Accepted for publication in Mind, Brain and Education.

The final publication is available at https://onlinelibrary.wiley.com/doi/10.1111/mbe.12324

Second language vocabulary learning through ecologically valid classroom practice can be indexed by Event-Related Potentials (ERPs). A conceptual replication study.

Authors: Bert Vandenberghe[1, 2], Maribel Montero Perez[3], Bert Reynvoet[4,5], Piet Desmet[1,2]

[1] Itec, imec research group at KU Leuven

[2] Department of Linguistics, KU Leuven

[3] Department of Linguistics, Ghent University

[4] Department of Psychology and Educational Sciences, KU Leuven

[5] Brain and Cognition, KU Leuven

Lay abstract

We investigated whether the EEG brain responses to newly learned second language (L2) vocabulary known from laboratory-based studies, could be replicated in a realistic L2 classroom setting. While treatments in laboratory-based studies involve deliberate attention to new lexical items during the learning phase, current L2 classroom practice often focuses on meaning processing and incidental vocabulary uptake. Results show that brain responses produced in laboratory settings can also be elicited through L2 classroom activities.

Abstract

Previous research has suggested that Event-Related Potentials (ERPs) can index form-, meaning-, and use-related aspects of second language (L2) vocabulary knowledge, through, respectively, a lexical decision task (LDT, targeting N400), a semantic relatedness task (SEMREL, targeting N400), and a grammatical judgment task (GJT, targeting P600/AN). Yet, those findings were established in the context of either laboratory-based or longitudinal word learning studies. Therefore, this conceptual replication study investigates whether LDT, SEMREL and GJT could reveal the aforementioned ERP components in the context of an ecologically valid short-term intervention representative of L2 classroom practice. Twenty Dutch-speaking participants performed reading and writing tasks in a computer-assisted learning environment. Target words were 20 real L2 French verbs. ERPs were recorded immediately after the learning sessions. Results showed N400 for form- and meaning-related knowledge, and P600 for grammatical use. We conclude that ERPs can index L2 vocabulary knowledge gained through ecologically valid classroom activities.

Introduction

In order to refine our understanding of L2 lexical knowledge, there has been a growing interest in exploring the value of measurement techniques coming from the field of cognitive psychology (Godfroid, 2020; Schmitt, 2019; Vandenberghe et al., 2019; 2021). Hence, neurocognitive techniques such as recording the electrical brain activity related to cognitive processing (i.e., recording ERPs or Event-Related Potentials) may have the sensitivity to shed light on how newly learned words are

processed in the brain (McLaughlin et al., 2004; Schmitt, 2010; Vandenberghe et al., 2019). Since ERP measures have the sensitivity to tap into L2 vocabulary knowledge dimensions that are not addressed through explicit measures (e.g., paper-and-pencil tasks), they have been called sensitive measures, that is, measures that gauge "different aspects of language processing that are not available to learners' conscious control or report" (Elgort, 2018, p. 4).

This study has to be situated within the context of a broader research project that investigates the impact of meaning-focused L2 instruction (i.e., engaging in meaningful communicative activities without a focus on new L2 vocabulary) and word-focused L2 instruction (i.e., directing learners' conscious attention to new vocabulary in either communicative or non-communicative situations, Laufer, 2017), by using sensitive measurement techniques (e.g., reaction time measures in Vandenberghe et al., 2021; ERPs in the present study). Within this global project, the role of the present study consists in exploring whether ERPs can actually be used in order to pick up L2 vocabulary learning gains as a result of an ecologically valid short-term L2 classroom intervention. The outcomes of the present study will inform a subsequent study that aims at comparing the ERP signature of ecologically valid meaning- and word-focused L2 teaching approaches.

Since most previous vocabulary ERP studies were either conducted in lab contexts, or through longitudinal interventions, the first aim of this study consists in assessing whether and to what extent ERPs can pick up L2 vocabulary learning that occurs during a short-term intervention in a realistic L2 classroom setting. Since ERPs have shown to be sensitive to form-, meaning-, and use-related aspects of word knowledge (Vandenberghe et al., 2019) when elicited through, respectively, a lexical decision task (LDT), a semantic relatedness task (SEMREL), and a grammatical judgement task (GJT), the second aim of this study is to investigate whether these three experimental paradigms could replicate findings established in previous studies.

In sum, this study was designed as a conceptual replication study, i.e., a study that attempts to confirm previous findings (i.e., findings established in a laboratory context or during longitudinal studies) using a set of specific methods (i.e., a realistic short-term L2 classroom intervention) that test the same idea (Diener & Biswas-Diener, 2020).

What is involved in word knowledge ?

Assessing the impact of instruction on vocabulary acquisition requires a precise conceptualization of what it means to know a lexical item. Nation's (2013) framework is considered to be the most comprehensive approach to the concept of word knowledge (Schmitt, 2019). In this

framework (Figure 1), vocabulary knowledge encompasses form-related (e.g., written form), meaning-related (e.g., word associations), and use-related (e.g., grammatical patterns in which a word is used) aspects of word knowledge. Additionally, learning a lexical item is an incremental process of gradually acquiring small amounts of form-, meaning-, and use-related word knowledge through repeated encounters with the new word (Nation & Webb, 2011).

<< INSERT FIGURE 1 HERE >>

Experimental tasks used in previous research

Experimental paradigms commonly used in psycholinguistic research may have the potential to address form-, meaning-, and use-related aspects of L2 word knowledge, as conceived in Nation's framework (Vandenberghe et al., 2019). The lexical decision task (LDT, i.e., participants have to decide whether a presented letter string is an existing word or not) has been used to tap into aspects related to word form. In a series of studies (e.g., Laszlo & Federmeier, 2009), it has been demonstrated that all types of orthographic input (e.g., words, pseudowords, acronyms) attempt to have access to meaning. N400 (a negative-going brainwave observed between 300-500 ms at central and parietal sites, and indicative of meaning processing) indexes the difficulty of achieving semantic access. As such, pseudowords, which have no semantic representation, are assumed to show N400 (indexing the difficulty of semantic access) when compared to well-known real words (assumed to have an easily accessible and solid semantic representation) (Figure 2).

<< INSERT FIGURE 2 HERE >>

In order to gauge meaning-related L2 vocabulary knowledge, the semantic relatedness task (SEMREL) has been used, in which the ERP responses to related and unrelated word pairs are compared (*chien-chat*, 'dog-cat', and *maison-soif*, 'house-thirst' in McLaughlin et al., 2004). In SEMREL, N400 is said to index the degree of integration difficulty with the preceding context, i.e., the first word of a word pair. Consequently, unrelated word pairs (*maison-soif*, 'house-thirst') present a more negative-going N400 deflection than related word pairs (*chien-chat*, 'dog-cat'). Additionally, in comparison to its canonical latency (300-500 ms post-stimulus), N400 for newly learnt words has sometimes been elicited in later time windows (e.g., 400-600 ms in Pu et al., 2016; 500-700 ms in Ojima et al., 2005; 500-700 in Mestress-Missé et al., 2007). This delayed latency has been explained as the signature of emerging meaning-related knowledge development (Mestres-Missé et al., 2007), in that it reflects

slower processing that will speed up with increasing proficiency (Ojima et al., 2005) Likewise, while the canonical topography for N400 shows a centro-parietal distribution, newly established lexical knowledge has repeatedly been indexed by N400 with a more anterior/frontal scalp distribution (Elgort et al., 2015; Mestress-Missé et al., 2007; Ojima et al., 2011; Pu, et al., 2016; Yum et al., 2014).

In order to measure use-related grammatical knowledge, the grammatical judgement task (GJT) (i.e., participants have to decide whether a sentence is grammatically correct or not) has been used to investigate the effect of syntactic violations (Morgan-Short, 2014). In this task, P600 (a positive-going brainwave observed between 500-900 ms, typically observed at parietal sites, and indicative of grammatical processing) has shown to be sensitive to violations such as phrase structure disruptions, verb argument structure violations or word category violations (e.g., the man in the *drinks a coffee, Rossi et al., 2006). P600 can be accompanied by anterior negativities (AN), i.e. negative-going brainwaves at anterior scalp sites believed to reflect processes linked with automaticity and the involvement of working memory (Morgan-Short et al., 2015).

In sum, since experimental paradigms such as LDT, SEMREL, and GJT have frequently been used in previous ERP research to yield N400, P600 and AN, the first objective of this conceptual replication study was to examine whether these paradigms could also elicit these components in the context of ecologically valid L2 classroom activities.

Ecologically valid L2 classroom instruction

L2 pedagogy has often promoted L2 learning through performing communicative classroom activities that reflect real-life usage situations (e.g., booking a flight, selecting relevant information on news sites, expressing one's opinion). Under this view, L2 learning is 'incidental' (Webb, 2020), that is, while attention can be allocated to linguistic forms, the overriding focus is on meaning processing. Yet, L2 vocabulary research has repeatedly shown that 'intentional learning' (i.e., with an explicit attentional focus on new lexical items) is a more effective method for learning new vocabulary (Laufer, 2017; Schmitt, 2008; Webb & Nation, 2017). In most previous laboratory-based ERP studies, participants were instructed to deliberately commit to memory new vocabulary. Therefore, in those studies, word learning was 'intentional', not 'incidental'. In Perfetti et al. (2005), for instance, participants had to study 60 rare words through flashcards with definitions for 45 minutes, while in Balass et al. (2010), participants were asked to study words through a 2 h – 2 h 30 self-paced computerized learning task. In Bakker et al. (2015), targets were presented with their definitions followed by a battery of training exercises (word form and definition matching, word form recall,

definition recall). In sum, since most previous laboratory-based vocabulary studies allocated a strong attentional focus to the target words, the present conceptual replication study intends to gauge whether classroom based incidental L2 vocabulary learning would be sufficiently effective to yield ERP responses that are indicative of form-, meaning-, or use-related L2 vocabulary learning gains.

Additionally, in L2 vocabulary research, most insights come from short-term intervention studies (Schmitt, 2020). Yet, the majority of previous ERP studies conducted in L2 classroom contexts adopted a longitudinal design (12 weeks in Choi et al., 2014; 12 weeks in Chun et al., 2012; 14, 60 and 140 hours of instruction in McLauglin et al., 2004; one year in Osterhout et al., 2019; 12 weeks in Soskey et al., 2016). Since the number of encounters with new lexical items plays a key role in successful L2 vocabulary learning (Laufer, 2017; Schmitt, 2020, Webb & Nation, 2017), the effects found in longitudinal studies cannot be transferred to a short-term intervention. As such, it is possible that vocabulary knowledge learned in the L2 classroom context can be detected by ERPs after some weeks or months, but not yet in the early stages of learning (McLaughlin et al.; 2004; Ojima et al., 2011; Osterhout et al., 2019; Yum et al., 2014).

Taken together, since most previous laboratory-based ERP studies were based on intentional learning and/or on longitudinal designs, the present study intended to investigate whether the experimental tasks used in those studies (LDT, SEMREL, GJT) could replicate the indices of L2 vocabulary learning in a different research context, that is, through a short-term L2 classroom intervention study representative of incidental vocabulary learning gained through realistic L2 classroom tasks.

Rationale and research questions

This conceptual replication study has to be situated within the context of a broader research project focusing on the value of sensitive measures in L2 vocabulary research. The present study was designed in order to explore whether L2 vocabulary learning that occurs during a short-term L2 classroom intervention could be picked up by ERPs. Since Nation's vocabulary knowledge framework encompasses form-, meaning-, and use-related aspects, LDT is used to measure word form-related knowledge, SEMREL is used to measure meaning-related knowledge, and GJT is used to measure use-related knowledge. The following research questions guided this study:

RQ 1. To what extent will LDT, SEMREL and GJT elicit indices of form-, meaning-, and use-related aspects of well-known L2 French words?

RQ 2. To what extent will LDT, SEMREL and GJT elicit indices of form-, meaning-, and use-related aspects of newly learned L2 French words through ecologically valid classroom activities ?

RQ 1 was included in order to ensure the validity of our paradigm. By doing so, we followed methodological recommendations suggesting that task validity can be ensured when results based on knowledge that participants fully master (i.e., high frequency L2 French vocabulary pertaining to the 1,000 most frequent French words, based on Lonsdale and Le Bras, 2009), produce outcomes that are interpretable in light of previous literature (Meulman, et al., 2016; Woodman, 2010). Therefore, with respect to RQ1, we hypothesize that for well-known L2 French words, LDT (form) and SEMREL (meaning) may elicit a canonical N400 (i.e. 300-500 ms with a central and parietal distribution), while GJT (use) may yield P600 in the 500 -900 ms time window and AN.

With respect to RQ 2, we predict that incidental vocabulary learning through a short-term intervention may yield the identical ERP components. Based on previous literature, these components may possibly be observed in a later time window (e.g., 400-600 ms or 500-700 ms) and showing a different scalp distribution (e.g., at more anterior electrode sites).

Method

Participants

Participants were 20 young adults (15 female, age = 18-25) who volunteered to take part in the experiment. All participants were L1 speakers of Dutch with intermediate to advanced knowledge of L2 French. Since the global proficiency level at the end of secondary school in Flanders for L2 French is B1 on the Common European Framework of Reference (CEFR, Council of Europe, 2001), we can assume that all participants minimally had attained this intermediate level. All participants gave written informed consent before participation in the study. The procedures, which conform to recognized standards, were approved by the KU Leuven Social and Societal Ethics Committee (SMEC).

Targets

Targets were 20 real low-frequency L2 French verbs (Figure 3) that were selected in the context of the global research project (Vandenberghe et al., 2021). The target words were selected from a pool of 51 candidates (i.e., presumably unfamiliar to B1 level learners) encountered at French online news sites. These verbs were tested on three levels. First, a multiple-choice meaning recognition test was administered in last-year secondary school students (N = 228, global B1 CEFR level). Second, a

comparable group of learners (*N* = 239) was asked to infer the meaning of the candidate targets through newspaper contexts. Third, experienced L2 French teachers (*N* = 22) assessed whether the meaning of 120 French verbs (including the potential targets) could be known by students at the end of secondary school. Results of the recognition and inferencing test, as well as the ratings from the teachers, were combined in order to obtain the final set of twenty target verbs.

<< INSERT FIGURE 3 HERE >>

Distribution of the 20 target verbs over frequency bands, based on www.lextutor.ca. K3: 1; K4:1 K5: 1; K6: 1; K7: 1; K8: 1; K9: 2; K10: 1; K11:1; K12: 2; K14: 3; K15: 2; K21: 1; K23: 1; not in the frequency list: 1.

Learning materials

In order to mimic a realistic language use context, a life-like online news site was created including ten articles based on authentic news items (Figure 4). By clicking on each item, participants were directed to a split-screen page showing the article on the left-hand side, and the activities related to the article on the right-hand side (Figure 5). Activities focused on reading comprehension and written production. Receptive activities were multiple-choice comprehension questions that were both meaning- and word-focused. For instance, in the article entitled "Scandale: une agence de tourisme américaine *éclabousse* la réputation de Disneyland" (Scandal: a travel agency *damages* the reputation of Disneyland), it was asked which statement would best fit the title: On *sauve / dégrade / améliore* la réputation de Disneyland ('The reputation of Disneyland is *saved / harmed / improved*'). Per article, the comprehension activities were followed by a writing activity that consisted in posting a short comment of about five lines for each news item while using the target items. Participants performed the learning session at their own pace, but were aware of a 90 minutes time frame.

<< INSERT FIGURE 4 HERE >>

<< INSERT FIGURE 5 HERE >>

Stimulus materials

LDT was used to measure form-related knowledge. Stimuli consisted of the 20 target verbs (*kiffer*), 20 well-known (i.e., pertaining to 1000 most frequent L2 French words, based on Lonsdale and Le Bras, 2009) French verbs (e.g., *passer*), and 40 pseudoverbs (following French phonotactics, **fisser*). Stimuli were matched for verb type (ending in *–er*), and length (number of syllables and characters).

SEMREL was used to measure meaning-related knowledge. Stimuli were word pairs consisting of either two well-known L2 words or a well-known L2 word and a target word. Stimuli were 80 sets of prime-target pairs containing 20 related and 20 unrelated pairs with known words or targets preceded by either a related or unrelated prime (*cuisiner–manger, *cacher–manger; in English: to cook–to eat, *to hide, to eat*)

GJT was used to measure use-related word knowledge. In the context of this study, use-related knowledge was operationalized as the lexicogrammatical properties of the target verbs. In *Le jeune kiffe ce style de musique électronique* ('The youngster likes this style of electronic music'), the verb *kiffe* ('likes') followed by a noun phrase is a grammatical continuation of the phrase. However, in *Le jeune kiffe *dit style de musique électronique* ('The youngster likes *says style of electronic music'), the verb **dit* ('says') following the verb *kiffe* ('likes') is an ungrammatical continuation of the phrase. Although phrase structure violations may be salient and overtly unacceptable to intermediate and advanced L2 French learners, it has to be noted that ERPs are measured upon presentation of the violation stimulus while performing the experimental task (i.e., **dit* in the above-cited example). Conversely, behavioral judgments are made after presentation of the last word of the sentence, which means that, contrary to ERP responses, behavioral responses are made on the basis of the integration of all the information contained in the sentence. As such, behavioral judgments are made after completion of processes related to resolving semantic and syntactic inconsistencies, or completing syntactic analyses (Stowe, et al., 2018).

Stimuli were 10 correct and 10 syntactically violated sentences containing a target verb (**Les manifestants agressifs narguent venir policiers de la capitale, '**The aggressive demonstrators mock come the capital's police force'), and 10 correct and 10 syntactically violated sentences containing a known verb (**Notre équipe gagne jouer match the football important*, '*Our team wins play the important football game*')*.

Procedures

EEG recording (90 minutes) took place immediately after the treatment. Participants were comfortably seated in a silent, dim light room and took the three experimental tasks. Based on Nation and Webb's (2011) recommendations, the three tasks were sequenced in order to avoid testing effects

The form-related LDT was administered first, since seeing the L2 word form is not informative about the meaning, nor about the lexicogrammatical properties of the target verb. In order to ensure that sentence contexts would not inform meaning-related knowledge, GJT (third task) was administered before SEMREL (second task).

Stimuli were displayed on a monitor (60 Hz refresh rate) located at 60-70 cm of the participants. In all tasks, in order to reduce alpha-wave activity and timing-dependent oscillations, stimulus onset asynchronies and inter-stimulus intervals varied randomly.

For LDT, participants were asked to decide whether stimuli (target, well-known verb, pseudoverb) were existing L2 French verbs. The trial sequence was as follows: fixation cross (1000 ms), blank screen (1200-1500 ms), stimulus (2000 ms), blank screen (1200-1500 ms), yes/no prompt "Is this an existing French word?" (until response). Stimuli were presented in random order.

For SEMREL, participants had to decide whether two words were related in meaning. The trial sequence was as follows: fixation cross (1000 ms), blank screen (1200-1500 ms), prime (1000 ms), blank screen (1200-1500 ms), target (2000 ms), blank screen (1200-1500 ms), yes/no prompt "Related meaning?" (until response). Stimuli were presented in pseudo-random order to ensure an interval between related and unrelated pairs containing the target verb.

For GJT, participants had to decide whether sentences were grammatically correct or not. Stimuli words were presented one at a time. ERPs were recorded at presentation of the violation stimulus. Trials were sequenced as follows: fixation (1000 ms), blank screen (1000-1400 ms), word (500 ms), blank screen (1000-1400 ms), etc., Yes/no prompt "Correct?" (until response). Sentences were presented in random order.

EEG recording and pre-processing

EEG was recorded from 64 Ag-AgCL-tipped electrodes (10-20 system). Six additional electrodes (both mastoids, outer canthi of both eyes, above and below the right eye) were used to monitor ocular artifacts,. Electrodes were adjusted to a stable offset lower than 20 mV. Recordings (2048 Hz) were made with Biosemi's Active-Two System. EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) were used for pre-processing. Data were down-sampled to 256 Hz and band-pass filtered (lower bound 0.1 Hz, upper bound 30 Hz). Bad channels and EEG episodes contaminated by paroxysmal muscular artifacts were removed before Independent Components Analyses (ICA). After pruning the continuous EEG, removed channels were interpolated. Data were re-referenced to the average of the scalp electrodes. Epochs were extracted from the continuous EEG as follows: -200 ms to 800 ms for LDT, -200 ms to 1000 ms for SEMREL, -200 ms to 1400 ms for GJT.

Epochs containing values exceeding -100 μV and +100 μV were removed. ERPs were quantified by computing mean amplitudes over participants per stimulus type, relative to a 200 ms pre-stimulus baseline.

Approach to data analysis

Following Luck's (2014) suggestion for electrode clustering, we adopted a 3 X *N* distribution, i.e., 3 vertically-oriented electrode clusters (i.e., left/midline/right) X *N* horizontally-oriented anterior-to-posterior clusters. In the context of language-related ERP research that targets N400, P600, and AN, three (i.e., $N = 3$) anterior-to-posterior clusters are relevant. With respect to the *anterior* cluster, an anteriorly-distributed N400 has been detected for new vocabulary, both in L1 and L2 (e.g., Elgort et al., 2015; Pu et al., 2016). Furthermore, P600 can be preceded or accompanied by anterior negativities (e.g., Morgan-Short et al., 2012). With respect to the *central* cluster, the canonical N400 effect is usually found over central and parietal sites. Finally, for the *posterior* cluster, P600 has repeatedly been reported over posterior sites. In order to maintain consistency across the three experimental tasks, we will use this clustering for analyzing the three experimental tasks. A similar 3 X 3 distribution has been adopted in other L2 experiments that targeted both N400 and P600 (e.g., Newman, et al., 2012).

Voltages were averaged across neighboring electrodes and combined following nine regions of interest (e.g., Newman et al., 2012), i.e., left anterior: AF3, F1, F3, F5; left central; FC1, FC3, FC5, C1, C3, C5; left posterior: CP1, CP3, CP5, P1, P3, P5, P03; midline anterior: AFz, Fz; midline central: FCz, Cz, midline posterior: CPz, Pz, POz, right anterior: AF4, F2, F4, F6, right central: FC2, FC4, FC6, C2, C4, C6; right posterior: CP2, CP4, CP6, P2, P4, P6, PO4.

Following common practice In L2-related experiments (Morgan-Short & Tanner, 2014), time windows were selected in light of previous literature, and then adjusted based on visual inspection of the averaged ERP waveforms. For N400 (LDT and SEMREL), we considered N400's canonical latency (300-500ms), as well as N400's latency observed in the case of newly learned words (400-600 ms and 500-700 ms; e.g., Elgort et al., 2015; Pu et al., 2016). For P600 and AN (GJT), we considered 500-900 ms for P600 and late AN .

For each of the three experimental tasks, a within-subjects repeated-measures ANOVA was conducted. Repeated factors involved Laterality (three levels: left, midline, right) and Anterior-posterior distribution (three levels: anterior, central, posterior). In all analyses, the factor Type involved two levels. For LDT, two contrasts were investigated: Type *Well-known L2 French word* vs. Type *Pseudoword,* and Type *L2 French target* vs. Type *Pseudoword*. For SEMREL, the Types *Related* vs. *Unrelated well-known L2 French word*, as well as the Types *Related* vs. *Unrelated newly learned L2*

French target word. For GJT, the Types *Correct* vs. *Violated sentences containing a well-known L2 French word*, as well as the Types *Correct* vs. *Violated sentences containing a L2 French target word.* *p*-values were significant at .05. Partial eta squared (η_p^2) was used as effect size (small, $\eta p^2 > .0099$, medium, $\eta p^2 > .0588$, and large, $\eta p^2 > .1379$, Richardson, 2011). Bonferroni-correction was used to adjust for multiple comparisons. When necessary, Greenhouse-Geisser correction for non-sphericity was applied.

Results

Behavioral results

Paired-sample *t*-tests were used in order to analyze the behavioral outcomes. For LDT (Table 1), scores for well-known L2 French verbs and were significantly better than the scores on the target verbs ($t_{(19)} = 6.31$, $p < .001$).

Table 1. *Accuracy LDT (%)*

	Known		**Target**		**Pseudo**	
	M	***SD***	***M***	***SD***	***M***	***SD***
***N*=20**	99.75	1.09	84.75	12.60	84.25	11.21

For SEMREL (Table 2), the outcomes for the well-known words were better than the outcomes for the target words for both the Related ($t_{(18)} = 5.90$, $p < .001$) and Unrelated ($t_{(18)} = 2.71$, $p = .01$) categories. Comparisons also showed that there were no differences between the Known Related and the Known Unrelated categories ($t_{(18)} = -1.21$, $p = .24$).

Table 2. *Accuracy SEMREL (%)*

	Known Related		**Known Unrelated**		**Target Related**		**Target Unrelated**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
***N*=19**	94.50	7.57	97.25	4.02	75.25	16.39	90.25	11.78

Note. Results of one participant were excluded because of improper task performance.

For GJT (Table 3), there were no differences between the outcomes for correct sentences containing a known verb and correct sentences containing a target verb ($t_{(19)} = .41$, $p = .07$). There were also no differences between the results on the incorrect sentences containing a known verb and the

sentences based on a target verb ($t_{(19)}$ = -.78, p = .44). Additionally, participants performed better at recognizing correctly and incorrectly built sentences based on well-known verbs, although with a borderline significance ($t_{(19)}$ = 2.13, p = .046). Yet, no differences were observed when comparing correct and incorrect sentences containing a target verb ($t_{(19)}$ = 1.07, p = .30).

Table 3. *Accuracy GJT (%)*

	Known Correct		**Known Violated**		**Target Correct**		**Target Violated**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
***N*=20**	97.50	7.50	85.33	24.33	95.79	13.56	88.42	23.21

ERP Results

With respect to the results of the statistical analyses, main effects and interactions involving the factor Type that reach significance will be reported.

LDT. Visual inspection of waveforms for well-known L2 French words vs. pseudowords (Figure 6, top) revealed more negative-going brainwaves for pseudowords than for real words. The difference between both Types started at about 300 ms and was sustained. The effect was detectable at the midline electrodes and at left and right hemisphere sites near the midline, with an anterior, central and posterior distribution. For target verbs vs. pseudoverbs (Figure 6, bottom), more negative-going brainwaves for pseudoverbs were observed at anterior and central midline electrodes, as well as at the anterior electrodes (both left and right).

For the comparison well-known L2 French words vs. pseudowords, analyses between 300-500ms (see also Figure 7, left) revealed a Type X Hemisphere interaction ($F_{(2, 38)}$ = 4.13, p = .024, ηp^2 = .18). A follow-up ANOVA showed that mean amplitudes in the left hemisphere and at the midline were significantly more negative (left: $F_{(1, 19)}$ = 8.47, p = .009, ηp^2 = .31; midline: $F_{(1, 19)}$ = 8.88, p = .008, ηp^2 = .32) for pseudowords (left: M = -.54 µV, SE = .24; midline: M = .02 µV, SE = .43) than for well-known L2 French words (left: M = .09 µV; SE = .17; midline: M = .09 µV, SE = .33).

For the comparison targets vs. pseudowords, analyses in the 300-500ms time window revealed no main effects nor interactions. Analyses in the subsequent 400-600 ms time window (see also Figure

7, right) showed a main effect of Type ($F_{(1, 19)}$ = 4.74, p = .042, ηp^2 = .20) with mean values significantly lower for pseudowords (M = -.12 µV, SE = .24) than for targets (M = .29 µV, SE = .18).

<< INSERT FIGURE 6 HERE >>

<< INSERT FIGURE 7 HERE >>

SEMREL. Visual inspection of the waveforms for related vs. unrelated well-known L2 French words (Figure 8, top), showed more negative-going waveforms for unrelated pairs starting at about 300 ms post-stimulus with a sustained effect. This pattern was observed at anterior, central and posterior midline electrodes, and over a broad region of the right hemisphere. At the left hemisphere, this pattern was detectable at anterior, central and posterior sites near the midline. For waveforms depicting related vs. unrelated targets (Figure 8, bottom), a similar pattern was detected, although only at anterior and central electrodes, with a more bilateral distribution.

<< INSERT FIGURE 8 HERE >>

With respect to the comparison between related and unrelated well-known L2 French words, analyses in the 300-500ms time window (Figure 9, left) revealed a Type X Hemisphere interaction ($F_{(2, 36)}$ = 17.58, p < .001, ηp^2 = .49). Follow-up ANOVAs showed that unrelated pairs were significantly more negative than related pairs (midline: $F_{(1, 18)}$ = 33.65, p < .001, ηp^2 = .65; right: $F_{(1, 18)}$ = 5.96, p = .025, ηp^2 = .25) over the midline (related: M = .69 µV, SE = .38; unrelated: M = -.73 µV, SE = .34) and at right hemisphere (related: M = .84 µV, SE = .32; unrelated: M = .16 µV, SE = .23) electrodes.

For the comparison between related and unrelated pairs with targets, analyses between 300-500 ms (Figure 9, right) showed a Type X Anteriority interaction ($F_{(2, 36)}$ = 17.96, p = .018, ηp^2 = .20). Follow-up ANOVAs showed that unrelated pairs were significantly more negative than related pairs (anterior: $F_{(1, 18)}$ = 10.71, p = .022, ηp^2 = .26; central: $F_{(1, 18)}$ = 10.71, p = .004, ηp^2 = .37) at anterior (related: M = -1.55 µV, SE = .62 ; unrelated: M = -2.77 µV, SE = .50) and central sites (related: M = -.34 µV, SE = .49; unrelated: M = -1.15 µV, SE = .38).

<< INSERT FIGURE 9 HERE >>

GJT. Violated sentences containing well-known L2 French verbs showed more positive-going waveforms between 500-900 ms, bilaterally distributed over posterior and central sites (Figure 10, top). At anterior sites, brainwaves reflecting violated sentences showed a bilateral negative-going deflection.

For sentences containing a target (Figure 10, bottom), violations produced a positive-going deflection between 500-900ms at and near the midline electrodes, with a frontal to posterior distribution.

<< INSERT FIGURE 10 HERE >>

For the comparison between correct and violated sentences based on well-known L2 French verbs, analyses showed a Type X Hemisphere X Anteriority interaction ($F_{(4, 76)} = 5.46$, $p = .008$, $\eta p^2 = .16$). Follow-up ANOVAs showed two tendencies (Figure 11, left). First, at midline and right posterior sites, voltages were significantly more positive (midline: $F_{(1, 19)} = 7.39$, $p = .014$, $\eta p^2 = .28$; right: $F_{(1, 19)} = 9.89$, $p = .005$, $\eta p^2 = .34$) for violated sentences (midline: $M = 2.55$ µV, $SE = .78$; right: $M = 2.03$ µV, $SE = .74$) than for unviolated sentences (midline: $M = .42$ µV, $SE = .59$; right: $M = .49$ µV, $SE = .56$). Second, at right anterior electrode sites, voltages were significantly more negative ($F_{(1, 19)} = 11.15$, $p = .003$, $\eta p^2 = .37$) for violated sentences ($M = -2.31$ µV, $SE = .67$) than for correct sentences ($M = .12$ µV, $SE = .53$). The comparison between correct and violated sentences based on targets showed (Figure 11, right) a Type X Hemisphere interaction ($F_{(2, 38)} = 5.91$, $p = .011$, $\eta p^2 = .24$). Follow-up ANOVAs indicated that, at midline sites, voltages were significantly more positive ($F_{(1, 19)} = 27.74$, $p < .001$, $\eta p^2 = .59$) for violated sentences ($M = 1.32$ µV, $SE = .39$) than for correct sentences ($M = -.72$ µV, $SE = .41$).

<< INSERT FIGURE 11 HERE >>

Discussion

This study aimed at investigating whether ERP findings known from previous research (N400 elicited through LDT for form-related word knowledge, N400 elicited through SEMREL for meaning-related word knowledge, P600, AN elicited through GJT for grammar-related word knowledge) could be replicated for L2 vocabulary knowledge gained through ecologically valid L2 classroom instruction.

RQ 1. To what extent will LDT, SEMREL and GJT elicit indices of form-, meaning-, and use-related aspects of well-known L2 French words?

Through the first research question, we intended to test the validity of the experimental paradigm, that is, investigating whether the three tasks would produce the expected components. In order to answer the first research question, we operationalized form-related knowledge as the elicitation of N400 when comparing the ERP responses to well-known and to pseudo-words, meaning-related knowledge as the elicitation of N400 when comparing the ERP responses to related and unrelated well-

known words, and use-related knowledge as the elicitation of P600/AN when comparing the ERP responses to correct and violated sentences based on well-known vocabulary.

We have argued that the experimental tasks would be valid if the expected ERP responses could be replicated for stimuli based on well-established knowledge (Meulman et al., 2016; Woodman, 2010). As expected, results indicated that outcomes from previous research could be replicated, i.e., N400 for LDT and SEMREL, and P600 accompanied by anterior negativities (AN) for GJT. Consequently, we can ascertain the validity of the paradigms used, since the experimental tasks can detect form-, meaning-, and use-related word knowledge aspects in the case of stimuli based on well-known L2 French words.

RQ 2. To what extent will LDT, SEMREL and GJT elicit indices of form-, meaning-, and use-related aspects of newly learned L2 French words through ecologically valid classroom activities?

Through the second research question, we intended to investigate whether we could detect new word learning. In order to answer the second research question, we operationalized form-related knowledge as the elicitation of N400 when comparing ERP responses to targets and to pseudo-words, meaning-related knowledge as the elicitation of N400 when comparing the ERP responses to related and unrelated word pairs containing a target word, and use-related knowledge as the elicitation of P600/AN when comparing the ERP responses to correct and violated sentences based on target verbs.

For form, visual inspection shows that the elicited N400 in the 400-600ms time window is distributed over frontal and fronto-central sites. A similar pattern, i.e., a frontally detected N400, was reported in previous literature and has been linked with recently established vocabulary knowledge (Elgort et al., 2015; Pu et al., 2016). Likewise, for meaning relatedness, a frontally and centrally distributed N400 effect was found in the 300-500ms time window. This pattern was bilaterally distributed over frontal sites and central sites and near the midline. For use, results show P600 for syntactically violated sentences based on L2 target verbs. However, the pattern is different from the typical centro-parietal distribution, in that P600 for newly learned words is more broadly distributed over frontal sites. P600 with a frontal instead of a central-posterior distribution has been linked with diagnosis, reanalysis and online revision of syntactic structure (Friederici, Hahne, & Saddy, 2002). The differential distribution of P600 that we observed may also point to more effortful processing and suggests that the lexicogrammatical knowledge of the target verbs is less robustly represented than the lexicogrammatical knowledge of the well-known verbs. Furthermore, anterior sites show negativities that could indicate late AN, but this pattern did not reach significance. Consequently, we can conclude that the ERP responses for stimuli based on newly learned words echo findings of previous word learning studies that have been linked with newly established or emerging L2 word knowledge (i.e., N400 in the 400-600 ms time window for LDT, instead of the canonical 300-500 ms, and the more

anterior distribution of N400 for SEMREL and P600 for GJT, instead of the canonical central-parietal distribution).

With respect to the incremental nature of word learning (i.e., each new encounter with a lexical item adds small amounts of knowledge to the already existing substrate), the results of the present study, in addition to previous research, suggest that the multidimensionality (i.e., amplitude, latency, and topography) of ERPs can index different stages involved in word learning, going from initial L2 vocabulary knowledge to well-established L2 vocabulary knowledge. Further research into the developmental trajectory of ERP patterns over time, would undoubtedly contribute to a better understanding of the theoretically fundamental concept of incremental L2 vocabulary knowledge. Furthermore, noticing, i.e., making the learner aware that there is something new to learn by focusing the attention on the target words (Laufer, 2020) is considered to be key for the efficiency of form-focused L2 approaches (e.g., Webb, 2020). From this perspective, ERP research may provide a more fine-grained understanding about the differential effects of learning conditions with either a lot (form-focused) or few (meaning-focused) opportunities for noticing. Third, our results also suggest that ERP measures can contribute to questions concerning the development of automatic processing. As such, deviations from the canonical distribution and latency are said to be indicative of lower proficiency and more effortful processing (Elgort et al., 2015, Frishkoff et al., 2010; Mestress-Missé et al., 2007; Pu et al., 2016). On the other hand, in sentence processing studies, the involvement of anterior components, combined with P600, can either suggest higher proficiency and automaticity (Morgan-Short, et al., 2012; Steinhauer & Drury, 2012), or more effortful processing and increased working memory demands when they are sustained (Baggio et al., 2008; King & Kutas, 1995; Reichle & Birdsong, 2014).

Additionally, the present conceptual replication study may also contribute to a better understanding of the relationship between behavioral results and neurocognitive outcomes. Given the paucity of ERP studies that have focused on the effects of instruction and the discrepancies between neurocognitive findings and behavioral results, Morgan-Short and Ullman (2014) claim that the current stage of L2 neurocognitive research does not allow for strong pedagogical conclusions and warrant more research that is driven by L2-pedagogy oriented research questions (Morgan-Short & Ullman, 2014; Rastelli, 2018). From this perspective, the present research intends to contribute to refining our understanding of the relationship between ERPs, L2 learning and behavioral performance.

Limitations

This conceptual replication study inevitably presents some limitations. First, participants were not pre-tested whether and to what extent they actually had previous knowledge about the target words.

We argued that the target words were sufficiently unfamiliar or infrequent. The behavioral results on LDT show that there was a significant difference regarding their ability to correctly recognize word forms, suggesting that the target word forms were not or less familiar than the well-known word forms at the outset of the study (which is plausible, given the lower frequencies of the targets). Yet, in the global research project with secondary school participants, prior knowledge about the targets will be assessed at the outset of the study by using a multiple-choice meaning recognition test. We acknowledge that this is a shortcoming in this conceptual replication study.

Second, on a more general level, this study intended to investigate the impact of L2 vocabulary instruction within an ecologically valid learning setting. Therefore, an online learning environment was developed consisting of activities that are representative of current L2 classroom practice. Yet, while the measurement instruments (LDT, SEMREL, GJT) are widely used formats in experimental psychology, these paradigms are not used as assessments instruments in the realistic L2 classroom. Moreover, ecologically valid L2 classroom assessment taps into both receptive (e.g., understanding the meaning of a word while reading) and productive (e.g., using new vocabulary in a conversation) knowledge. Due to the limitations of the ERP technique (e.g., avoiding noise coming from muscular activity), the tests used in this study only tapped into receptive knowledge. Consequently, from a pedagogical perspective that emphasizes language use through skill, the measurement tools that have been used may suffer from a lack of face validity, i.e., not being perceived as valuable because of the absence of a familiarity link between L2 classroom tasks and experimental tasks (Nation & Webb, 2011).

Conclusion

This conceptual replication study had two main goals. First, investigating whether and to what extent ecologically valid L2 classroom learning could produce the ERP responses elicited in laboratory settings and longitudinal studies, and, second, testing the validity of the experimental tasks (LDT, SEMREL, GJT). Given the elicitation of N400 and P600 for newly learned L2 French targets, we can conclude that a short-term treatment based on an ecologically valid L2 learning environment can replicate ERP responses that have been elicited in previous research. In line with previous findings, deviations from the canonical topography (more anteriorly distributed N400 and P600) and the delayed time-window for N400 (400-600 ms), have repeatedly been reported in the case of newly established lexical knowledge. In addition to the above discussed limitations, the significance of these findings for the main study with secondary school learners is twofold. First, the learning environment developed in the context of the global research project, containing both form-focused and meaning-focused learning paths, can be used for ERP research. Second, the experimental tasks that were used can be considered as valid ERP measurement instruments.

Conflict of interest

The authors have no conflict of interest.

Funding information

This work was funded by the Research Foundation Flanders (FWO), grant G064116N.

References

Bakker, I., Takashima, A., Van Hell, J.G., Janzen, G., & McQueen, J. (2015). Tracking lexical consolidation with ERPs: lexical and semantic-priming effects on N400 and LPC responses to newly learned words. *Neuropsychologia, 79*, 33-41. https://doi.org/10.1016/j.neuropsychologia.2015.10.020

Baggio, G., van Lambalgen, M., & Hagoort, P. (2008). Computing and recomputing discourse models: an ERP study. *Journal of Memory and Language, 59*, 36-53. https://doi.org/10.1016/j.jml.2008.02.005

Balass, M., Nelson, J.R., & Perfetti, C.A. (2010). Word learning: an ERP investigation of word experience effects on recognition and word processing. *Contemporary Educational Psychology, 35*, 126-140. https://doi.org/10.1016/j.cedpsych.2010.04.001

Choi, S., Kim, J., & Ryu, K. (2014). Effects of context on implicit and explicit lexical knowledge: an event-related potential study. *Neuropsychologia, 63,* 226-234. https://doi.org/10.1016/j.neuropsychologia.2014.09.003

Chun, E., Choi, S., & Kim, J. (2012). The effect of extensive reading and paired-associate learning on long-term vocabulary retention: an event-related potential study. *Neuroscience Letters, 521,* 125-129. https://doi.org/10.1016/j.neulet.2012.05.069

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

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

Diener, E., & Biswas-Diener, R. (2020). The replication crisis in psychology. In R. Biswas & E. Diener (Eds.), *Noba Textbook Series*: Psychology. Champaign, IL: DEF Publishers. Retrieved from http://noba.to/q4cvydeh

Elgort, I. (2018). Technology-mediated second language vocabulary development: a review of trends in research methodology. *Calico Journal, 35,* 1-29. https://doi.org/10.1558/cj.34554

Elgort, I., Perfetti, C.A., Rickles, B., & Stafura, J.Z. (2015). Contextual learning of L2 word meanings: second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning. *Language, Cognition and Neuroscience, 30,* 506-528. https://doi.org/10.1080/23273798.2014.942673

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

Frishkoff, G.A., Perfetti, C., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology, 35,* 376-403. https://dx.doi.org/10.1080%2F87565641.2010.480915

Godfroid, A. (2020). Sensitive measures of vocabulary knowledge and processing: expanding Nation's framework. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 433-453). New York/London: Routledge.

King, J.W., & Kutas, M. (1995). Who did what and when? Using word- and clause-level ERPs to monitor working memory usage in reading. *Journal of Cognitive Neuroscience, 7,* 376-395. https://doi.org/10.1162/jocn.1995.7.3.376

Laszlo, S., & Federmeier, K.D. (2009). A beautiful day in the neighborhood: an event-related potential study of lexical relationship and prediction in context. *Journal of Memory and Language, 61,* 326-338. https://doi.org/10.1016/j.jml.2009.06.004

Laufer, B. (2020). Evaluating exercises for learning vocabulary. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 351-368). New York/London: Routledge.

Laufer, B. (2017). The three "I"s of second language vocabulary learning. Input, Instruction, Involvement. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning. Volume III* (pp. 343-354). New York/London: Routledge.

Lonsdale, D., & Le Bras, Y. (2009). *A Frequency Dictionary of French Core Vocabulary for Learners*. London/New York: Routledge.

Lopez-Calderon, J., & Luck, S.J. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience,* 8:213. https://doi.org/10.3389/fnhum.2014.00213

Luck, S.J. (2014). *An Introduction to the Event-Related Potential Technique. Second Edition*. Cambridge (MA): The MIT Press.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience, 7,* 703-704. https://doi.org/10.1038/nn1264

Mestres-Missé, A., Rodríguez-Fornells, A., & Münte T.F. (2007). Watching the brain during meaning acquisition. *Cerebral Cortex, 17,* 1858-1866. https://doi.org/10.1093/cercor/bhl094

Meulman, N., Seton, B.J., Stowe, L.A., & Schmid, M.S. (2016). EEG and event-related brain potentials. In M.S. Schmid, S.M. Berends, C. Bergmann, S.M. Brouwer, N. Meulman, B. Seton, S.

Sprenger, & Stowe L. *Designing Research on Bilingual Development. Behavioral and Neurolinguistic Experiments* (pp. 81-107). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-11529-0

Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: a field reaching its potential. *Annual Review of Applied Linguistics, 34*, 15-36. https://doi.org/10.1017/S026719051400004X

Morgan-Short, K., Faretta-Stutenberg, M., & Bartlett-Hsu, L. (2015). Contributions of event-related potential research into issues in explicit and implicit second language acquisition. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* (pp. 349-383). Amsterdam/Philadelphia : John Benjamins Publishing Company. https://doi.org/10.1111/psyp.12283

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*, 933-947. https://doi.org/10.1162/jocn_a_00119

Morgan-Short, K., & Tanner, D. (2014). Event-related potentials (ERPs). In J. Jegerski, & B. VanPatten (Eds.), *Research Methods in Second Language Psycholinguistics* (pp. 127-152). New York: Routledge

Nation, I.S.P. (2013). *Learning Vocabulary in Another Language. Second Edition.* Cambridge: Cambridge University Press.

Nation, I.S.P., & Webb, S. (2011). *Researching and Analyzing Vocabulary.* Boston: Heinle.

Newman, A.J., Tremblay, A., Nichols, E.S., Neville, H.J., & Ullman, M.T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience, 24*, 1205-1223. https://dx.doi.org/10.1162%2Fjocn_a_00143

Ojima, S., Nakamura, N., Matsuba-Kurita, H., Hoshino, T., & Hagiwara, H. (2011). Neural correlates of foreign-language learning in childhood: a 3-year longitudinal ERP study. *Journal of Cognitive Neuroscience, 23*, 183-199. https://doi.org/10.1162/jocn.2010.21425

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. https://doi.org/10.1162/0898929055002436

Osterhout, L., Pitkänen, I., McLaughlin, J., & Zeitlin, M. (2019). Event-related potentials as metrics of foreign-language learning and loss. In M. Schmid, & B. Köpke (Eds.), *The Oxford Handbook of Language Attrition (pp. 403-416).* Oxford: Oxford University Press.

Perfetti, C.A., Wlotko, E.W., & Hart, L.A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1281-1292. https://doi.org/10.1037/0278-7393.31.6.1281

Pu, H., Holcomb, P.J., & Midgley, K.J. (2016). Neural changes underlying early stages of L2 vocabulary acquisition. *Journal of Neurolinguistics, 40,* 55-65. https://doi.org/10.1016/j.jneuroling.2016.05.002

Rastelli, S. (2018). Neurolinguistics and second language teaching: a view from the crossroads. *Second Language Research, 34,* 103-123.

https://doi.org/10.1177/0267658316681377

Reichle, R.V., & Birdsong, D. (2014). Processing focus structure in L1 and L2 French. L2 proficiency effects on ERPs. *Studies in Second Language Acquisition, 36*, 535-564. https://doi.org/10.1017/S0272263113000594

Richardson, J.T.E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*, 135-147. https://doi.org/10.1016/j.edurev.2010.12.001

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329-363. https://doi.org/10.1177/1362168808089921

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: a research agenda. *Language Teaching, 52*, 261-274. doi:10.1017/S0261444819000053

Soskey, L., Holcomb, P.J., & Midgley, K.J. (2016). Language effects in second-language learners: a longitudinal electrophysiological study of Spanish classroom learning. *Brain Research, 1646*, 44-52. https://dx.doi.org/10.1016/j.brainres.2016.05.028

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

Stowe, L.A., Kaan, E., Sabourin, L., & Taylor, R.C. (2018). The sentence wrap-up dogma. *Cognition, 176*, 232-247. https://doi.org/10.1016/j.cognition.2018.03.011

Vandenberghe, B., Montero Perez, M., Reynvoet, B., Desmet, P. (2019). The role of Event-Related Potentials (ERPs) as sensitive measures in L2 vocabulary acquisition research. *Journal of the European Second Language Association, 3*, 35-45. http://doi.org/10.22599/jesla.60

Vandenberghe, B., Montero Perez, M., Reynvoet, B., Desmet, P. (2021). Combining explicit and sensitive indices for measuring L2 vocabulary learning through contextualized input and word-focused instruction. *Studies in Second Language Acquisition, 43*, 1009-1039. http://doi.org/10.1017/S0272263120000431

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 225-239). New York/London: Routledge.

Webb, S., & Nation, I.S.P. (2017). *How Vocabulary is Learned.* Oxford: Oxford University Press.

Woodman, G.F. (2010). A brief introduction to the use of event-related potentials (ERPs) in studies of perception and attention. *Attention, Perception, & Psychophysics, 72*, 2031-2046. https://doi.org/10.3758/APP.72.8.2031

Yum, Y., Midgley, K. J., Holcomb, P. J., & Grainger, J. (2014). An ERP study on initial second language vocabulary learning. *Psychophysiology, 51*, 364-373. https://doi.org/10.1111/psyp.12183