first three factors each accounted for more than 6% variance: TF1 (29%) peaks at 992 ms, TF2 (21%) at 500 ms, and TF3 (11%) at 260 ms. The following spatial ICA yielded five SFs for each TF. Topography reconstructed from the three TFs shows that the only interpretable component is a prefrontal positivity for TF2 (Figure S3, left column). However, none of its five SFs by themselves showed a topography resembling that of TF2. Further inspection revealed that SF1 and SF2 each captures half of the prefrontal positivity (Figure S3, right column). We thus selected the electrodes representing the frontal positivity using both TF2SF1 and TF2SF2 as constraints. Note that SF1 and SF2 do not reflect an eye activity because the SFs were extracted based on the TFs. An eye-activity interpretation would mean that most participants moved their eyes within a limited time window where the time samples carry adequately high factor loadings in a TF – an improbable scenario. As to the time course, the L2 effect was delimited by TF2 and by using time samples with factor loadings of 0.6 or higher, which gives a time window of 424–612 ms. The ROI includes the following 20 electrodes: E1, E2, E3, E8, E12, E14, E19, E20, E22, E23, E24, E25, E26, E27, E28, E32, E48, E122, E123, E126.

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

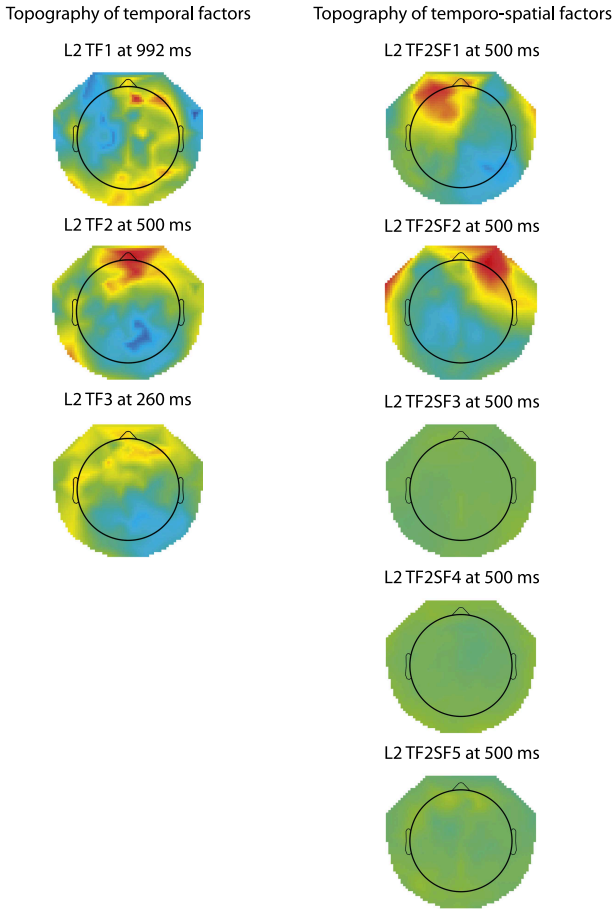


Figure S3. The Topography of the data reconstructed from the first three TFs at their peaking time (left column): The topography of the factor loadings (rescaled to microvolts) of each spatial factor extracted from TF2 at 500 ms

Address for correspondence

Zhiyin Renee Dong
Department of Languages, Literatures and Cultures
University of Delaware
30 East Main Street
Newark, DE 19716
United States
rdong@udel.edu

Accepted Manuscript
Version of record at: https://doi.org/10.1075/lab.20058.don

Co-author information

Chao Han
Department of Linguistics and Cognitive Science
University of Delaware
hanchao@udel.edu

Gabriella Hermon
Department of Linguistics and Cognitive Science
University of Delaware
gaby@udel.edu

Arild Hestvik
Department of Linguistics and Cognitive Science
University of Delaware
hestvik@udel.edu