# How Native-like is L2 Processing?: An ERP study of Tense Morphology and Filler-Gap Dependencies

Zhiyin Renee Dong, Arild Hestvik, Gabriella Hermon

University of Delaware

Author Note

Zhiyin Renee Dong, Department of Foreign Languages and Literatures, University of Delaware
Arild Hestvik, Department of Linguistics and Cognitive Science, University of Delaware
Gabriella Hermon, Department of Linguistics and Cognitive Science, University of Delaware

This research was supported in part by the University of Delaware China Alumni Award and the University of Delaware Arts and Sciences Faculty Research Award

Correspondence concerning this paper should be addressed to Zhiyin Renee Dong, Department of Foreign Languages and Literatures, University of Delaware, Newark, DE 19716. Telephone: +1-302-831-0230; Fax: +1-302-831-6459; Email rdong@udel.edu.

Abstract

This paper investigates whether adult L2 speakers can perform on-line processing of syntax and morpho-syntax in a native-like way by examining their parsing of filler-gap (FG) dependencies and tense morphology in two ERP experiments. The results indicate that although the learners can use decomposition rules and produce native-like brain responses for tense morphology, their processing of FG dependencies is qualitatively different from that of native speakers in that L2 speakers resort to semantic information without building syntactic traces. Furthermore, this parsing pattern is not affected by proficiency or working memory differences. Our results suggest partial but not full support for theories of "shallow processing" in L2. We conclude that native-like L2 online parsing depends on the target structure, and L2 shallow processing may be limited to those instances in which L2 learners would have to reconstruct abstract elements (such as traces).

## Introduction

An important research topic in Second Language Acquisition (SLA) is whether adult second language learners adopt a fundamentally different parsing mechanism from that of native speakers to process syntax and morpho-syntax online. Some researchers maintain that first language (L1) and second language (L2) processing largely share the same system (e.g., Schwartz & Sprouse, 1996; Sabourin & Stowe, 2008; Bates & MacWhinney, 1987), and non-native L2 processing can be explained by factors such as the level of proficiency (e.g., Larson-Hall, 2006; Ojima, Nakata, & Kakigi, 2005), L1 influence (e.g., Weber & Cutler, 2004), or the higher demand in L2 processing for cognitive resources such as working memory (e.g., McDonald, 2006). In contrast, other researchers argue that L1 and adult L2 processing differ qualitatively (e.g., Weber-fox & Neville, 1996; Pakulak & Neville, 2011; Hawkins & Chan, 1997). For instance, The Shallow Structure Hypothesis (SSH) (e.g., Clahsen & Felser, 2006a, b) states that adult L2 learners cannot compute fully detailed syntactic representations during online comprehension, regardless of their proficiency level, working memory capacity, and L1 influence. Instead, L2 processing is largely guided by semantic, pragmatic, and contextual information instead of structural principles. While there have been a large number of studies testing the degree to which L2 grammatical processing is native-like, the results to date are far from conclusive (e.g., Schwartz & Sprouse, 2008; Omaki & Schulz, 2011). However, what has also become increasingly evident is that L2 processing might interact with the type of syntactic construction under investigation. While the L2 studies on relative clause attachment preferences tend to generate non-native-like patterns (e.g., Jegerski, 2010; Rah & Adone, 2010), other studies on constructions like gender and number agreement have produced more consistent, native-like results (e.g., Ojima et al., 2005; Dowens, T. Guo, J. Guo, Barber & Carreiras 2011). In addition,

the roles of L2 individual factors such as proficiency and working memory in L2 processing has not been clearly specified (e.g., see van Hell & Tokowicz , 2010 for the proficiency effect; Juffs & Harrington, 2011 for the working memory effect). This is partially due to the fact that the interactions among these individual factors are not always considered, and very rarely are they examined across syntactic structures. In addressing these gaps in our current knowledge of the field, this paper presents two Event Related Potentials (ERP) experiments examining how advanced Chinese speakers of English process inflectional morphology and Filler-Gap (FG) dependency constructions. In addition to exploring the issue of whether there is a fundamental L1-L2 processing difference, these experiments also investigate whether L2 processing patterns vary across structures and how individual L2 factors interact with brain responses. Our findings suggest that L2 online parsing patterns vary for different syntactic structures, and that the L2 parsing mechanism is semantics-driven and underuses structural cues only when the target structure involves highly abstract syntactic elements such as traces.

## L2 Processing of Filler-Gap Dependencies

### L2 Processing of FG dependencies and Experiment I rationale

FG dependencies refer to the relation between a dislocated sentence constituent (typically referred to as the Filler) and its originating position (the Gap), typically where a verb assigns this item its thematic role (Phillips & Wagers, 2007), as in sentences like *The Lady (the Filler) that the doctor treated__ (the Gap) yesterday for a minor cut was from England*. FG dependencies are structurally complex and their processing requires the effective online use of abstract rules such as movement and trace positing. It is thus ideal for examining the online processing of grammatical features.

The current research on L2 processing of FG dependencies and sentence processing has produced mixed results: while some studies suggest that L2 learners have access to complex structural representations (e.g., Juffs, 2006) and can make use of many L1 parsing routines such as Active Filler Strategy (AFS) (Clifton & Frazier, 1989) in native-like ways (e.g., Williams, Möbius & Kim 2001; Williams, 2006), there is also contradicting evidence suggesting that learners are not as sensitive to structural information in online parsing (e.g., Weber-Fox & Neville, 1996; Pakulak & Neville, 2011; Hawkins & Chan, 1997). It has also been observed that L2 learners tend to rely more on semantics and plausibility information (e.g., Williams et al., 2001; Clahsen & Felser, 2006b) than native speakers. It particular, whether L2 learners can use abstract traces in FG dependencies has attracted a lot of attention. Some scholars, in line with the view that L2 processing is fundamentally different, have presented evidence arguing that L2 speakers resolve FG dependencies by integrating the dislocated item directly with the verb that subcategorizes for it, instead of using a trace as native speakers do (e.g., Marinis, Roberts, Felser, & Clahsen, 2005). Such claims, however, are contradicted by other studies (Rodríguez, 2008, Dekydtspotter et al., 2008).

Since the issue of whether the L2 parsing of FG dependencies uses pure structural cues like abstract traces or relies on non-syntactic elements bears directly on the nature of L2 online grammatical parsing, it is necessary to further the investigation with a variety of experimental methodologies and paradigms. The current studies adopt the method of ERP, which is highly suitable because of its excellent temporal resolution and its high sensitivity to highly automatic and sometimes subconscious language processes in comparison to behavioral measures (e.g., Dowens et al., 2011). In addition, ERP indices reveal the nature (syntactic vs. semantic) of the language processes investigated and are therefore very useful in examining whether the L2

parser uses structural cues (such as traces) or relies on semantic information (e.g., thematic role assignment) for online processing.

For instance, the ERP index N400, a central-parietal negative-going voltage shift that typically starts at 250-500 milliseconds (ms) and peaks at 400 ms after the violating item, indicates semantic incongruities and violations associated with a verb's arguments (e.g., Kutas & Federmeier, 2000; Frisch, Hahne, & Friederici, 2004). The Early Left Anterior Negativity (ELAN) and its related Left Anterior Negativity (LAN), however, are related to syntactic and morpho-syntactic processing. The ELAN is a negative voltage shift in the 150-250 ms range after the target stimulus, often found in the frontal region of the scalp and more pronounced on the left hemisphere. It is related to phrase structure violations (e.g., Neville, et al., 1991; Isel, 2007), or structure building that occurs extremely early (first-pass processing), and is affected by predictions concerning the word category of an upcoming item (Lau, Stroud, Plesch & Phillips, 2006). The LAN is a negative voltage shift that occurs between 300-500 ms after the onset of the violation and is obtained in the anterior position, commonly on the left side, but sometimes bi-laterally or even on the right side more than on the left. The LAN effect is often related to morpho-syntactic rule violations and processing problems (e.g., Friederici, 2002). Another syntax-related component is the P600, a positive-going voltage wave obtained between 500-600 ms and 800-900 ms post-onset of the stimulus in the parietal region of the scalp. It is often observed for various syntactic anomalies including phrase structure violations, and morpho-syntactic violations (e.g., Hagoort, Brown & Groothusen, 1993). Complex syntactic structures such as filler-gap dependencies and "reanalysis" triggered by Garden Path sentences can also elicit a P600 (e.g., Hagoort et al., 1993; Friederici, 2002). These components thus can help to

identify which information source(s) are used in L2 online computation in comparison to that of L1.

In addition to adopting a suitable methodology, it is also important to use a paradigm designed specifically to test the use of abstract traces in FG formation. Among various studies in the L1 FG processing research that suggest abstract trace positing (e.g., Love 2007; Gibson & Warren, 2004), the ERP studies conducted by Hestvik et al. (2007; 2012) are particularly suitable for replication with L2 speakers. The studies presented auditory stimuli such as *The zebra that the hippo kissed * the camel on the nose ran far away.* The extra noun after the subcategorizing verb *kissed* illicitly fills the assumed gap site. When the parser attempts to restore the NP *the zebra*, ungrammaticality is recognized and the corresponding ERP component should result. If the relation between the verb and the filler is formed based on argument structure, thematic role assignment, or other semantic information, then the violation should generate the ERP component N400 at the filled gap site in comparison to the control condition. However, no N400 was found by Hestvik et al. (2007; 2012). Instead, the ELAN, indicative of structural violations, was obtained.

Based on the Neurophysiological Time Course model for syntactic processing proposed by Friederici and her colleagues (Friederici 1995; Friederici, Hahne & Mecklinger 1996), Hestvik et al. (2007) argued that the timing of multiple ERP components for any grammatical violation is crucial for the interpretation. The ELAN is typically observed only 100-200 ms after the offending item, suggesting problems in first-pass, highly automatic structure building. At this early stage, the parser only attends to the minimal structural information such as word category specifications. The second stage concerns the evaluation of the verb's argument structure and semantic fit, during which the related violations would elicit a LAN and/or an N400. In the last

stage, the parser reanalyzes (if necessary) and conducts the final consolidation of information (integration), when the P600 usually appears. This is why the P600 is often associated with the "second-pass" syntactic processing (e.g., Kaan, Harris, Gibson & Holcomb, 2000). Trace building occurs in the earliest stage as part of phrase structure building, as it is an identical copy of the moved element that is not phonetically realized (Chomsky, 1986). Therefore, in Hestvik et al. (2007), a trace is built as soon as the verb *kissed* is processed, filling in the projected direct object position. The parser then expects the next item to be of a different word category than NP. However, as it encounters the extra, unexpected NP, ungrammaticality occurs and the ELAN results. If the violation was recognized in the later stage, when the verb's arguments and thematic role assignment are evaluated, the N400 should appear. In addition, studies investigating ERP correlates of combined syntactic/semantic violations showed that problems in the early stage such as word category violations often block further syntactic processing. Consequently, in cases with both phrase structure violations and semantic anomalies, only an ELAN will be obtained, and not an N400 (Friederici et al., 1996).

In sum, Hestvik et al. (2007; 2012) provide L1 evidence for FG trace positing and an ideal paradigm for examining whether L2 learners use traces (as native speakers do) or resort to nonstructural information (a non-native-like strategy) to resolve FG dependencies. If an ELAN or LAN is also obtained for the L2 learners, then they are similar to native speakers in using abstract syntactic rules to build FG dependency traces. Given that L2 processing is memory taxing (e.g., McDonald), a LAN can be treated as a "delayed" ELAN, as found by Hestvik et al. (2012) for the native speakers with lower WM capacities. However, if the L2 learners can only resort to verb argument structure, thematic role assignment, and other semantic information to form the FG dependency without incorporating syntactic details, then the N400 should result,

indicating that they treat the ungrammaticality as a semantic violation (i.e., failure to integrate an extra argument). A third potential outcome with only the P600 but no anterior negativity would suggest that while there is a significant difference between L1 and L2 parsing, L2 learners can still make use of some "second-pass" syntactic rules.

## Method

**Participants.** 57 subjects (38 female and 19 male) participated in Experiment I. One subject was excluded from all data analyses because of data collection errors. Two more subjects were excluded from the ERP data analyses for high bad trial percentages (see the data processing section below for details). The average age of the participants is 24.4 years (SD= 2.53, Range=18-30). They are all late Chinese learners of English who studied English mostly in classroom settings, with an average length of formal English instruction of 10 years and 2 months. The average age of the first exposure to English (English class) is 10 (SD=3.73, Range 3-15). None of them had lived in an all-English-speaking environment before age 14. Prior to the experiment, they had lived in English-speaking countries for an average of 35.9 months (SD=19.5, Range=7-96). None of the subjects have any neurological impairment, and all except one are right-handed. They were paid $20-$40 for their participation[1] and gave informed consent before the experiment.

**Overall procedures.** Experiment I has 4 components, administered to the subjects in the following order: (1) the Versant English proficiency test, (2) the working memory test, (3) the ERP task vocabulary drill and practice run, followed by electrode net application and the EEG recording session, and (4) the Paper-and-Pencil Acceptability Task (see below). The entire experimental session with all four tasks lasted for approximately two and a half hours.

[1] Depending on the experiment duration.

**Proficiency test.** The English proficiency level was determined by the results of the Versant English Test (Pearson Plc). Versant English is a fully automated spoken English test administered over the phone or computer. It has been widely used around the world for admission and job placement purposes, and its validity is high (e.g., Bernstein & Cheng, 2007). Versant English aligns with The Common European Framework of Reference (CEFR) and its U.S. counterpart the American Council on the Teaching of Foreign Languages (e.g., Bernstein & De Jong, 2001; Baztán, 2008) and is well correlated with established proficiency tests such as the TOEFL iBT Speaking Test. The subjects scored an average of 59.2 out of 80 points on this test (SD= 8.5, Range=47-80), indicating that on average they are advanced-low[2] speakers of English by the guidelines of the CEFR and ACTFL. Three proficiency groups were thus constructed, and Table 1 below shows the alignment between the proficiency guidelines of the CEFR and ACTFL:

Table 1

Versant English scores and L2 proficiency levels by the ACTFL and CEFR standards

| Versant | CEFR | ACTFL | Proficiency Index in this study | Number of participants in Exp. |
|---|---|---|---|---|
| 78-80 | C2 | Advanced-high/Superior | HIGH | 2 |
| 69-78 | C1 | Advanced-mid | HIGH | 15 |
| 58-68 | B2 | Advanced-low | MID | 24 |
| 48-57 | B1 | Intermediate-high | LOW | 14 |
| 38-47 | A2 | Intermediate-mid | LOW | 2 |

[2] The ACTFL Proficiency Guidelines define advanced-low speakers as "are able to handle a variety of communicative tasks. They are able to 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).

**Working memory test.** The working memory test used in current study is an audio version of the Harrington and Sawyer (1992) reading span test[3], a reliable L2 version of the most widely used WM test, the Self Paced Reading Test (RST) by Daneman and Carpenter (1980) (e.g., Martin & Ellis). A total of 42 sentences, divided into 4 levels with three sets of sentences at each level, were used as stimuli. Half of the sentences are ungrammatical and have a disrupted word order in the middle or at the end of the sentence. The subjects are asked to make grammaticality judgments while also remembering the sentence-final words. The number of sentences contained in each set increases over the levels, making it more and more difficult to recall all the final words in a set. The WM test was programmed and delivered in E-prime (Schneider, Eschman, & Zuccolotto, 2002) in order to collect data for the response time, accuracy of the acceptability judgment, and the number of words correctly recalled.[4] On average, the participants were highly accurate (96.8% accuracy, Range=92%-100%) with grammaticality judgments and recalled 31.6 words out of the total 42 words (SD=0.6, Range=19-41). The median listening span test score was 33. The participants were assigned to the low WM group if their score was lower than 33 and to the high WM group if their score was equal to or higher than 33. This procedure resulted in 27 low WM subjects and 29 high WM subjects.

**Acceptability Task.** To better interpret the online ERP measures, it is important to examine the L2 participants' off-line grammatical knowledge of long-distance filler-gap dependencies. A paper-and-pencil acceptability task in the format of a questionnaire was therefore administered to the L2 participants after the ERP session, in which they rated the

---

[3] All stimuli were read by a female native speaker of English at a normal speaking rate and were digitally recorded using 16 bit resolution and a 22,050 kHz sampling rate.

[4] A lab assistant was on site to administer the test and to record the final words manually as well.

sentences' acceptability on a 7-point scale (1 being completely not acceptable and 7 being perfectly acceptable). A group of 37 native English speakers (22 female, 15 male) were recruited to be the acceptability test controls. All of them are monolingual undergraduate students from the University of Delaware, with an average age of 19.8 (SD=2.2, Range=18-29). All of them gave informed consent and completed the background questionnaire. For their participation, they were awarded a small amount of extra credit in their first-year Chinese foreign language course.

30 sentences structurally identical to the stimuli used in the ERP tasks were constructed with different vocabulary items and were used in the Acceptability Task. Twelve of these sentences are target items, among which 6 are the same as the items in the ungrammatical condition in the ERP test, adopted from Hestvik et al. (2012). These sentences should receive a low acceptability score if the rater is sensitive to the "filled gap" violation. The other six items are grammatical and are identical to the items in the other ERP test conditions (see ERP test materials below for details). In addition, 18 filler sentences of various degrees of acceptability (Sprouse and Almeida, 2012) were incorporated. The target sentences were counter-balanced and randomly distributed among the filler sentences.

After the data were collected for both the L1 and L2 groups, repeated measures ANOVAs (grammaticality x proficiency/WM Groups) were conducted for the Chinese group. Their average rating for the grammatical sentences is 5.95 (SD=0.87), and the average for the ungrammatical ones is 2.59 (SD=1.24). The ANOVA revealed a main effect of grammaticality ($F(1,56)= 296.7$, $P<0.001$), confirming that the learners are sensitive to the filled-gap violations. Further examination of the proficiency group variable showed a significant interaction between proficiency group and grammaticality ($F(2,54)=3.775$, $p<0.05$), such that the higher the proficiency, the bigger the difference in ratings between grammatical and ungrammatical

sentences. The L2 ratings were marginally affected by WM capacity, as the ANOVA revealed a marginally significant interaction between grammaticality and WM group (F (1,55)=3.65, p=0.061). When compared to the native speakers, the L2 group rated the sentences similarly, as seen in Figure 1 below. For the L1 group, the mean acceptability rating for the grammatical sentences is 6.09 (SD=0.48), and for the ungrammatical sentences the mean rating is 1.95 (SD=0.08).

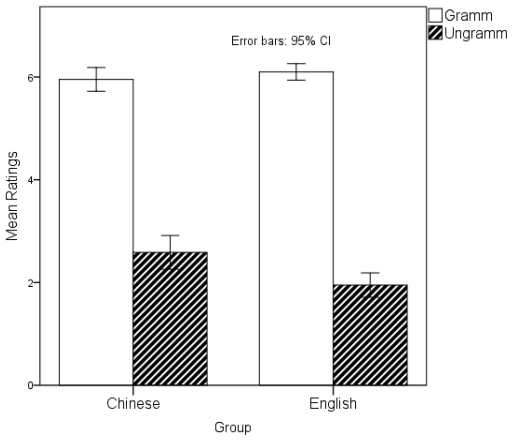



Figure 1. Acceptability data for the L1 and L2 Groups.

Although a between group repeated measures ANOVA revealed a significant interaction between group (English vs. Chinese) and grammaticality (F (1,92)=8.5, p=0.004), suggesting that the English native speakers are more sensitive to the violations than the Chinese group, no significant between-group difference was found by the ANOVA (F (1,92)=3.09, p=0.083). When the L2 subjects with low proficiency were removed, the ANOVA reported an insignificant interaction between group and grammaticality (F (1,75)=3.57, p=0.07), and when only the high proficiency L2 learners were compared to the native speakers, the ratings of the two groups became identical, as the grammaticality x group interaction is highly insignificant (F (1,50) =0.15, p=0.7). These results confirm that the L2 group has the necessary grammatical knowledge

to process the complex filler-gap dependency structures, and L1 and L2 judgments of grammaticality are highly comparable when L2 proficiency is controlled.

**The ERP test.** The stimuli and the design of the ERP experiment are based on those of Hestvik et al. (2012), with a slight modification to the comprehension question. Four experimental conditions were included, as summarized in Table 2 below:

Table 2

Example sentences from each experimental condition

| | **Condition** | **Sample sentence** |
|---|---|---|
| A. | Ungrammatical | The zebra that the hippo kissed *the camel on the nose ran far away. |
| B. | Adjunct | The weekend that the hippo kissed the camel on the nose, it was humid. |
| C. | Object | The zebra said that the hippo kissed the camel on the nose and then ran far away. |
| D. | Trace | The zebra that the hippo kissed on the nose ran far |

The critical comparison is between the ungrammatical condition (hereafter UNGRAMM) and the adjunct (ADJUNCT) condition. The two conditions are identical before and in the critical region (the verb *kissed* and the noun immediately after it), except that in the UNGRAMM condition the extra noun *the camel* causes a "filled-gap" ungrammaticality. The object (OBJECT) and trace (TRACE) conditions are included as fillers, to reduce expectations about ungrammaticality. Eight lists (4 in script A and 4 in script B) of 32 sentences each were constructed, resulting in a total of 256 experimental sentences. These sentences were counter-balanced across scripts and the delivery was randomized and counter-balanced across subjects

(see Hestvik et al., 2012 for details). Each sentence presentation was followed by a comprehension question (See Hestvik et al. (2012) for details[5]).

***ERP EEG procedure.*** After the ERP task instruction and some practice, the subjects were then drilled on some of the vocabulary items that were considered difficult by pilot study participants. After they were considered proficient with the task procedure and the vocabulary, the electrode net was applied and the subject was seated in a chair with a table top in a sound-attenuating booth. The participants were instructed to listen to each sentence and the comprehension question following it, delivered by a computer. The two pictures described above then showed up in the left and right lower corners of the computer screen. The subjects were asked to indicate their picture selection by pressing the buttons in the corresponding locations on a response box. The experiment was programed using the E-Prime software (Schneider et al., 2002) and was divided into 4 blocks of 64 sentences randomly presented to the subjects. The subjects were offered a break between blocks. The entire EEG recording session took about one hour and fifteen minutes.

***Data collection, EEG recording, and data analyses.*** The EEG was recorded with 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. Cz online was used as a reference, and the electrode impedances were kept below 50kΩ. The continuous EEG was divided into epochs of 1400 ms for each trial by time locking to the onset of the critical noun phrase (at the beginning of the article *the*). Baseline correction was performed using a 200 ms baseline period (before the onset of the noun phrase) as

[5] The comprehension questions were adopted from the Hestvik et al. (2007, 2012) studies. A minor modification was introduced in the present study. Since the ungrammaticality in the UNGRAMM condition could cause confusion and negatively affect the subjects' judgment accuracy, we coded comprehension questions for the UNGRAMM condition as correctly answered no matter what responses the participants gave.

a reference signal value. For artifact correction, the bad channels were replaced, eye blinks were subtracted, and then the eye movements were corrected using ICA Dien (2010). This order of procedures could potentially cause the eye blink channels (some with substantial voltage fluctuation) to be registered as bad channels and replaced instead of corrected. The data was therefore re-examined by conducting the artifact detection in the reverse order. The grand average patterns resulting from these two orders were determined by visual inspection to be identical. Thus, the voltage data derived via the initial data processing procedure was kept.

The artifact correction and bad channel replacement resulted in the removal of an average of 22.5% of the trials per subject (SD=0.034). Subjects with more than 30% bad trials were not included in the ERP analysis, leaving 54 subjects for the ERP data analyses. The subjects were distributed across three proficiency groups, High (17), Mid (21), Low (16), and two working memory groups, HighWM (29) and LowWM (25). The data for all subjects were included in the behavioral data analyses, and trials that were not responded to correctly were also included in order to prevent excess loss of trials and power. The ERP average was computed for each condition, each subject, and each relevant electrode site and referenced to the average voltage of all electrodes.

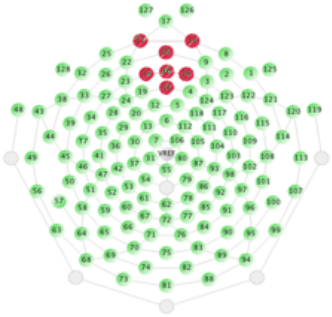
For the behavioral data, the accuracy of the responses was recorded via E-Prime (Schneider et al., 2002) and submitted to a 2x2 mixed factorial repeated measures ANOVA, with condition (3) (excluding the UNGRAMM condition) as a within-subject measure. When the Chinese behavioral data were interpreted in comparison to the English results reported in Hestvik et al. (2012)[6], the Mann-Whitney U test, the non-parametric version of the regular independent t

[6] The authors of this paper were able to obtain the behavior data of 30 native speakers from the Hestivk et al. (2012) study to facilitate direct comparison of the L1 and L2 speakers.

test, was used because of a normality issue and the significantly different sample sizes across studies.

For the voltage data, the entire set of electrodes was divided into regions based on initial visual inspection, such that an average over those regions could be computed and used as one of the dependent measures in the ANOVA. Two regions were identified based on the initial visual inspection of the grand average line plot: one at the frontal-central region of the scalp, which shows a positivity (Frontal-Central), another showing a widely distributed negativity in the mostly central region with some extension to the posterior area (Central). Figure 2 shows the arrangement of electrodes on the electrode net used:

**Frontal-Central Positivity**
Electrodes: 10, 11, 14, 15, 16, 18, and 21



**Central Negativity**
Electrodes: 7, 30, 31, 36, 37, 41, 42, 46, 47, 51, 52, 53, 54, 55, 60, 61, 78, 79, 80, 85, 86, 87, 92, 93, 97, 98, 101, 102, 103, 104, 105, 106, 108, and Cz

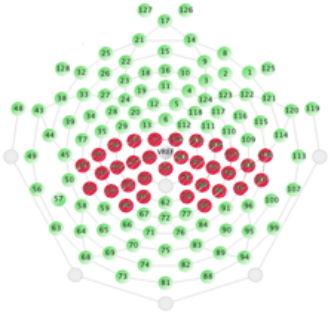


Figure 2. The electrode distribution in two critical regions (highlighted) of the 128- Channel HCGSN v.1.0 net

Another dependent measure for the voltage data analyses is time window. A total of seven 200 ms time bins were constructed for the 1400 ms epoch starting from the onset of the critical noun phrase. The mean amplitude over these time bins was computed for each electrode region for each subject and condition. The average amplitude obtained was submitted to a

mixed-factorial repeated measures ANOVA with time x condition (2) (and laterality (2) when necessary) as within-subject factors. To measure the effect of working memory capacity and proficiency level, WM group (2, high and low) and proficiency level (3, high, mid, and low) were included as between-group factors. When appropriate, *p*-values were adjusted by using Greenhouse-Geisser (1959) correction for violation of the assumption of sphericity. In addition, significant interactions between condition, time, and electrode region were followed up by planned orthogonal contrast analyses. Lastly, regression analyses were conducted between voltage data and Versant score and working memory score.

**Results.**

**Behavioral results**. The overall mean accuracy for the Chinese group is 79% (SD=7.7%). There is a significant difference among conditions (ANOVA F (1,55)=87.34, p<0.00). A post-hoc Scheffe test showed that the accuracy rate in the OBJECT condition is significantly lower than the other two conditions, which do not differ significantly from each other. This pattern is observed with the native speakers as well, as reported in Hestvik et al. (2012). The mixed-factorial repeated measures ANOVA showed that there was a main effect of WM group (F (1, 55)=11.002, p=0.002) and condition (F (2, 110)=42.49, p<0.001), as well as a significant interaction between WM group and condition (F (2, 110) =5.25, p=0.007), such that the high WM group performed better than the low WM group in all three conditions, and in particular in the OBJECT condition. To examine the role of proficiency, a main effect was found for proficiency group (F (2, 54) =7.603, P=0.001), as well as for condition (F (2, 108)=39.26, P<0.001), such that the high proficiency subjects were more accurate than the low proficiency subjects. Further regression analyses confirmed that there was a weak but significant correlation between WM score and accuracy rate ($R^2$=0.25, F (1, 55)=18.3, P=0.001). There was also a

significant correlation between Versant score (proficiency) and accuracy rate (R2=0.26, F (1,55)=18.87, P=0.000).

The overall accuracy rate of the Chinese subjects (79%, SD= 7.7%) is slightly lower than that of the native English speakers, who showed 86% (SD=5%) accuracy in the three conditions combined. The direct comparison between the two groups is illustrated in Figure 3:

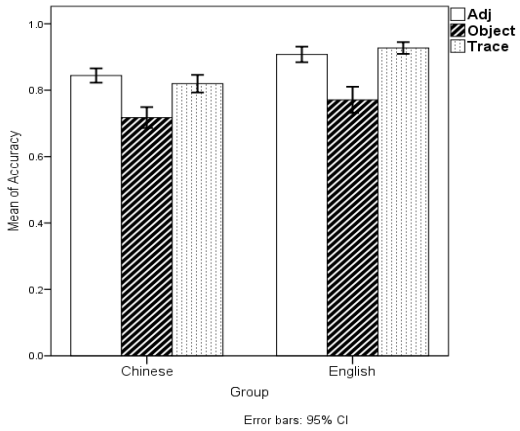



Figure 3. Comprehension question accuracy rate by condition for L1 and L2 groups

A non-parametric version of the independent samples t test, the Mann-Whitney U test[7], rendered a Z value of -1, p=0.3, indicating that the difference between the native speaker and learner groups is not significant. Another similarity shared by the native speaker and L2 groups is that both groups were significantly less accurate in the OBJECT condition than in the other two conditions, which are not very different from each other.

**ERP voltage results.** Visual inspection for a grammaticality effect revealed very different ERPs for the L2 subjects from those of the native speakers. As shown in Figure 4, the red regions represent positive voltage shifts (positivity) and the blue regions represent the

[7] The Mann-Whitney U test was used for the following reasons. First, the sample sizes are very different for the native speakers (30) and the learners (56), and the ratio exceeds the 1.5 that is recommended for regular t tests. Secondly, the distribution of the accuracy scores among the native speakers group is not normal (a Shapiro-Wilk test yielded a *p*-value of 0.02).

negative voltage shifts (negativity). As described earlier, there is no ELAN or LAN observed for the L2 speakers. A widespread N400-like negativity is observed in the central posterior region, and a positivity is evident in the front-central region of the scalp, slightly spreading into the anterior regions of both hemispheres. No P600 was observed.

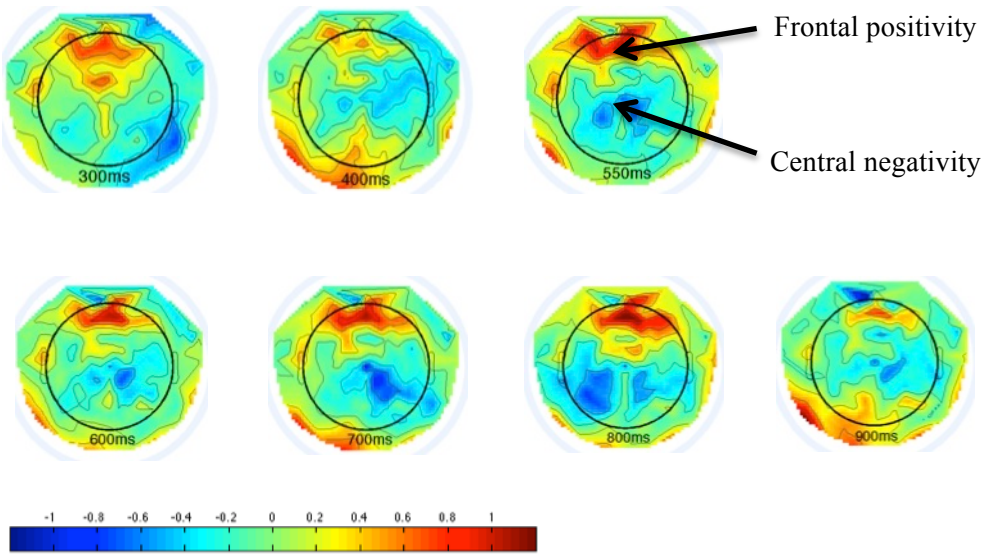



Figure 4. Grand average top plot for the grammaticality effect in difference wave forms (ADJUNCT control condition voltage minus UNGRAMM condition voltage)

To rule out the possibility that potential ELAN/LAN effects were masked by the variability in proficiency level and working memory capacity, additional analyses by proficiency and WM group were conducted based on the mean voltage of the electrodes in the left anterior region computed for each 100 ms time window starting from 300 ms to 1000 ms. The ungrammatical condition elicited positivity in all three proficiency groups in largely the same pattern over the time span. No negativity was found when the WM groups were separated either, with the low WM group actually demonstrating more positivity than the high WM group. The ANOVAs yielded no significant group effect ($F(1,52)=1.47$, $p=0.231$), no interaction between

WM group and grammaticality (F (1, 52)=0.683, p=0.412), and no three-way interaction among WM group, grammaticality, and time (F (6, 312)=2.044, p=0.06), suggesting that the difference in WM capacity doesn't change the positivity obtained in the LAN area.

Figure 5 below shows the two ERPs, the central negativity (in the right panel) and the frontal central positivity (in the left panel), as they develop over the time interval. The negativity in the central region that starts at around 250-300 ms after the onset of the offending extra noun phrase peaks at around 500 ms and is sustained into the later time windows. In the frontal-central area, a clear positivity in the UNGRAMM condition was observed in the frontal-central region between 300 ms and 1000 ms:

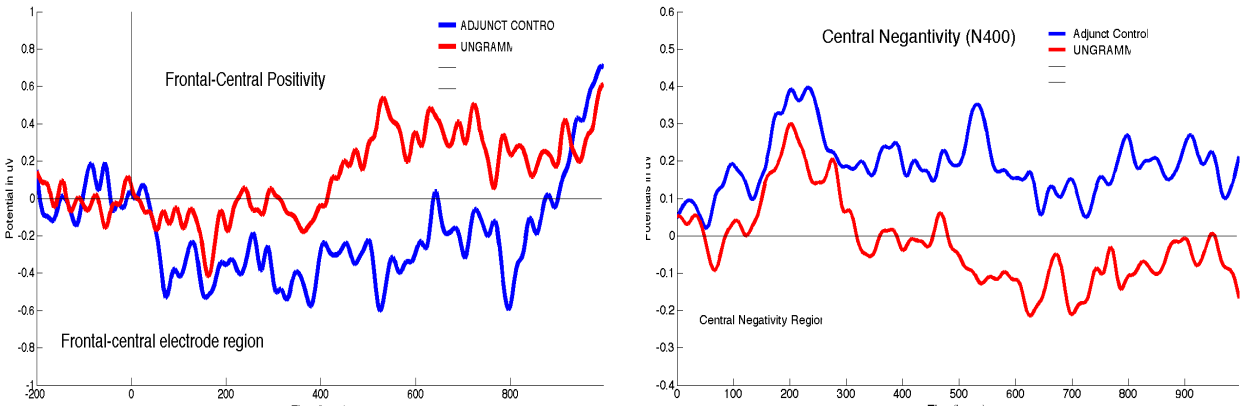



Figure *5*: The central negativity (on the right) and frontal-central positivity (on the left) waveforms for the grammatical (blue) and ungrammatical (red) conditions

The central region mean voltage was computed for each 100 ms time window starting from 300 ms to 1000 ms for each subject and each condition. An ANOVA revealed a main effect of condition (F (1, 53)=11.11, p<0.001), suggesting that there is a significant difference between the grammatical and ungrammatical conditions from 300 ms on. Planned orthogonal tests reported significance in every time window starting at 300 ms (t=3.02, p<0.001 for the 300-400

ms window) and on. Although the observed negative component seems to maximize slightly later and lasts longer than the typical N400 found in native speakers (Kutas & Fedemeier, 2011), it has been reported that L2 learners and bilinguals tend to have a delayed N400 (e.g., Ardal, Danald, Meuter Muldrew & Luce, 1990). Its more central as opposed to central-posterior distribution is expected for auditory stimuli and tasks with picture identification (Kutas & Fedemeier, 2011). Therefore, this ERP can be functionally considered to be the N400.

To inspect how proficiency and working memory affected the N400, the between-subject factors proficiency (3) and WM (2) were added to the above ANOVA separately. No significant interaction was found between the N400 and proficiency group ($F(2, 51)=0.78$, $p=0.4$), nor between the negativity and the WM group (($F(1, 52)=1.38$ $p=0.25$). By visual inspection, the N400 seems to be the most prominent for the High proficiency group, especially in the 500-700 ms time window, and least visible for the Mid proficiency group participants. Similarly, it is the high working memory group that had a larger N400, although the difference is not statistically significant. Proficiency (Versant) scores and WM scores were then run with the voltage average, and no significant correlation was found for proficiency ($r=0.021$, $F(2, 53)=0.02$, $p=0.88$), or WM capacity ($R^2=0.024$, ($F(1, 53)=1.367$, $p=0.248$), confirming that these individual differences don't predict the voltage changes.

For the frontal-central positivity, an ANOVA revealed a significant grammaticality effect ($F(1, 53)=4{,}677$, $p=0.035$). In addition, the interaction between time and grammaticality is also significant ($F(6, 52)=2.308$, $P=0.034$), demonstrating that the ungrammaticality effect changes over time. Planned orthogonal contrast tests revealed that the positivity became significant during the 500-600 ms time window ($t=-2.286$, $p=0.006$), approached significance during the 600-700 ms time window ($t=-1.94$, $p=0.057$), became significant again during the 700-800 ms

time window (t=-2.126, p=0.03), and then fell slightly under significance during the 800-900 ms time window (t=-1.96, p=0.055). The positivity could be summarized as largely significant from 500-900 ms after the onset of the extra noun phrase. Figure 7 shows the frontal positive average waveform and the pairwise t tests (two-tailed) for the voltage difference scores for each time window.

Adding WM and proficiency as between-subjects factors, an ANOVA reported that neither WM (F (1, 52)=1.416, p=0.24) nor proficiency (F (2, 51)=0.879, p=0.21) interact with the positivity, although visual inspection seems to indicate that the positivity is mostly carried by the high WM group and the High and Low proficiency groups. Furthermore, no significant associations were found between either WM score or proficiency score (Versant score) and the ERP measures, as the results of regression analyses show ($R^2$=0.007, p>0.05; $R^2$=0.02, p>0.05, respectively).

Lastly, no P600 was observed. To rule out delayed potential positivity in the central posterior region, an ANOVA was run on the last 4 100 ms time windows (600 ms and on) and reported no significant time effect (F (3, 53)=0.862, p=0.462) and no grammaticality effect (F (1, 53)=1.284, p=0.262). A planned orthogonal contrast analysis showed that the positivity is significant only in the last time window of 900-1000 ms (t=-2.238 p=0.029). This is clearly different from that of a typical P600, which starts around 400-500 ms after the onset of the violation and peaks around 600 ms. The ANOVA didn't find any interaction between proficiency group and the grammaticality effect (F (2, 51)=0.19 P=0.828), nor a between-group effect (F (2, 51)=0.21, p=0.811). Furthermore, no interaction between WM and proficiency and grammaticality was found (F (1, 52)= 0.181, p= 0.672). A significant between-WM group effect ((F (1, 52) =4.63, p=0.036) was found, and according to visual inspection the difference is due to

how the positivity changes over the time windows for the two groups. Regression analyses indicated no significant correlation between either Versant score ($R2=0.015$, $F(2, 53)=0.012$, $p=0.9$) or working memory score ($R2=0.039$, $F(1, 53)=0.083$, $P=0.73$) and voltage changes.

**Experiment I discussion**

The behavioral findings of Experiment I suggest that proficiency-controlled L2 participants were no different from the native speakers in both their off-line and online behavioral responses. They have the appropriate grammatical knowledge to process long-distance FG dependencies. Furthermore, positive correlations between proficiency level, WM capacity, and accuracy for online and off-line responses were also found, confirming that L2 offline performance improves as a function of the two individual difference factors under consideration. The ERP patterns of the L2 participants, however, are dramatically different from those of the native speakers when compared to the results of the Hestvik et al. (2007, 2012) studies. First, there was no ELAN or LAN found for the L2 speakers in the anterior region of either hemisphere, regardless of their proficiency level and working memory capacity. In addition, no P600 was obtained for the L2 speakers either. More importantly, a negative-going potential at the central-posterior region was detected starting at 250-300 ms after the offending extra noun phrase and peaked during the 500-600 ms window. As explained earlier, this potential fits into the profile of the N400 generated among L2 learners/bilinguals for tasks involving picture judgments and auditory stimuli. Based on the off-line and behavioral data, it is very unlikely that the L2 participants didn't recognize the violations in the UNGRAMM condition. The absence of the ELAN/P600 and the N400 in combination thus seems to suggest that the L2 learners simply didn't treat the violation as syntactic in nature. Instead, what possibly happened was that the L2 learners formed the filler-gap dependency based on semantic-driven heuristics,

specifically verb argument requirements, without positing the abstract trace. Furthermore, when these components were analyzed together with proficiency level and WM capacity, no reliable interactions nor correlations were found, suggesting that these two factors had no affect on the L2 learners' ERP measures. This pattern contrasts sharply with that of the behavioral results reviewed above, which indicated that the L2 performance was indeed affected by individual differences. Taken altogether, the combined brain response patterns and the lack of a modulation effect from WM capacity/proficiency suggest that the L2 processing of FG dependencies is categorically different from L1 parsing in lacking trace building and in exhibiting an over-reliance on non-structural information.

A novel ERP finding of Experiment I is the frontal central positivity, whose timing and amplitude appear to correspond with, but are not identical to, those of the N400. A similar frontal positivity, named the frontal Post-N400 Positivity (PNP), has been reported in the literature in association with the N400 (See Van Petten & Luka, 2012 for a detailed review). The PNP appears when there is a highly unlikely *final* word (high vs. low cloze probability) (e.g., Fedemeier & Kutas, 2005), and when lexical predictions are not confirmed (Delong, Urbach, Groppe & Kutas, 2011; Thornhill & Van Petten, 2012). The frontal positivity obtained in the current experiment could thus be explained by the "unexpected lexical item" account, which is also in line with the above conclusion that the learners used semantic/lexical information to resolve the filler-gap dependencies. Having demonstrated that the L2 parsing of FG dependencies is indeed non-native like, we now turn our attention to morpho-syntactic processing and the L1-L2 difference in that area.

## L2 Processing of Tense Morphology

### L2 processing of tense morphology and Experiment II rationale

How native-like L2 processing is has also been considered and examined for morpho-syntactic parsing. For tense morphology, it has been proposed that while native speakers decompose regular forms such as *walked* into stem + *-ed* affix by rule application[8], L2 speakers rely exclusively on semantics-based whole-form memorization (i.e., they memorize *walked* as one piece), due to their incomplete access to morphological structure and their inability to efficiently use morpho-syntactic rules online (e.g., Neubauer & Clahsen, 2009; Silva & Clahsen, 2008; Clahsen & Neubauer, 2010). However, this claim was contradicted by results from several studies that found evidence for L2 decomposition (See Gor, 2010 for a review, e.g., Portin, Lehtonen, & Laine, 2007; Gor and Cook, 2010; Pliatsikas & Marinis, 2011). In addition to the inconclusive findings, most of the studies to date presented the inflected forms in isolation, which is not the natural way in which they are processed (Paradise, 2004). Thus, the current study replicates the ERP experiment conducted by Hestvik et al. (2009), which adopts an experimental paradigm in which the inflected forms are presented in naturally occurring contexts.

Before the design and predictions of Experiment II can be discussed, it is important to review the established correlations between ERP components and tense morphology parsing. LAN effects have been obtained in the *contextual violations paradigm*, in which subjects are presented with stimuli containing correct tensed verb forms, but in an inappropriate context, as in *Yesterday I *walk/walked to school* (Newman, Ullman, Pancheva, Waligura & Neville, 2007, Stenhauer & Ullman, 2002a). The N400 component, which typically indicates semantic

[8] There is a debate over how verb type interacts with decomposition in the research on L1 tense morphology processing. One approach proposes that all inflected forms, including irregular forms in English, are decomposed (e.g., Taft & Forster, 1975; Halle & Marantz, 1994). Experimental evidence in support of this account was found by a wide range of studies (e.g., Kielar & Joanisse, 2010; Solomyak & Marantz, 2010). The current paper assumes this full decomposition account.

anomalies, has been found to reflect declarative memory processes such as lexical access (Ullman, 2001b, Weyerts et al., 1997). The P600 was reported in relation to morpho-syntactic violations (Dowens et al., 2011; Carreiras, Pattamadilok, Meseguer, Barber & Devlin, 2012) and was observed in the contextual violation paradigm (e.g., Newman et al., 2007; Stenhauer & Ullman, 2002a).

In Hestvik et al. (2009), the subjects listened to sentences such as *Yesterday I *walk /walked to school* and *Yesterday I *eat/ate a banana.*[9] For **Yesterday I walk to school,* the listener expects a past tense marker upon encountering the verb. If decomposition occurs in regular inflected verbs, the missing *–ed* would violate the affixation rule and a LAN-type effect would result. That is precisely what Hestvik et al. observed for the native speakers. Furthermore, Hestvik et al. found the same LAN effect for the irregular verbs, suggesting that decomposition occurs for both verb types. Given this paradigm, if L2 learners decompose much less or not at all and rely mostly on whole-word memorization, they should NOT produce any LAN effects for regular verb violations. Instead, they should produce an N400, indicative of violations of lexical access, if the regular verb past tense forms were indeed memorized. For irregular verbs as in *Yesterday I *eat/ate a banana*, it is expected that the L2 speakers will generate an N400 as well, since storage is the primary processing mechanism for both regular and irregular forms. If, however, the LAN is obtained for the regular verbs but the N400 is found for the irregular verbs, then the L2 speakers decompose only the regular verbs and memorize the irregular forms. The third possibility is that the LAN is observed for both forms, in which case we argue that the L2 speakers decompose both forms and are native-like in tense processing.

---

[9] The testing material in Hestvik et al. (2009) is identical to that used by Newman et al. (2007), but was presented in auditory form.

## Method

**Participants.** Thirty-two Chinese speakers of English from the University of Delaware participated in this experiment. One subject decided to not continue before the ERP recording started. The data of two other participants were excluded because of experiment errors. The average age of the remaining participants (9 male and 20 female) is 23 years (SD=2.56, Range=20-30). Fourteen of these subjects also participated in Experiment I. All of the subjects are native speakers of Mandarin Chinese and acquired English mostly in a classroom setting. The average age of first exposure to English is 10.5 years (SD= 2.7, Range=6-15). None of them lived in an English-speaking country prior to the age of 16. By the time of the experiment, they had lived full-time in the U.S. and/or other English-speaking countries for an average of 34 months (SD=16, Range=8-79 months). All subjects reported having normal hearing and normal to corrected vision. None of them have any neurological impairment, and all except one are right-handed. They were paid $20 for their participation and gave informed consent before the experiment.

Each L2 speaker's proficiency level was determined by administering the Versant English test used for Experiment I. All scored above 49 out of 80, with an average score of 62.6 (SD=7.4, Range=47-77). The L2 participants are proficient speakers of English and are on average at the level of Advanced-low. Two groups of proficiency were constructed based on the mean score 62, resulting in a High proficiency group with 16 subjects and a Low Proficiency group with 13 subjects. Table 3 below shows the proficiency alignment:[10]

[10] Working memory is not measured in this experiment because WM effects are usually found for target structures with great complexity such as long-distance dependencies and Garden Path sentences (Juffs & Harrington, 2012).

Table 3*:*

Versant English scores and L2 proficiency levels by the ACTFL and CEFR standards

| Versant Scores | CEFR | ACTFL | Number of L2 participants |
|---|---|---|---|
| 69-78 | C1 | Advanced-mid | 6 |
| 58-68 | B2 | Advanced-low | 15 |
| 47-57 | B1 | Intermediate-high | 8 |

**Materials.** The stimuli for this experiment consisted of 320 simple declarative English sentences, in which 56 regular verbs and 56 irregular verbs were used. These are the same stimuli used in Hestvik et al. (2009). The verb type (irregular vs. regular) and tense (past vs. present) creates a 2 x 2 design, illustrate in Table 4:

Table 4:

Tense x VerbType Design with Yesterday

| Tense | **Verb type** | |
|---|---|---|
| | Irregular | Regular |
| Past (grammatical) | Yesterday, I ate a banana. | Yesterday, I walked to school. |
| Present (ungrammatical) | * Yesterday, I eat a banana. | *Yesterday, I walk to school. |

To control for the possibility that the difference in ERP signatures is due to the difference in tense (present versus past), instead of ungrammaticality, a third factor, context (Null Context versus Yesterday) was added. The ERP signatures to *I walk to school/I eat a banana* will be compared to, for example, those of *I walked to school/I ate a banana*. This allows us to confirm that the ERP in the ungrammatical cases is truly elicited by the tense expectation violations and not by the difference between tenses.

***Procedure***. After the consent and the net application, participants were seated in a chair in a sound-attenuating booth. The participants were instructed to listen to each stimuli sentence and determine whether it described an event in the past, in the present, or didn't make sense (ungrammatical condition), and to press the corresponding buttons on a response box. The 320 sentences were distributed over 4 lists, and a given critical verb appeared only once in each list. In any given list, the verb appeared in one of the 4 possible combinations of tense and context. The stimuli were pseudo-randomized so that there were no more than 2 consecutive ungrammatical sentences in any given list. Verb type and grammaticality were counter-balanced across lists, and all subjects heard the sentences in the same order. The participants first practiced on a set of 6 trial sentences. They had to reach 75% accuracy with their judgments before they were allowed to start the experiment session. The 320 experimental sentences were presented successively, in 4 blocks of 20 trials. Between the blocks there was a brief pause. The subjects were offered a break after 2 blocks. The entire experiment took around 40-45 minutes to complete.

***Data collection, EEG recording, and data analyses.*** The accuracy and reaction time (RT) of the behavioral responses were collected via E-Prime (Schneider et al., 2002), which was also used to present the stimuli. Both the accuracy and RT data were submitted to a mixed-factorial repeated measures ANOVA, with verb type (2), tense (2), and context (2) as within-subject measures. For the voltage data, EEGs were recorded in the same way as in Experiment I. The continuous EEG was divided into epochs of 1400 ms for each trial by time-locking to the *onset* of the critical verb. Baseline correction was performed by using a 200 ms baseline period (before the onset of the verb) as a reference signal value. Following the baseline correction, epochs with artifacts were rejected, which resulted in the average removal of 21.3%

of the trials per subject (SD=19.7%, Range=0.6%-45.9%). Subjects with more than 50% bad trials were not included in the ERP analysis. However, trials that were not responded to correctly were kept, in order to prevent excess loss of trials and power. The ERP average was computed for each condition, each subject, and each electrode site. The entire set of electrodes was divided into regions, such that an average over those regions could be computed and used as one of the dependent measures in the ANOVA. A total of eight electrode sites were used for Experiment II based on the factors of ANTERIORITY (anterior vs. posterior electrodes), LATERALITY (left vs. right hemisphere, excluding the midline electrodes), and DORSALITY (inferior vs. superior electrodes). For instance, the left anterior inferior region contains electrodes 18, 21, 22, 23, 25, 26, 27, 32, 33, 34, 38, 39, 40, 43, 44, 45, and 48 (red-colored), electrodes 127 and 128 monitored eye activity, and the left anterior superior region contains 7, 12, 13, 19, 20, 24, 28, 29, 30, 35, and 36. Collapsing these two creates the typical LAN area.

A total of 7 200 ms time bins were constructed for the 1400 ms epoch starting from the onset of the critical verb. The mean amplitude over these time bins was computed for each electrode region for each subject and condition. The average amplitude obtained was submitted to a mixed- factorial repeated measures ANOVA with time x region x condition (Tense, Verb type) as within-subject factors. When appropriate, *p*-values were adjusted by using Greenhouse-Geisser (1959) correction for violation of the assumption of sphericity. In addition, significant interactions between conditions, time, and electrode region were followed up by either pairwise t tests or planned orthogonal contrast analysis. Lastly, two separate sets of analyses were carried out for the ERP data. The conditions with context (i.e., with the adverb *yesterday*) were analyzed separately from the conditions with a null context (i.e., without *yesterday*).

**Experiment II Results**

**Behavioral results.** The Chinese group was highly accurate with their responses to the comprehension questions and yielded an average of 87.9% correct (SD=0.04). For the response time (RT), the L2 subjects took an average of 722.63 ms (SD=213 ms) to give their judgments. Notably, the performance of the Chinese participants was negatively affected by ungrammaticality, as expected, but especially so with the regular verbs. For accuracy, the Chinese group made significantly more mistakes in the ungrammatical condition (84% vs. 90%; $F(1, 28)=38.95$, $p<0.01$), and this difference is more pronounced for the regular verbs (86% accurate) in comparison to irregular verbs (94% accurate). This is confirmed by an interaction effect (grammaticality x verb type; $F(1, 28)=19.59$, $p<0.01$). The L2 speakers were also reliably slower in responding *only* for the regular verbs in the ungrammatical condition (761 ms vs. 684 ms; $F(1, 28)=15.33$, $p<0.01$).

**ERP results.** Before the analyses are reviewed, it is important for critical time windows to be established. Recall that the entire length of the EEG recording is 1400 ms, divided into seven 200 ms time bins. The ERP is time-locked to the onset of the verbs, and the average verb duration is 548 ms (SD=89 ms) for the regular verbs, and 513 ms (SD=95 ms) for the irregular verbs. This means that any LAN effect, if found, would not show up until 813 ms (513 ms+300 ms) for the irregular verbs, and until 848 ms (548 ms+300 ms) for the regular verbs. The relevant time windows are therefore 800-1000 ms, 1000-1200 ms, and 1200-1400 ms. For the grammaticality effect, visual inspection reveals a clear negativity in the brain responses to the tense violation with respect to the grammatical controls in the left anterior region, as shown in Figures 7 and 8:

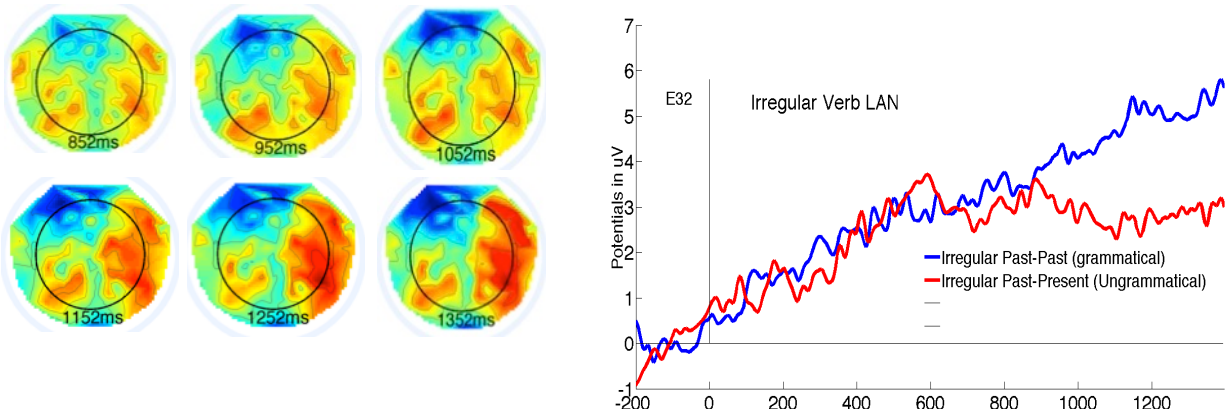



Figure 7. The LAN effect for irregular verbs: Difference wave topo plot on the left and electrode E32 (a representative electrode in the typical LAN region) on the right

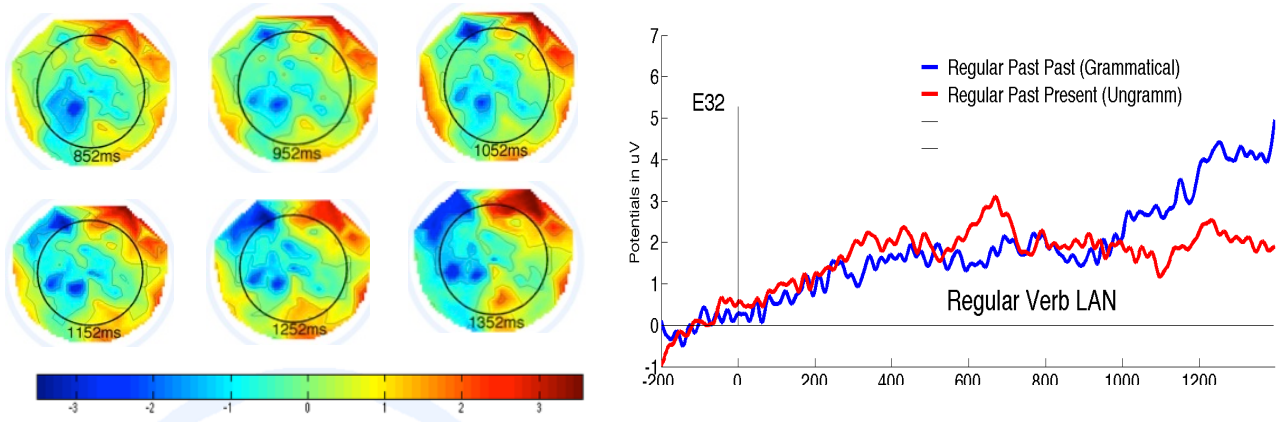



Figure 8. The LAN effect for regular verbs: Difference wave topo plot on the left and electrode E32 (a representative electrode in the typical LAN region) on the right

When both verb types are combined, the negativity starts at around the 850-1000 ms time window (300-350 ms after the violation), and is consistent with a typical LAN effect. In the right anterior region, positivity for the present tense (ungrammatical condition) over the past tense (grammatical condition) is observed. Such a reversed pattern compared to the brainwaves on the left anterior side is commonly found with the LAN and was obtained for the native English group as well in Hestvik et al.'s study (). The brain responses of the Chinese speakers didn't

show any pattern similar to an N400 or P600. The mean voltage was computed by region, time, and tense (grammaticality) and then was submitted to a mixed-factorial ANOVA (time (3) x region (4) x grammaticality (2)). There is a significant tense (grammaticality) x region interaction ($F(3,78)=4.89$, $p<0.01$), indicating that the tense effect differs by region, and a three-way interaction time x region x grammaticality ($F(6,156)=3.38$, $P<0.01$), reflecting that the grammaticality effect becomes larger over the course of the 800 ms to 1200 ms window. Given the *a priori* hypothesis, orthogonal contrast analysis revealed that the LAN effect was statistically significant in the 1000-1200 ms time window (400 ms after the violation) ($t=6.09$, $p=0,02$) and after, but not in the initial 800-1000 ms window. Thus, the onset of the Chinese LAN is considered to be relatively late (400-500 ms after the detection of the violation). To examine the precise locus of the negativity, the brain responses obtained in the inferior region were compared to those in the superior (higher, closer to the crown) region of the left hemisphere. An ANOVA (sub-region x time x tense) yielded no significant interaction between sub-regions and tense ($F(1,26)=0.066$, $p=0.8$), suggesting that the LAN was not focused in a particular sub region of the typical LAN area.

The LAN effect is bigger in terms of the voltage difference between the irregular verbs and the regular verbs. A significant LAN effect was obtained from 1000 ms and on (1000 ms: $t=-2.43$, $p=0.02$; 1200 ms: $t= -2.78$, $p=0.01$) for the irregular verbs. For the regular verbs, however, although the shape and time course of the negativity waves for regular verbs look exactly like a typical LAN, it didn't reach statistical significance for all time windows according to pairwise t tests. In the last time window, the pairwise t test shows near-significant values: $t=1.46$, $p=0.07$. Interestingly, an ANOVA didn't yield a statistically significant interaction between verb type and tense ($F(1,26)=3.38$, $p=0.07$) for the Chinese group. When proficiency is considered, visual

inspection suggests that the LAN (grammaticality effect) starts about 400 ms later for the Low proficiency group. Further examination by verb types clearly revealed that this is mostly caused by the low proficiency group's responses for regular verbs, in which the ungrammaticality induced positivity instead of negativity, suggesting they could have failed to detect the violation for regular verbs. This pattern seems to explain the insignificant regular verb LAN reported above. However, an ANOVA found no significant between-proficiency group effect ($F(1, 25)=0.07$, $p>0.05$), nor a proficiency group vs. grammaticality interaction ($F(1, 26)=1.14$, $p>0.05$). A simple regression analysis of Versant score on the LAN size for all time windows combined also yielded no significant correlation between the two factors ($R^2=0.132$, $F=2.122$, $p>0.1$), both when verb types are combined and separated.

Lastly, the null-context conditions (i.e., *I ate a banana/walked to school vs. I eat a banana/walk to school*) were inspected separately to ensure the negativity patterns obtained for the context condition are due to grammaticality alone. Neither a visual nor a statistically significant difference was found between the two no context waveforms from 600 ms and on, confirming that the previously reported LAN/LAN-type effect was indeed due to ungrammaticality.

**Experiment II Discussion**

As discussed above, the L2 speakers were highly accurate with their behavioral responses at an average of 87.9% accuracy, suggesting they have the grammatical knowledge of tense morphology. As expected, the Chinese group took longer (average 722.63 ms) than the English group (average 566.4 ms) to give judgments. However, the response time (RT) was affected in very similar ways for both groups by context, verb type and tense, such that the null context conditions took longer than the context conditions, the present tense conditions took longer than

the past tense condition, and the irregular verbs took longer than the regular verbs (Hestvik et al., 2009). The Chinese group thus is highly comparable to the native group in behavioral responses. The notable difference between the groups seems to be that the L2 speakers were negatively affected by the ungrammatical condition in both RT and accuracy, especially for regular verbs.

The Chinese ERP results also largely converge with those of English speakers as reported in Hestvik et al. (2009): first, neither groups produced the N400 or the P600. In particular, the Chinese group generated the LAN or a LAN-type effect indicative of morpho-syntactic rule violation for both verb forms, as did the native speakers. Secondly, the two groups' LAN patterns are very similar in terms of changes over the time course and the direction of the corresponding waveforms, such that the negativity obtained in the left anterior region for both verb types increased over the critical time windows (200-600 ms after detection of the tense violation). In addition, the LAN effect is bigger for the irregular verbs than the regular verbs for both groups.

The main differences between the L2 and L1 ERP responses mostly lie in the LAN onset and amplitude. When both verb types are considered in combination, the L2 LAN became significant in the 1000-1200 ms window (400-500 ms after the violation), 200-300 ms later than the native speakers as reported in Hestvik et al. (2009). The amplitude of the LAN for the native speakers is also bigger: for irregular verbs, the maximum LAN effect (in voltage difference) is around 7 uV for the L1 speakers and 3.2 uV for the L2 speakers. For the regular forms, the maximum voltage is around 4 uV as compared to the L2 2.6 uV. This suggests that the native speakers produced an earlier and stronger LAN, probably due to a greater sensitivity to the violation. Additionally, the regular verb negativity for the Chinese group didn't reach statistical

significance, although its location, time course, and direction match those of a typical LAN , and it approaches significance in the last time window (600 ms after violation).

There are a number of possible explanations for the small amplitude: First, the regular forms are less "learnt" than the irregular forms. The irregular verbs, though proven to be fully decomposed by the results discussed here, still have a stronger memory component than the regular verbs in that the brain has learned to match the different affixes to the correct stems in a more individual fashion (Halle & Maranz, 1994). Therefore, irregular verbs are inherently more reinforced than regular verbs. It then follows that the violations can be expected to generate a larger effect. Secondly, the regular verb tense marker (and hence the violation of it) is harder to detect acoustically. Given the increased cognitive load involved in processing an L2 (e.g., McDonald, 2006), it is very likely that the lack of acoustic saliency and reinforcement make the processing of regular verb violations extra challenging for the late learners, which resulted in a smaller violation effect. The behavioral data goes along with the above analysis in that the L2 speakers performed significantly worse in the violation condition in both accuracy and response time, but more so for the regular verbs than the irregular verbs. Lastly, the Chinese LAN is not concentrated in a particular sub area within the typical LAN quadrant, while it has been argued that the LAN generated by the native speakers (e.g., Hahne et al., 2006) could be more focused in the inferior region. The above L1-L2 effect locus, onset, and amplitude differences, however, are not qualitative in nature. In general, the Chinese group produced comparable behavioral and ERP results to those of the English group. In particular, the Chinese LAN and LAN-type effect obtained for the tense violations, in combination with the absence of the N400 component, suggest that L2 learners can effectively use decomposition in processing morphologically complex words. We conclude that L2 tense morphology processing is largely native-like.

## General Discussion

To test whether there exists a fundamental L1-L2 processing difference and to examine how L2 speakers use different information sources in online parsing, this paper assesses late L2 learners' ability to decompose morphologically complex forms and to build traces when computing filler-gap dependencies in real time sentence processing. In addition, L2 parameters proposed to affect L2 processing, such as proficiency and WM capacity, were also carefully measured and examined together with the L2 brain responses. We see from Experiment I that while the proficiency-controlled L2 data are native-like for off-line judgments and online behavioral tasks, the L1-L2 brain responses differ categorically in that the L2 speakers failed to build an abstract trace and resorted to semantics to resolve the FG dependency. Furthermore, such an L2 processing pattern does not correlate with individual factors such as the proficiency level and WM capacity of the subject. These patterns fully support the claim that L2 online computation is semantic-driven and lacks structural details (e.g., Clahsen & Felser, 2006a, b). In contrast, the L2 speakers produced comparable ERP signatures to the native speakers in processing tense morphology, suggesting sensitivity to morphological structure and effective online use of morpho-syntactic rules.

The findings of the two experiments in combination are problematic for accounts maintaining that the L1-L2 processing difference is quantitative rather than qualitative in nature. These accounts typically attribute the L1-L2 difference to L2-specific factors other than age of acquisition, such as proficiency, L1 transfer, and resource constraints. We saw that proficiency cannot explain the discrepancies between the current experiments since the subjects in these two studies were matched in proficiency level. Moreover, although working memory was not specifically measured in Experiment II, the chances of the participants in Experiment II having a

greater WM capacity than those in Experiment I are very small, given the large sample sizes of both studies. Crucially, in the FG dependency experiment, we saw that the qualitatively different L2 brain responses were not affected by proficiency or working memory capacity.

Another factor that could also be argued to cause the non-native brain responses of the learners when resolving FG dependencies is L1 interference. Chinese is a *wh*-in-situ language (cf. Li and Thompson, 1981), and whether movement is involved in the derivation of Chinese *wh*-questions is still being debated, with most researchers agreeing that at least adjunct *wh*-elements have to undergo movement at LF (cf. Huang, Li, & Li, 2009). In addition, Chinese has overt filler-gap dependency structures with a dislocated item, and the relative clause (RC) seems to be a filler-gap structure in Chinese. A movement analysis has been proposed for both Chinese RCs (e.g., Shyu, 1996; Hsu, 2008), and existing experimental findings, though limited, indicate that the processing of Chinese RCs and other overt FG dependencies (e.g., topic structures) are similar to those in other languages, including English (e.g., Lin & Garnsey, 2011; Packard, Ye and Zhou, 2011). In addition, while it is true that Chinese and English relative clauses are not the same, there are significant L1-L2 differences regarding morphological features, as tested in Experiment II, because Chinese is known to have very limited overt morphology (e.g., Jiang & Zhou, 2009). Nevertheless, the Chinese subjects managed to overcome the potential negative L1 influence to achieve native-like processing. L1 transfer thus cannot be the decisive factor in determining the nature of L2 processing.[11]

[11] Another potential factor is the quality of L2 lexical access, which has been proposed to be less automatic/complete than that of native speakers. It was found that L2 learners tend to have difficulty processing at the sentence level due to the unfamiliar vocabulary (e.g., Koda, 2005). The vocabulary used in Experiment I included only high frequency words, and the same materials have been used to test children (Epstein & Hestvik, to appear) who produced the bilateral AN. While it is nevertheless possible that some animal names might be unfamiliar to the adult L2 learners, the vocabulary used in Experiment II is not any less difficult/infrequent, especially the past participle forms (e.g., spin, spun). Therefore, under-routinized lexical access cannot be the reason for the qualitatively different ERP patterns between the two experiments.

While our findings argue against the view that L1 and L2 parsing share the same system, they are not in line with accounts like the Shallow Structure Hypothesis, which maintains that L2 parsing is structurally shallow in a permanent and global fashion. The results of Experiment II clearly indicate that L2 learners can effectively use morpho-syntactic rules in real time, and that non-native processing does not apply to all components of the grammar. Existing L2 morpho-syntactic data also confirm this point, as native-like processing patterns have been obtained for L2 learners by a large number of L2 ERP sentence processing studies, particularly those dealing with grammatical agreement (see Tolentino & Tokowicz, 2011 for a review). Outside of morpho-syntax, there is also evidence suggesting native-like L2 syntactic parsing, although on a relatively restricted range of syntactic constructions such as phrase structure and verb subcategorization. While native speakers produce the ELAN/LAN +P600, advanced L2 speakers produce a P600 and occasionally the ELAN component (e.g., Hahn, 2001; Isel, 2007; Rossi et al., 2006; Pakulak & Neville, 2011; Kotz et al., 2009).

From a theoretical point of view, given the fact that the parser *always* mediates the grammar (i.e., all input used to develop the grammar is filtered through the parser) and the grammar in return informs the parser (e.g., Gregg, 2003), it is unclear how a fundamentally non-native parser (i.e., a globally shallow processor that remains so throughout all the acquisition stages) can be used to build the target grammar with all its richness. Equally unexplainable is why and how the complete grammar fails persistently to develop a fully functional parser. Thus, assuming that the parser and grammar are one system (Lewis and Phillips, 2013), it would be reasonable to propose that (1) L2 parsing strategies are structure dependent, and (2) L2 non-native processing, at least at the end stage of acquisition, is limited to only a few situations, perhaps where inadequate structure building does not compromise meaning computation.

Consider the cognitive processing of tense morphology: it requires a relatively transparent feature matching operation similar to the Agree operation used for grammatical agreement. To be more precise, assuming that decomposition always takes place, this operation is triggered by some surface cues (e.g., morphological marking on the subject in subject-verb agreement, a time adverb for tense information as in Experiment II) that syntactically link the different sentence constituents (e.g., Molinaro et al. 2013). Since those cues are overt and meaning-bearing (i.e., *-ed* can be related to events/actions that happened in the past), the learners could "notice" these cues or, in the case of communication break-down, the absence of them. Such a mechanism is one of the critical components of acquisition (e.g., Gregg, 2003; Bley-Vroman, 1989). As the learners relate these cues to the corresponding morpho-syntactic rule (such as memorizing a list of lexical variations attached to certain structures), they are able to use those rules more efficiently. In contrast, structures like filler-gap dependencies require the use of an abstract trace for the complete computation of the syntactic representation. These operations and elements are (1) completely devoid of semantic meaning, and (2) opaque at the surface level (i.e., they lack a phonological reflex). Crucially, the meaning computation is still accurate without the trace[12], and the learners cannot "notice" the use of the rule and relate it to any sort of meaning-form mismatch. In fact, the learners might be motivated by efficiency to use a meaning-based routine. Thus, the processing of such constructions could remain permanently shallow.

[12] For example, consider the sentence *[The purse [$_{cp}$ Op$_i$ that [IP Allison misplaced t$_i$]]] is very expensive.* There is a person whose name is Allison. She misplaced a purse and the purse is expensive. While the native speakers build a hierarchical structure with a null operator and the trace in the gap position, the L2 speakers, with a declined ability to induce structure due to the age constraint, compute the meaning by segmenting the incoming information into chunks via thematic role assignment and by associating modifiers to semantically appropriate head phrases. These chunks are integrated into the existing semantic representation in an incremental fashion (Clahsen & Felser, 2006). Therefore, as soon as *misplaced* is processed, the Agent (*Allison*) and the Theme (*the purse*) are identified. The L2 speakers then work out that Allison misplaced a purse, and later that the purse is expensive (not *Allison* because that would not make sense). Thus, the meaning of the sentence is successfully computed, albeit via a non-native mechanism.

Under this view, it is explainable why proficiency sometimes predicts a more native-like parsing profile for morpho-syntax processing (as also shown in McLaughlin et al. (2010), Steinhauer, White & Drury (2009)) but at other times it does not affect L2 processing, as in the case of FG dependencies in Experiment I.

Following this line of thinking, some predictions for L2 processing of various structures can be made. For example, constructions involving VP ellipsis as in *John defended himself and Bill did too*, could trigger non-native shallow processing among L2 speakers as well. This is because the elided VP is syntactically represented throughout the stages of the derivation, but has no phonological value (e.g., Sag, 1976). The L2 speakers, not capable of building full abstract syntactic structures, would process the covert elements via semantic reconstruction (such as computing a pro form) instead of building structure inside the elided VP. Nevertheless, the L2 speakers still effectively compute the meaning of the sentence. Contrary to the trace positing in FG dependencies and VP ellipsis, verb subcategory information (like agreement and tense decomposition) could be processed in a native-like fashion by L2 learners, due to the overt cues evident in the surface form (e.g., the presence of prepositions, for instance).

In conclusion, this paper provides novel neurophysiological evidence demonstrating that while L2 learners are able to effectively use syntactic information in a native-like fashion for some structures, they resort to a fundamentally non-native parsing strategy guided mostly by semantics for other structures. We propose that 1) the distinct linguistic features of the structures and 2) whether shallow processing has any consequence for the accuracy of the semantic computation determine the L2 parsing strategy. Assuming the view that grammatical theories and language processing models describe a single cognitive system (Lewis and Phillips, 2013), it

is stressed here that L2 non-native/shallow processing is of limited scope and persists only when it does not interfere with successful meaning computation.

There are, of course, limitations to the current studies. To validate the point that the L2 parsing strategy is construction-specific, more structures, such as other types of FG dependencies and referential dependencies, need to be investigated in future studies. Another promising research avenue is to consider the effect of L2 instruction type. A distinction can be made regarding whether the learner received explicit, memorization-based training or implicit, immersion-style instruction. The participants in the present studies had the former kind of instruction. However, when the learners received implicit teaching delivered 100% contextualized in the target language, syntactic rules are supposed to be derived by the learners via the processing of naturalistic input that is carefully chosen to focus on the target structure, and lexical items are taught mostly via visual cues and context instead of translation into the native language. A few recent studies (e.g., Morgan-Short, Steinhauer, Sanz & Ullman 2012; Dussias, 2003; Frenck-Mestre, 2002, Pilatsikas & Marinis, 2011) have shown that L2 exposure and training types can shape L2 processing strategies and neuro-cognition. For future research, it would be beneficial to test late learners who received implicit training from very early on to explore the role of training type in L2 processing. The results will not only inform the research on L2 processing overall but also that of L2 acquisition and pedagogy.

**References**

Ardal, S., Donald, M. W., Meuter, R., Muldrew, S., & Luce, M. (1990). Brain responses to semantic incongruity in bilinguals. *Brain and language*, *39*(2), 187-205.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157-194). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baztán, A. M. (2008). La evaluación oral: una equivalencia entre las guidelines de ACTFL y algunas escalas del MCER (doctoral thesis). Universidad de Granada. p. 469–70.

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). Taipei, ROC: The 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). Florence, KY: Routledge.

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning. *Linguistic perspectives on second language acquisition*, *4*, 1-68.

Carreiras, M., Pattamadilok, C., Meseguer, E., Barber, H., & Devlin, J. T. (2012). Broca's area plays a causal role in morphosyntactic processing. *Neuropsychologia*, 50(5), 816-820.

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

Clahsen, H., & Felser, C. (2006)a. Grammatical processing in language learners. *Applied Psycholinguistics,* 27, 3–42.

Clahsen, H., & Felser, C. (2006)b. How native-like is non-native language processing? *Trends in Cognitive Sciences*, 10, 564–570.

Clahsen, H. & K. Neubauer (2010). Morphology, frequency, and the processing of derived words in native and non-native speakers. *Lingua* 120: 2627–263.

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). Dordrecht, The Netherlands: Kluwer.

Daneman,M. & P. A. Carpenter (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior 19.4, 450–466.

Dekydtspotter, L., Donaldson, B., Edmonds, A. C., Fultz, A. L., & Petrush, R. A. (2008). Syntactic and prosodic computations in the resolution of relative clause attachment ambiguity by English-French learners. *Studies in Second Language Acquisition*, *30*(04), 453-480.

Delong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual ERP responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203-1207.

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.

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.

Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners. *Studies in Second Language Acquisition*, *25*(4), 529-557.

Epstein, B., Hestvik, A. (2013). ERPs reveal atypical processing of subject vs. object Wh-questions in children with Specific Language Impairment. To appear in *International Journal of Language & Communication Disorders*.

Federmeier, K. D., & Kutas, M. (2005). Aging in context: Age ‐ related changes in context use during language comprehension. *Psychophysiology*, *42*(2), 133-141.

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

Frenck-Mestre, C. (2002). An on-line look at sentence processing in the second language. In R. R. Heredia & J. Altarriba (Eds.), *Bilingual sentence processing* (pp. 217-235). Amsterdam: Benjamins.

Friederici, A. D. (1995). The time course of syntactic activation during language processing: A model based on neuropsychological and neurophysiological data. *Brain and language*, *50*(3), 259-281.

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, 1219–1248.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6(2), 78-84.

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.

Gibson, E., & Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, 7, 55-78.

Gor, K. (2010). Introduction. Beyond the obvious: Do second language learners process inflectional morphology?. *Language Learning*, 60(1), 1-20.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24(2), 95-112.

Gregg, K. (2003). SLA theory construction and assessment. In C. Doughty & M. Long (Eds.). *Handbook of second language acquisition* (pp. 831-865). Oxford: Blackwell

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

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *Journal of Psycholinguistic Research*, *30*(3), 251-266.

Hahne, A., Mueller, J., & Clahsen, H. (2006). Morphological processing in a second language: Behavioral and event-related brain potential evidence for storage and decomposition. *Journal of Cognitive Neuroscience, 18*, 121–134.

Halle, M., & Marantz, A. (1994). Some key features of Distributed Morphology. In A. Carnie & H. Harley (Eds.), *MIT Working Papers in Linguistics 21: Papers on phonology and morphology*. Cambridge, Massachusetts: MIT Department of Linguistics and Philosophy.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*(1), 25-38

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, 187–226.

Hestvik, A., Maxfield, N., Schwartz, R. G., & Shafer, V. L. (2007). Brain responses to filled gaps. *Brain and Language,* 100(3), 301-316.

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.

Hestvik, A., Shafer, V., Schwartz, R. G., Ullman, M., Neumann, Y., & Rinker, T.(2009). Brain responses to contextually ungrammatical verb inflection. Manuscript in preparation.

Retrieved from http://udel.edu/~hestvik/publications/Hestvik-Shafer-et-al-PastTense-BrainResearch.pdf

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, Cambridge University Press.

Isel, F. (2007): Syntactic and referential processes in second-language learners: event-related brain potential evidence. *Neuroreport 18*, 1885–89.

Jegerski, J. (2010). Ultimate attainment in second language acquisition: Near-native sentence processing in Spanish. ERIC.

Jiang, X., & Zhou, Z. (2009). Processing different levels of syntactic hierarchy: An ERP study on Chinese. *Neuropsychologia*, 47(5), 1282–1293.

Juffs, A. (2006). Grammar and parsing and a transition theory. *Applied Psycholinguistics, 27,* 66-69.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, 44(02), 137-166

Kaan, E., Harris, A., Gibson, E., & Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and cognitive processes,* 15(2), 159-201.

Kielar, A., & Joanisse, M. F. (2010). Graded effects of regularity in language revealed by N400 indices of morphological priming. *Journal of cognitive neuroscience*, *22*(7), 1373-1398.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.

Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language, 109*(2), 68-74.

Kutas, M., Federmeier, K.D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences* 4, 463–470.

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.

Larson-Hall, J. (2006). What does more time buy you? Another look at the effects of long-term residence on production accuracy of English /r/and/l/ by Japanese speakers. *Language and Speech*, 521–548.

Lau, E., Stroud, C., Plesch, S., & Phillips, C. (2006). The role of structural prediction in rapid syntactic analysis. *Brain and Language, 98*, 74–88.

Lewis, S. & Phillips, C. (2013). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*.

Li, C., & Thompson, S. (1981). A functional reference grammar of Mandarin Chinese.

Lin, Y., & Garnsey, S. M. (2011). *Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin* (pp. 241-275). Springer Netherlands.

Love, T. E. (2007). The processing of non-canonically ordered constituents in long distance dependencies by pre-school children: A real-time investigation. *Journal of Psycholinguistic research*, 36, 191-206.

Marinis, T., L. Roberts, C. Felser, & H. Clahsen (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, 27, 53-78.

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(03), 379-413.

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.

McLaughlin, J., Tanner, D., Pitkänen, I., Frenck‐Mestre, C., Inoue, K., Valentine, G., & Osterhout, L. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Language Learning*, *60*(s2), 123-150.

Molinaro, N., Barraza, P., & Carreiras, M. (2013). Long-range neural synchronization supports fast and efficient reading: EEG correlates of processing expected words in sentences. *NeuroImage*.

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.

Neubauer, K., & Clahsen, H. (2009). Decomposition of inflected words in a second language: An experimental study of German participles. *Studies in Second Language Acquisition*, 31(3), 403–435.

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

Newman, A. J., Ullman, M. T., Pancheva, R., Waligura, D. L., & Neville, H. J. (2007). An ERP study of regular and irregular English past tense inflection. *NeuroImage,*34(1), 435-445.

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

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.

Packard, J. L., Ye, Z., & Zhou, X. (2011). Filler-gap processing in Mandarin relative clauses: Evidence from event-related potentials. In *Processing and producing head-final structures* (pp. 219-240). Springer Netherlands.

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.

Paradis, M. (2004). A Neurolinguistic Theory of Bilingualism. Amsterdam: Benjamins.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. *Oxford handbook of psycholinguistics*, 739-756.

Phillips, C., & Wagers, M. (2007). Relating structure and time in linguistics and psycholinguistics. *Oxford handbook of psycholinguistics*, 739-756.

Pliatsikas, C., & Marinis, T. (2011). Processing of regular and irregular past tense morphology in highly proficient L2 learners of English: A self-paced reading study. *Applied Psycholinguistics*.

Rah, A., & Adone, D. (2010). Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels. *Studies in Second Language Acquisition*, 32(1), 79-109.

Rodríguez, G. A. (2008). *Second Language Sentence Processing: Is it fundamentally different?*. ProQuest.

Rossi, S., Pasqualetti, P., Zito, G., Vecchio, F., Cappa, S. F., Miniussi, C., ... & Rossini, P. M. (2006). Prefrontal and parietal cortex in human episodic memory: an interference study by repetitive transcranial magnetic stimulation. *European Journal of Neuroscience*, *23*(3), 793-800.

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

Sag, I. A. (1976). "Deletion and Logical Form". *PhD Thesis, Massachusetts Institute of Technology* (New York: Garland Publishing).

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

Shyu, S. I., & Sybesma, R. (1996). The syntax of focus and topic in Mandarin Chinese. *Letters & Replies*, *2*(4), 11.

Silva, R., & Clahsen, H. (2008). Morphologically complex words in L1 and L2 processing: Evidence from masked priming experiments in English. *Bilingualism: Language and Cognition*, 11, 245–260.

Solomyak, O., & Marantz, A. (2010). Evidence for early morphological decomposition in visual word recognition. *Journal of Cognitive Neuroscience*, *22*(9), 2042-2057.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*, *48*(3), 609-652.

Steinhauer, K., & Ullman, M. T. (2002). Consecutive ERP effects of morpho-phonology and morpho-syntax. *Brain and Language*, 83, 62-65.

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

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, *14*(6), 638-647.

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.

Tolentino, L. C., & Tokowicz, N. (2011). Across Language, space and time. *Studies in Second Language Acquisition*, *33*(01), 91-125.

Ullman, M. T. (2001)b. A neurocognitive perspective on language: The declarative/procedural model. *Nature reviews neuroscience*, *2*(10), 717-726.

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.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language*, 50(1), 1-25.

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

Weyerts, H., Penke, M., Dohrn, U., Clahsen, H., Munte, T.F. (1997). Brain potentials indicate differences between regular and irregular German plurals. *Neuroreport* 8, 957–962.

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

Williams, J. (2006). Incremental interpretation in second language sentence processing. *Bilingualism: Language and Cognition*, 9, 71-88.